UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/024,967	12/30/2004	Daniel Egnor	0026-0130	7261

44989 7590 01/13/2012

HARRITY & HARRITY, LLP
11350 Random Hills Road
SUITE 600
FAIRFAX, VA 22030

EXAMINER	
HWA, SHYUE JIUNN	
ART UNIT	**PAPER NUMBER**
2163	

MAIL DATE	DELIVERY MODE
01/13/2012	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE BOARD OF PATENT APPEALS
AND INTERFERENCES

Ex parte DANIEL EGNOR and GEETA CHAUDHRY

Appeal 2009-012635
Application 11/024,967
Technology Center 2100

Before JEFFREY S. SMITH, ERIC B. CHEN, BRUCE R. WINSOR, *Administrative Patent Judges*.

SMITH, *Administrative Patent Judge*.

DECISION ON APPEAL

STATEMENT OF THE CASE

This is an appeal under 35 U.S.C. § 134(a) from the Examiner's final rejection of claims 1-29, which are all the claims pending in the application. We have jurisdiction under 35 U.S.C. § 6(b).

We affirm.

Invention

Appellants' invention relates to a system that determines documents that are associated with a location, identifies a group of signals associated with each of the documents, and determines authoritativeness of the documents for the location based on the signals. Abstract.

Representative Claim

1. A method comprising:

identifying a set of documents, as candidate documents, that are all associated with a same geographic location;

identifying signals associated with the candidate documents;

determining a measure of authoritativeness of the candidate documents for a business at the location based on the signals; and

processing the candidate documents based on their measures of authoritativeness for the business at the location.

Prior Art

Agoni	US 2002/0133374 A1	Sep. 19, 2002
Getchius	US 6,643,640 B1	Nov. 4, 2003
Nye	US 2004/0064334 A1	Apr. 1, 2004

Examiner's Rejections

Claims 1-4, 6-8, 10, 12-18, 20-22, 24, and 26-28 stand rejected under 35 U.S.C. § 103(a) as being unpatentable over Getchius and Agoni.

Claims 5, 9, 11, 19, 23, and 25 stand rejected under 35 U.S.C. § 103(a) as being unpatentable over Getchius, Agoni, and Nye.

Claim 29 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over Getchius and Nye.

Claim Groupings

Based on Appellants' arguments in the Appeal Brief, we will decide the appeal on the basis of claims 1, 4, 6-8, 10, and 29.

FINDINGS OF FACT

We rely on, and adopt as our own, the findings of fact set forth by the Examiner in the Final Rejection and Examiner's Answer.

ANALYSIS

Section 103 rejection of claims 1-3, 5, 9, 11-17, 19, 23, 25, and 26- 28

Appellants contend that the combination of Getchius and Agoni does not teach "determining a measure of authoritativeness of the candidate documents for a business at the location based on the signals" as recited in claim 1. App. Br. 9-11; Reply Br. 4-10. The Examiner finds that Getchius teaches this limitation. Ans. 6; 23-28. We agree with the Examiner.

Appellants contend that Getchius does not teach "identifying signals associated with the candidate documents," therefore, Getchius cannot teach determining authoritativeness based on "signals associated with the candidate documents." App. Br. 12. The Examiner finds that the combination of Getchius and Agoni teaches identifying signals associated with the candidate documents. Ans. 7. We agree with the Examiner. We further find that the term "identifying signals associated with the candidate documents" as recited in claim 1 encompasses identifying any data associated with the candidate documents, such as the data discussed in col. 28, ll. 7-11 of Getchius.

We sustain the rejection of claim 1 under 35 U.S.C. § 103 for the reasons set forth by the Examiner in the action from which this appeal is taken and the reasons set forth by the Examiner in the Examiner's Answer in response to the Appellants' Appeal Brief. We concur with the conclusion reached by the Examiner.

Appellants present arguments for independent claims 14 and 15 (App. Br. 26-37) similar to those presented for claim 1, which we find unpersuasive. Appellants have not presented arguments for separate

patentability of claims 2, 3, 5, 9, 11-13, 16, 17, 19, 23, and 25-28, which thus fall with corresponding independent claims 1 and 15.

Section 103 rejection of claims 4 and 18

Appellants contend that the combination of Getchius and Agoni does not teach "determining documents that are linked to by the candidate documents, and identifying the determined documents as candidate documents." App. Br. 15-16; Reply Br. 10-12. The Examiner finds that the combination of Getchius and Agoni teaches the limitations of claim 4. Ans. 8, 29-30. We agree with the Examiner.

We sustain the rejection of claim 4 under 35 U.S.C. § 103 for the reasons set forth by the Examiner in the action from which this appeal is taken and the reasons set forth by the Examiner in the Examiner's Answer in response to the Appellants' Appeal Brief. We concur with the conclusion reached by the Examiner.

Appellants present arguments for claim 18 (App. Br. 38-39) similar to those presented for claim 4, which we find unpersuasive. We sustain the rejection of claim 18 under 35 U.S.C. § 103.

Section 103 rejection of claims 6 and 20

Appellants contend that the combination of Getchius and Agoni does not teach "determining a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes: generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate

documents that point to the candidate document" as recited in claim 6. App. Br. 16-19; Reply Br. 13-15. The Examiner finds that the combination of Getchius and Agoni teaches this limitation. Ans. 8, 30-31. We agree with the Examiner.

We sustain the rejection of claim 6 under 35 U.S.C. § 103 for the reasons set forth by the Examiner in the action from which this appeal is taken and the reasons set forth by the Examiner in the Examiner's Answer in response to the Appellants' Appeal Brief. We concur with the conclusion reached by the Examiner.

Appellants present arguments for claim 20 (App. Br. 39-42) similar to those presented for claim 6, which we find unpersuasive. We sustain the rejection of claim 20 under 35 U.S.C. § 103.

Section 103 rejection of claims 7 and 21

Appellants contend that the combination of Getchius and Agoni does not teach "identifying anchor text associated with links to the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes: generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location" as recited in claim 7. App. Br. 19-21; Reply Br. 15-17. The Examiner finds that the combination of Getchius and Agoni teaches this limitation. Ans. 8-9, 31-32. We agree with the Examiner.

We sustain the rejection of claim 7 under 35 U.S.C. § 103 for the reasons set forth by the Examiner in the action from which this appeal is

taken and the reasons set forth by the Examiner in the Examiner's Answer in response to the Appellants' Appeal Brief. We concur with the conclusion reached by the Examiner.

Appellants present arguments for claim 21 (App. Br. 42-44) similar to those presented for claim 7, which we find unpersuasive. We sustain the rejection of claim 21 under 35 U.S.C. § 103.

Section 103 rejection of claims 8 and 22

Appellants contend that the combination of Getchius and Agoni does not teach "identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes: generating an authoritative score for one of the candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location" as recited in claim 8. App. Br 21-24; Reply Br. 17-20. The Examiner finds that the combination of Getchius and Agoni teaches this limitation. Ans. 9, 32-33. We agree with the Examiner.

We sustain the rejection of claim 8 under 35 U.S.C. § 103 for the reasons set forth by the Examiner in the action from which this appeal is taken and the reasons set forth by the Examiner in the Examiner's Answer in response to the Appellants' Appeal Brief. We concur with the conclusion reached by the Examiner.

Appellants present arguments for claim 22 (App. Br. 44-46) similar to those presented for claim 8, which we find unpersuasive. We sustain the rejection of claim 22 under 35 U.S.C. § 103.

Section 103 rejection of claims 10 and 24

Appellants contend that the combination of Getchius and Agoni does not teach "determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes: increasing the measure of authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location" as recited in claim 10. App. Br. 24-25; Reply Br. 20-22. The Examiner finds that the combination of Getchius and Agoni teaches this limitation. Ans. 9-10, 33-34. We agree with the Examiner.

We sustain the rejection of claim 10 under 35 U.S.C. § 103 for the reasons set forth by the Examiner in the action from which this appeal is taken and the reasons set forth by the Examiner in the Examiner's Answer in response to the Appellants' Appeal Brief. We concur with the conclusion reached by the Examiner.

Appellants present arguments for claim 24 (App. Br. 46-48) similar to those presented for claim 10, which we find unpersuasive. We sustain the rejection of claim 24 under 35 U.S.C. § 103.

Section 103 rejection of claim 29

Appellants contend that the combination of Getchius and Nye does not teach the limitations of claim 29. App. Br. 56-61. The Examiner finds that the combination of Getchius and Nye teaches the limitations of claim 29. Ans. 17-22. We agree with the Examiner.

We sustain the rejection of claim 29 under 35 U.S.C. § 103 for the reasons set forth by the Examiner in the action from which this appeal is

taken and the reasons set forth by the Examiner in the Examiner's Answer in response to the Appellants' Appeal Brief. We concur with the conclusion reached by the Examiner.

DECISION

The rejection of claims 1-4, 6-8, 10, 12-18, 20-22, 24, and 26-28 under 35 U.S.C. § 103(a) as being unpatentable over Getchius and Agoni is affirmed.

The rejection of claims 5, 9, 11, 19, 23, and 25 under 35 U.S.C. § 103(a) as being unpatentable over Getchius, Agoni, and Nye is affirmed.

The rejection of claim 29 under 35 U.S.C. § 103(a) as being unpatentable over Getchius and Nye is affirmed.

No time period for taking any subsequent action in connection with this appeal may be extended under 37 C.F.R. § 1.136(a). *See* 37 C.F.R. § 41.50(f).

AFFIRMED

tj

INFORMATION DISCLOSURE CITATION PTO-1449	CUSTOMER NUMBER 44989	ATTORNEY'S DKT NO. 0026-0130	APPLICATION NO. 11/024,967
		APPLICANT(S) Daneil Egnor et al.	
		FILING DATE December 30, 2004	GROUP 2163

U.S. PATENT DOCUMENTS

EXAMINER'S INITIALS	PATENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE

FOREIGN PATENT DOCUMENTS

EXAMINER'S INITIALS	PATENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Partial Translation	
						Yes	No
	JP 2000-348041 A (w/ English Abstract)	12-15-00	Japan			X	
	JP 2003-524259 A	08-12-03	Japan			X	
	JP 2004-227165 A (w/ English Abstract)	08-12-04	Japan			X	
	JP 2000-250931 A (w/ English Abstract)	09-14-00	Japan			X	
	JP 2003-173280 A (w/ English Abstract)	06-20-03	Japan			X	
	JP 2003-067419 A (w/ English Abstract)	03-07-03	Japan			X	
	WO 01/63479 A1	08-30-01	WIPO				

OTHER DOCUMENTS (Including Author, Title, Date, Pertinent Pages, Etc.)

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant(s).

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-250931

(43)Date of publication of application : 14.09.2000

(51)Int.Cl. G06F 17/30
G06F 17/27

(21)Application number : 11-053137

(22)Date of filing : 01.03.1999

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

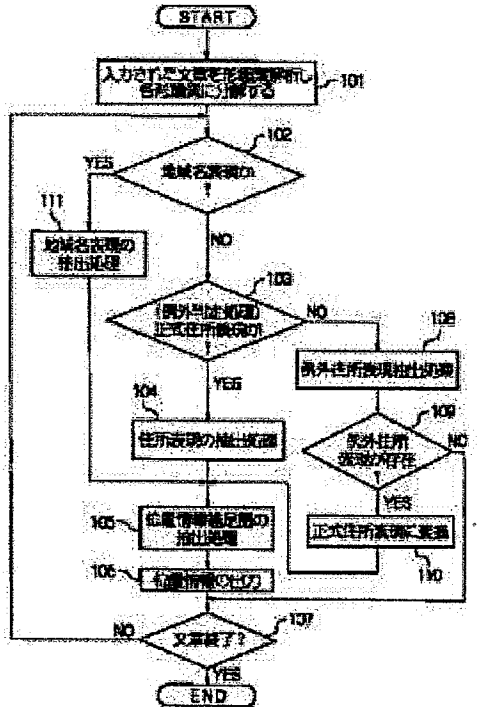
(72)Inventor : SUGIURA HIRONOBU
TSUCHIYA HIDEYUKI

(54) DEVICE AND METHOD FOR AUTOMATIC EXTRACTION OF POSITIONAL INFORMATION AND RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To automatically extract the positional information in the expressions of addresses and regional names which are included in a document with high probability.

SOLUTION: An inputted sentence is decomposed into morphemes (101), and every morpheme is compared with regional name expressions (102). An exception decision processing is carried out to decide whether the expression including a morpheme is a formal address expression (103). If a formal address expression is confirmed, the morpheme is successively compared with all address expressions of Japan and an address expression is extracted (104). If no formal address expression is confirmed in exception decision processing, the coincidence is retrieved between those morphemes and the exceptional address expressions and an exceptional address expression is extracted. Then the names of a prefecture, a city and a district are added to the extracted exceptional address expression and this address expression is converted into a formal address expression (108-110). If a positional information complementary word is included within six words of the extracted position information, the positional information including the positional information



complementary word is outputted (106).

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]An automatic extracting method of position information for extracting position information included in an inputted text characterized by comprising the following.
Processing decomposed into two or more morphemes by dividing an inputted text per 1 or two or more character strings.
Processing judged for whether one morpheme in said each morpheme is made into a retrieval object morpheme, and this retrieval object morpheme is in agreement with a character string of regional name expression registered beforehand.
Exception decision processing which judges **** as whether expression containing said retrieval object morpheme is a formal address expression.
In said exception decision processing, it is said retrieval object morpheme.

[Claim 2]Processing to which said exception decision processing carries out identity retrieval of said retrieval object morpheme and a character string of an all-prefectures name registered beforehand, Processing which makes the next morpheme of said retrieval object morpheme a new retrieval object morpheme when said retrieval object morpheme really uses an all-prefectures name in identity retrieval with a character string of said all-prefectures name, Processing which performs identity retrieval of said retrieval object morpheme and a character string of a cities, towns and villages division name registered beforehand, Processing judged as expression containing said retrieval object morpheme being a formal address expression when said retrieval object morpheme is in agreement with a character string of a cities, towns and villages division name in processing which performs identity retrieval with said cities, towns and villages division name, In identity retrieval with a character string of said all-prefectures name, said retrieval object morpheme is not in agreement with a character string of an all-prefectures name, And an automatic extracting method of the position information

according to claim 1 which comprises processing judged as expression containing said retrieval object morpheme not being a formal address expression when a retrieval object morpheme is not in agreement with a character string of a cities, towns and villages division name in processing which performs identity retrieval with said cities, towns and villages division name.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]It is related with the automatic extracting method for extracting automatically position information included in the text constituted by the electronized text, such as address expression or regional name expression, and a device.
[0002]
[Description of the Prior Art]There was a method of whether conventionally, the character string which shows the position information specified by the user as a method of searching the position information included in the text is contained in the text, and searching the whole sentence of a text. In order to look for the position information specified by a user, the whole sentence of a text will be searched with this method. However, in this method, search time will also increase substantially with the increase in the target amount of texts. Therefore, when extracting beforehand position information included in the text, such as address expression or regional name expression, and searching, shortening search time is called for by investigating only the extracted position information.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL FIELD

[Field of the Invention]It is related with the automatic extracting method for extracting automatically position information included in the text constituted by the electronized text, such as address expression or regional name expression, and a device.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art]There was a method of whether conventionally, the character string which shows the position information specified by the user as a method of searching the position information included in the text is contained in the text, and searching the whole sentence of a text. In order to look for the position information specified by a user, the whole sentence of a text will be searched with this method. However, in this method, search time will also increase substantially with the increase in the target amount of texts. Therefore, when extracting beforehand position information included in the text, such as address expression or regional name expression, and searching, shortening search time is called for by investigating only the extracted position information.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

EFFECT OF THE INVENTION

[Effect of the Invention]As explained above, this invention has the effect that time to search the position information in a text is shortened substantially, when becoming possible to extract automatically address expression or regional name expression described in the text.
[0066]When this invention is applied to a geographic information system, it has the effect of becoming possible to stick information on a geographic information system automatically, by extracting position information automatically out of information, including a newspaper article etc.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]In conventional technology, there was a problem that position information on various expressions included in the text could not be extracted.
[0005]An object of this invention is to provide the automatic extracting apparatus and method of position information that the position information included in the text can be extracted with high probability.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

MEANS

[Means for Solving the Problem]To achieve the above objects, an automatic extracting method of position information on this invention, Processing decomposed into two or more morphemes by dividing a text which is the automatic extracting method of position information for extracting position information included in an inputted text, and was inputted per 1 or two or more character strings, Processing judged for whether one morpheme in said each morpheme is made into a retrieval object morpheme, and this retrieval object morpheme is in agreement with a character string of regional name expression registered beforehand, Exception decision processing which judges **** as whether expression containing said retrieval object morpheme is a formal address expression, When it judges that expression which contains said retrieval object morpheme in said exception decision processing is a formal address expression, Address extracting processing which extracts address expression from said text by performing identity retrieval of said retrieval object morpheme and a character string of address expression of the Japan whole country registered beforehand one by one, Exception address expression is extracted by performing identity retrieval of said retrieval object morpheme and a character string of exception address expression registered beforehand, when it judges that expression which contains said retrieval object morpheme in said exception decision processing is not a formal address expression, Exception address extracting processing which adds a character string or an omitted county name of a "prefecture" omitted by this extracted exception address expression or a "city", and is changed into a formal address expression, To each morpheme within fixed numbers, from an end of regional name expression and address expression which were extracted. It searches whether the same character string as a position information supplementary word registered beforehand exists, and when it exists, it comprises position information supplementary word extracting processing which makes position information this regional name expression or from address expression to position information supplementary word.

[0007]This invention decomposes an inputted text into a morpheme, and extracts a local expression or address expression by carrying out identity retrieval to an address name of the whole country which consists of each morpheme, a regional name registered beforehand or an all-prefectures name, a cities, towns and villages division name, Oaza and a common-name name, and a character and a ** name. And when address expression is exception address expression with which exception address expression or a county name to which a character string of a "prefecture" or a "city" was abbreviated from a formal address expression was omitted. It judges with it not being a formal address expression in exception decision processing, and is made to change into a formal address expression by compensating exception address expression with a "prefecture", a "city", or a county name. Therefore, in an automatic extracting method of position information on this invention, while being able to extract automatically position information included in a text, such as a formal address expression, address expression which is not formal, and regional name expression, with high probability, position information including a position information supplementary word can be extracted.

[0008]An automatic extracting method of position information on this invention comprises:

Processing to which said exception decision processing carries out identity retrieval of said retrieval object morpheme and a character string of an all-prefectures name registered beforehand.

Processing which makes the next morpheme of said retrieval object morpheme a new retrieval object morpheme when said retrieval object morpheme really uses an all-prefectures name in identity retrieval with a character string of said all-prefectures name.

Processing which performs identity retrieval of said retrieval object morpheme and a character string of a cities, towns and villages division name registered beforehand.

Processing judged as expression containing said retrieval object morpheme being a formal address expression when said retrieval object morpheme is in agreement with a character string of a cities, towns and villages division name in processing which performs identity retrieval with said cities, towns and villages division name, In identity retrieval with a character string of said all-prefectures name, said retrieval object morpheme is not in agreement with a character string of an all-prefectures name, And processing judged as expression containing said retrieval object morpheme not being a formal address expression when a retrieval object morpheme is not in agreement with a character string of a cities, towns and villages division name in processing which performs identity retrieval with said cities, towns and villages division name.

[0009]An automatic extracting method of position information on this invention comprises:

Processing to which said address extracting processing makes the next morpheme of said

retrieval object morpheme a new retrieval object morpheme.
Said retrieval object morpheme.
Processing which performs identity retrieval with a character string of Oaza and a common-name name registered beforehand.
Processing which makes the next morpheme of said retrieval object morpheme a new retrieval object morpheme when said retrieval object morpheme is in agreement with a character string of Oaza and a common-name name in identity retrieval with a character string of said Oaza and common-name name, Processing which performs identity retrieval of said retrieval object morpheme and a character string of eye ** [a character and] were registered beforehand, Processing which makes the next morpheme of said retrieval object morpheme a new retrieval object morpheme when said retrieval object morpheme is in agreement with a character string of eye ** [character and] in identity retrieval with a character string of eye said [character and **], When said retrieval object morpheme is not in agreement with a character string of Oaza and a common-name name in identity retrieval with a character string of said Oaza and common-name name, Or when said retrieval object morpheme is not in agreement with a character string of eye ** [character and] in identity retrieval with a character string of eye said [character and **], Or in identity retrieval with a character string of said Oaza and common-name name, said retrieval object morpheme is in agreement also with a character string of Oaza and a common-name name, And when said retrieval object morpheme is in agreement also with a character string of eye ** [character and] in identity retrieval with a character string of eye said [character and **], Address item extracting processing extracted noting that it is an address item [in / for this number / address expression], when it judges whether the next morpheme of a retrieval object morpheme is a number, and it is a number, and a morpheme of an extracted all-prefectures name, Processing which connects a morpheme of a cities, towns and villages name, a morpheme of Oaza and a common name, a morpheme of eye ** [a character and], and a morpheme of an address item, and is considered as one address expression.

[0010]An automatic extracting method of position information on this invention said exception address extracting processing, Exception address type decision processing which judges whether expression judged that is not a formal address expression in said address extracting processing is exception address expression with which a county name was omitted from a formal address expression, When judged with it being exception address expression with which a county name was omitted from address expression with a formal expression which contains a retrieval object morpheme in said exception address type decision processing, said retrieval object morpheme, Processing which carries out identity retrieval of the character string to which a "prefecture" was abbreviated from an all-prefectures name registered

beforehand, Processing which carries out identity retrieval of said retrieval object morpheme and the character string to which a "city" was abbreviated from a city name registered beforehand, Processing which makes the next morpheme of said retrieval object morpheme a new retrieval object morpheme when said retrieval object morpheme of a "city" corresponds with an omitted character string in identity retrieval with a character string to which the above "city" was abbreviated, In name-of-a-person decision processing which carries out identity retrieval and said name-of-a-person decision processing with a character string of expression used when describing a name of a person registered beforehand to be said retrieval object morpheme, a retrieval object morpheme, It has the processing judged as expression containing a morpheme to be examined not being address expression when in agreement with a character string of expression used when describing a name of a person, and when a morpheme to be examined is not in agreement with a character string by which a "city" was abbreviated also to a character string to which a "prefecture" was abbreviated.
[0011]An automatic extracting method of position information on this invention said exception address extracting processing, When judged with it not being exception address expression with which a county name was omitted from address expression with a formal expression which contains a retrieval object morpheme in said exception address type decision processing, A towns-and-villages name which was in agreement in identity retrieval with a cities, towns and villages division name based on an all-prefectures name searched in identity retrieval with a character string of an all-prefectures name, Identity retrieval with a character string which omitted only a group name from a formal county-towns-and-villages name registered beforehand is performed, It has further processing with which a county name omitted by considering it as address expression which had a character string of a formal county-towns-and-villages name corresponding to the towns-and-villages name searched instead of congruous towns-and-villages names is compensated, and the processing which makes the next morpheme of said retrieval object morpheme a new retrieval object morpheme.

[0012]An automatic extracting method of position information on this invention comprises:
A morpheme which has said position information supplementary word extracting processing within fixed limits from an end of address expression or regional name expression.
Processing which performs identity retrieval with a position information supplementary word registered beforehand.
Processing which extracts even morphemes congruous from said address expression or regional name expression as one position information when in agreement in identity retrieval processing with said position information supplementary word.
Processing which returns a retrieval object morpheme six words ago when not in agreement in identity retrieval processing with said position information supplementary word.

[0013]
[Embodiment of the Invention]Next, the embodiment of this invention is described in detail with reference to Drawings.
[0014]Drawing 1 is a block diagram showing the composition of the automatic extracting apparatus of the position information on one embodiment of this invention.
[0015]The automatic extracting apparatus of the position information on this embodiment The address database (DB) 10, The regional name expression database (DB) 20 and the position information supplementary database (DB) 30, It comprises the name-of-a-person determination database (DB) 40, the morphological-analysis part 50, the regional name expression extraction part 60, the address expression extraction part 70, the exception address expression extraction part 80, and the position information supplementary word extraction part 90.
[0016]The address database 10 comprises the address table 11, the all-prefectures correspondence table 12, the city correspondence table 13, and the county-towns-and-villages correspondence table 14, as shown in drawing 2.
[0017]The all-prefectures field where all-prefectures names, such as "Tokyo", are registered as the address table 11 is shown in drawing 3, It comprises the cities, towns and villages division field where Tokyo's 23 Wards names, such as village names, such as names of towns, such as city names, such as "Yokohama", and "Hayama-cho", and "Hakuba-mura", and "Shinjuku-ku", are registered, the Oaza and the common-name field used as address expression of the low rank of a cities, towns and villages division name, and the character and the ** field which serve as a low-ranking address expression further. And the combination of these four fields can express now all the addresses of the Japan whole country.
[0018]The all-prefectures correspondence table 12 comprises the all-prefectures abbreviation field where the character string by which the "capital", the "way", the "prefecture", and the "prefecture" were abbreviated to the all-prefectures field from the all-prefectures name is registered corresponding to the all-prefectures field, as shown in drawing 4 (a).
[0019]The city correspondence table 13 comprises the city field and the city abbreviation field where the character string to which the "city" was abbreviated is registered from the city name corresponding to the city field, as shown in drawing 4 (b).
[0020]The county-towns-and-villages correspondence table 14 comprises the all-prefectures field, the county-towns-and-villages field where the county name and the towns-and-villages name belonging to the county were registered, and the county abbreviation field where only the towns-and-villages name to which the county name was abbreviated from the county-towns-and-villages name field was registered, as shown in drawing 5. And the all-prefectures field, and the county-towns-and-villages field and the county abbreviation field are matched,

respectively.

[0021]As the regional name expression database 20 is shown in drawing 6 (a), regional name expression which is not address expression of "western part of Japan", the "Kanto district", etc., etc. is registered.

[0022]The position information supplementary word which is expression for the position information supplementary database 30 to supplement with the position information on the "neighborhood", the "neighborhood", an "eastern part", a "southern part", etc. as shown in drawing 6 (b) is registered.

[0023]Expression of the title etc. which are used when describing names of persons, such as "appearance", "**", a "suspect", and a "supervisor", as the name-of-a-person determination database 40 is shown in drawing 6 (c) is registered.

[0024]The morphological-analysis part 50 is decomposed into each morpheme by conducting the morphological analysis of the inputted text using the dictionary in which the character string was registered beforehand. Here, a morpheme is the unit which divided the text for every character string of one or some, and is a character string comparable as a word fundamentally. All the character strings registered into the all-prefectures field which constitutes the address table 11, the cities, towns and villages division field, Oaza and the common-name field, and a character and the ** field are also registered into the dictionary used in a morphological analysis. Although the minimum cost method which is a method of determining the size of the morpheme decomposed is used in the case of a morphological analysis, the minimum in the minimum cost method is set up so that the unit of each character string of eye ** [all prefectures, a cities, towns and villages division, Oaza and a common name, a character, and] may not be subdivided any more. for example, the thing to do for the morphological analysis of the text "the traffic accident occurred in the prefectural road in Kamakura-shi, Kanagawa" -- Kanagawa being "Prefecture", being "Kamakura", "it being alike", "it being able to set", and a "prefectural road" -- "it is " -- two or more morphemes referred to as a "traffic accident", "**", "generating", and "having carried out" are obtained.

[0025]The regional name expression extraction part 60 is performing regional name expression search by comparing each morpheme decomposed by the morphological-analysis part 50 with the character string registered into the regional name expression database 20.

[0026]The address expression extraction part 70 performs exception decision processing which is the judgment of whether a regional name is address expression with each formal morpheme which was not in agreement in the regional name expression extraction part 60, When each morpheme judges with it being a formal address expression, address expression is extracted from the text by comparing each morpheme with the character string registered into the address table 11 of the address database 10. In this exception decision processing, when the "prefectural" character string is not contained in the name of a prefecture, and the "city"

character string is not contained in the city name, it is judged with it not being a formal address expression. When the character string which comes to the next of a name of a prefecture even when a formal all-prefectures name is searched is a character string of the towns-and-villages name of the low rank of a county name, it judges with it being address expression with which the county name display was omitted, and it judges with it not being a formal address expression.

[0027]Each morpheme judged as the exception address expression extraction part 80 not being a formal address expression in the address expression extraction part 70, By searching the character string registered into the all-prefectures abbreviation field of the all-prefectures correspondence table 12, or the city abbreviation field of the city correspondence table 13, required exception address expression where a "prefecture" or a "city" is omitted is extracted, the character string of a "prefecture" or a "city" is compensated, and it changes into a formal address expression.

[0028]When judged with the county name being omitted in the exception decision processing in the regional name expression extraction part 60, the exception address expression extraction part 80, A county name is added to exception address expression which judged the county name omitted by searching the county-towns-and-villages correspondence table 14 based on the all-prefectures name and towns-and-villages name which were searched in exception decision processing and with which the extracted county name was omitted, and it changes into a formal address expression.

[0029]In the case of the character string by which the next morpheme of each morpheme or the morpheme of the 2nd word is registered into the name-of-a-person determination database 40, the exception address expression extraction part 80 judges with it being a name of a person, and is kept from extracting a retrieval object morpheme as address expression to it. It is avoidable to extract accidentally address expression required with "Kamakura" from the character string of the name of a person "Taro Kamakura" in exception address extracting processing by this.

[0030]When the position information on address expression or regional name expression is extracted in the regional name expression extraction part 60, the address expression extraction part 70, and the exception address expression extraction part 80, the position information supplementary word extraction part 90, Retrieval processing of a position information supplementary word is performed by comparing the morpheme which is within the limits of six words from the end of the position information with the character string registered into the position information supplementary database 30. And when a position information supplementary word is extracted, from position information to the searched position information supplementary word is newly outputted as position information.

[0031]Next, operation of the automatic extracting apparatus of the position information on this

embodiment is explained in detail with reference to the flow chart of drawing 7 - drawing 11.
[0032]First, with reference to drawing 7, operation of the whole automatic extracting apparatus of the position information on this embodiment is explained.
[0033]The morphological analysis of the inputted text is conducted in the morphological-analysis part 50, and it is decomposed into each morpheme. (Step 101).
[0034]In the regional name expression extraction part 60, comparison search of each morpheme decomposed in the morphological-analysis part 50 is first carried out with the character string registered into the regional name expression database 20 (Step 102). When judged with the morpheme of a retrieval object being in Step 102, and being regional name expression, the regional name expression extraction part 60 extracts the morpheme as regional name expression (Step 111).
[0035]In Step 102, when judged with the morpheme of a retrieval object not being regional name expression, in order to take out only a formal address expression, it is judged in the address expression extraction part 70 whether it is a formal address expression (Step 103).
[0036]When judged with it being a formal address expression in Step 103, extracting processing of address expression is performed in the address expression extraction part 70 (Step 104).
[0037]When judged with it not being a formal address expression in Step 103, exception address expression extracting processing is performed in the exception address expression extraction part 80 (Step 108), and it is changed into (Step 109) and a formal address expression when exception address expression exists. In exception address extracting processing, when an exception address does not exist, the exception address expression extraction part 80 judges with a retrieval object morpheme not being a thing about position information. And after it is checked whether any processing of texts is completed (Step 107), processing is performed for the following morpheme as a retrieval object morpheme.
[0038]If the extracting processing of the regional name expression in Step 111, the extracting processing of the address expression in Step 104, and the conversion process of a formal address expression of the exception address expression in Step 110 are performed, The position information supplementary word extraction part 90 performs retrieval processing of a position information supplementary word by comparing the morpheme which is within the limits of six words from the end of the extracted position information with the character string registered into the position information supplementary database 30. And when a position information supplementary word is extracted, the position information acquired from position information by making even the searched position information supplementary word into new position information (Step 105) is outputted (Step 106).
[0039]And after it is checked whether any processing of texts is completed (Step 107), processing is performed by making the following morpheme into a retrieval object morpheme.

[0040]Next, the exception decision processing (Step 103) in drawing 7 is explained in more detail using the flow chart of drawing 8.

[0041]In the address expression extracting processing part 70, identity retrieval of each morpheme and the character string registered into the all-prefectures field of the address table 11 is performed, It is referred to as "B", in setting a certain flag to "A" and not being in agreement, while storing the morpheme in the arrangement 1 and shifting one retrieval object morpheme, when in agreement (Step 201).

[0042]Next, the address expression extracting processing part 70 performs identity retrieval of each morpheme and the character string registered into the cities, towns and villages division field of the address table 11 (Step 202). In identity retrieval processing of Step 202, when a retrieval object morpheme and the character string registered into the cities, towns and villages division field are in agreement, it judges with it being a formal address expression, and processing is moved to Step 104.

[0043]In coincidence processing of Step 202, when a retrieval object morpheme and the character string registered into the cities, towns and villages division field are not in agreement, it opts for the next processing based on the processing result in Step 201 (Step 203). In Step 201, when a retrieval object morpheme and a character string are in agreement in identity retrieval with the all-prefectures field, specifically (when a flag is "A"), It judges with it being a formal address expression, and processing is advanced to Step 103, when not in agreement, it judges with it being exception address expression, and processing is advanced to Step 108 (Step 203). (when a flag is "B")

[0044]Next, the extracting processing (Step 104) of the address expression in drawing 7 is explained in more detail using the flow chart of drawing 9.

[0045]First, as for the address expression extraction part 70, the present retrieval object morpheme is 1 face ** (Step 504) about a retrieval object morpheme, after storing the morpheme in the arrangement 2, since it is a cities, towns and villages name in a formal address expression.

[0046]And identity retrieval of a retrieval object morpheme, and the Oaza and the common-name field of the address table 11 is performed (Step 505), when in agreement, a retrieval object morpheme is stored in the arrangement 3, and it is 1 face ** (Step 506) about a retrieval object morpheme. Identity retrieval of a retrieval object morpheme, and the character and the ** field of the address table 11 is performed (Step 507), similarly, when in agreement, a retrieval object morpheme is stored in the arrangement 4, and it is 1 face ** (Step 508) about a retrieval object morpheme.

[0047]In one of the identity retrieval of Step 505 or Step 507, when not in agreement, extracting processing of an address item is carried out to the next of processing of Step 508 (Step 509). In the extracting processing of an address item, when the next morpheme of the

morpheme of address expression is a number, the number is considered as address item expression. In all the rows of number-morphological-number or number-morphological-number-morphological-number **, the morpheme of address expression considers only a number as address item expression.

[0048]finally, the address expression extraction part 70 receives each morpheme and the extracted address item which are stored in the arrangement 1-4 -- it connects and address expression extracting processing is ended as one address expression (Step 510). Only the character string stored in the all-prefectures field or the cities, towns and villages division field can serve as a head of address expression by this address expression extracting processing, and the character string stored in other fields cannot become a head of address expression.

[0049]Next, the exception address expression extracting processing (Step 108) in drawing 7 and exception address expression existence decision processing (Step 109) are explained in more detail using the flow chart of drawing 10.

[0050]According to this embodiment, address expression with which the character of the "prefecture" or the "city" is omitted, and address expression with which the county name is omitted are processed as exception address expression. In the exception address expression extraction part 80, address expression judged to be exception address expression judges which type it is in Step 401, In being exception address expression with which the character of the "prefecture" or the "city" is omitted, it processes Steps 301-304, and in being exception address expression with which the county name is omitted, it processes Steps 402 and 403.

[0051]The extracting processing of exception address expression with which the character of the "prefecture" or the "city" is omitted first is explained.

[0052]The exception address expression extraction part 80 performs meaning search with a retrieval object morpheme and the all-prefectures abbreviation field of the all-prefectures correspondence table 12 (Step 301). In Step 301, when not in agreement, meaning search with a retrieval object morpheme and the city abbreviation field of the city correspondence table 13 is performed (Step 302).

[0053]When in agreement in Step 301, and when in agreement in Step 302, the retrieval object morpheme is stored in arrangement, and it is 1 face ** (Step 303) about a retrieval object morpheme. And name-of-a-person decision processing which performs a retrieval object morpheme and the following morpheme, and identity retrieval with the name-of-a-person determination database 40 is performed (Step 304). being judged with it being exception address expression, when judged with it not being name-of-a-person expression in the name-of-a-person decision processing in Step 304 -- processing -- Step 110 -- advancing .

[0054]The all-prefectures abbreviation Field **** abbreviation field and formal address expression are matched, if judged with address expression by the above-mentioned search, a "prefecture" or a "city" will be given and extracted exception address expression will be

changed into a formal address (Step 110).

[0055]When judged with it being name-of-a-person expression in the name-of-a-person decision processing in Step 304, and when not in agreement in Step 302, it judges that the character string expressed by the retrieval object morpheme is not formal address expression or exception address expression, either, and it advances processing to that of Step 107.

[0056]By this processing, the exception address expression "Kanagawa" and "Yokohama" is changed into the formal address expression "Kanagawa Prefecture" and "Yokohama", respectively, for example.

[0057]Next, the extracting processing of exception address expression with which the county name is omitted is explained.

[0058]First, the exception address extraction part 80 performs processing compensated with the county name omitted using the county-towns-and-villages corresponding field 14. In this case, the exception address extraction part 80 narrows down a retrieving range from the all-prefectures name information retrieved at Step 201, Identity retrieval of the towns-and-villages names congruous in Step 202 and the character string registered into the group abbreviation field is performed, Processing compensated with the county name omitted by considering it as address expression which had the character string of the county-towns-and-villages field corresponding to the towns-and-villages name searched instead of the congruous towns-and-villages names is performed (Step 402). And the exception address extraction part 80 shifts one retrieval object morpheme (Step 403).

[0059]The processing in this step 402 is concretely explained using the case where the all-prefectures name searched in Step 201 is "Nagano Prefecture", and the towns-and-villages name searched in Step 202 is "Hakuba-mura." First, the exception address extraction part 80 performs identity retrieval of the character string and "Hakuba-mura" which were registered into the group abbreviation field whose all-prefectures field is "Nagano Prefecture." And the character string of "Hakuba-mura" is transposed to the character string "Hakuba-mura, Kita-Azumi-gun" of the county-towns-and-villages field corresponding to the congruous county abbreviation fields. "Nagano Hakuba-mura" which is exception address expression is transposed to "Hakuba-mura, Kita-Azumi-gun, Nagano-ken" by this processing.

[0060]Finally, the flow chart of drawing 11 is used and the extracting processing (Step 105) of the position information supplementary word in drawing 7 is explained in more detail.

[0061]The position information supplementary word extraction part 90 performs identity retrieval of the morpheme which is in the range of less than six words from the end of address expression or regional name expression, and the position information supplementary database 30 (Step 702). In Step 702, when the congruous character strings exist, even the morpheme extracted from extracted address expression or regional name expression in Step 702 is extracted as one position information (Step 703). In Step 702, when the congruous character

strings do not exist, the position information supplementary word extraction part 90 returns a retrieval object morpheme six words ago, and ends position information supplementary word extracting processing.

[0062]The place which performed automatic extracting of the position information included in the homepage (300 pages) in a newspaper article (1000) report and the Internet using the automatic extracting apparatus in this embodiment, The rate of automatic extracting in a homepage which was able to obtain 95.2% in the newspaper article and was able to obtain the rate of automatic extracting of 80.1% in the homepage fell rather than the newspaper article at the homepage because the character may be given not as text but as picture information.

[0063]Thus, in the automatic extracting apparatus of the position information on this embodiment, while being able to extract automatically position information included in the text, such as a formal address expression, address expression which is not formal, and regional name expression, with high probability, position information including a position information supplementary word can be extracted.

[0064]Although not shown in a figure, a data processing device (computer), memory storage, an input/output processor, and the recording medium that recorded the program for performing the automatic extracting method can constitute the automatic extracting apparatus of this embodiment. This recording medium may be a recording medium of a magnetic disk, semiconductor memory, or others. This program is read into a data processing device from a recording medium, controls operation of a data processing device, and performs processing performed by the morphological-analysis part 50 in drawing 1, the regional name expression extraction part 60, the address expression extraction part 70, the exception address expression extraction part 80, and the position information supplementary word extraction part 90. And memory storage is constituted by the address database 10, the regional name expression database 20, the position information supplementary database 30, and the name-of-a-person determination database 40, and an input/output device outputs the position information extracted from the input and text of text data for extracting position information.

Preparation ******.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a block diagram showing the composition of the automatic extracting apparatus of the position information on one embodiment of this invention.
[Drawing 2]It is a figure showing the data structure of the address database 10 in drawing 1.
[Drawing 3]It is a figure showing the data structure of the address table 11 in drawing 2.
[Drawing 4]It is a figure (drawing 4 (b)) showing the data structure of the figure (drawing 4 (a)) showing the data structure of the all-prefectures correspondence table 12 in drawing 2, and the city correspondence table 13.
[Drawing 5]It is a figure showing the data structure of the county-towns-and-villages correspondence table 14 in drawing 2.
[Drawing 6]It is a figure (drawing 6 (c)) showing the structure of the figure (drawing 6 (a)) showing the data structure of the regional name expression database 20 in drawing 1, the figure (drawing 6 (b)) showing the data structure of the position information supplementary database 30, and the name-of-a-person determination database 40.
[Drawing 7]It is a flow chart which shows operation of the automatic extracting apparatus of the position information on drawing 1.
[Drawing 8]It is the flow chart which showed the exception decision processing (Step 103) in drawing 7 in more detail.
[Drawing 9]It is the flow chart which showed the address expression extracting processing (Step 104) in drawing 7 in more detail.
[Drawing 10]It is the flow chart which showed the exception address expression extracting processing (Step 108) in drawing 7, and exception address expression existence decision processing (Step 109) in more detail.
[Drawing 11]It is the flow chart shown in the extracting processings (Step 105) of the position information supplementary word in drawing 7 in detail.

[Description of Notations]
10 Address database (DB)
11 Address table
12 All-prefectures correspondence table
13 City correspondence table
14 Group-towns-and-villages correspondence table
20 Regional name expression database (DB)
30 Position information supplementary database (DB)
40 Name-of-a-person determination database (DB)
50 Morphological-analysis part
60 Regional name expression extraction part
70 Address expression extraction part
80 Exception address expression extraction part
90 Position information supplementary word extraction part
101-111 Step
201-203 Step
301-304 Step
401-403 Step
504-510 Step
702-704 Step

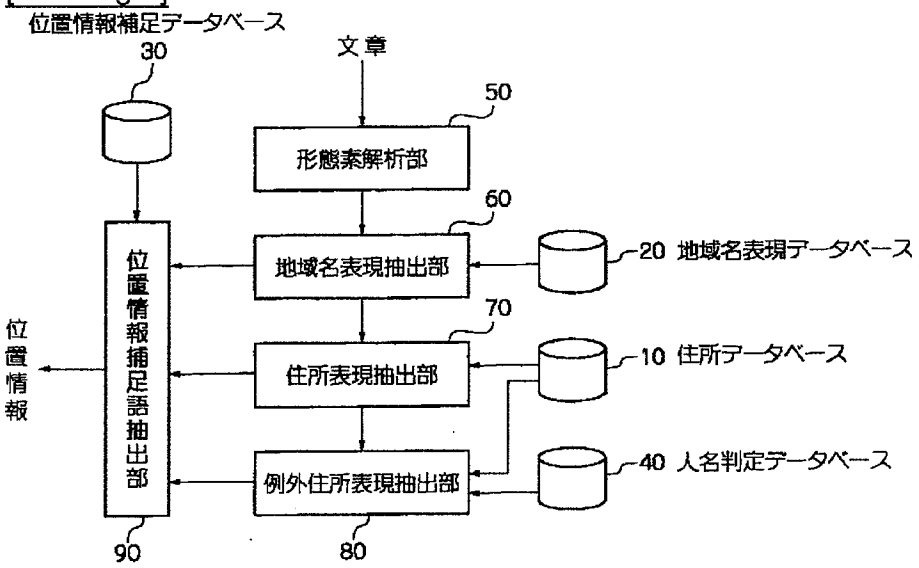
[Translation done.]

* NOTICES *

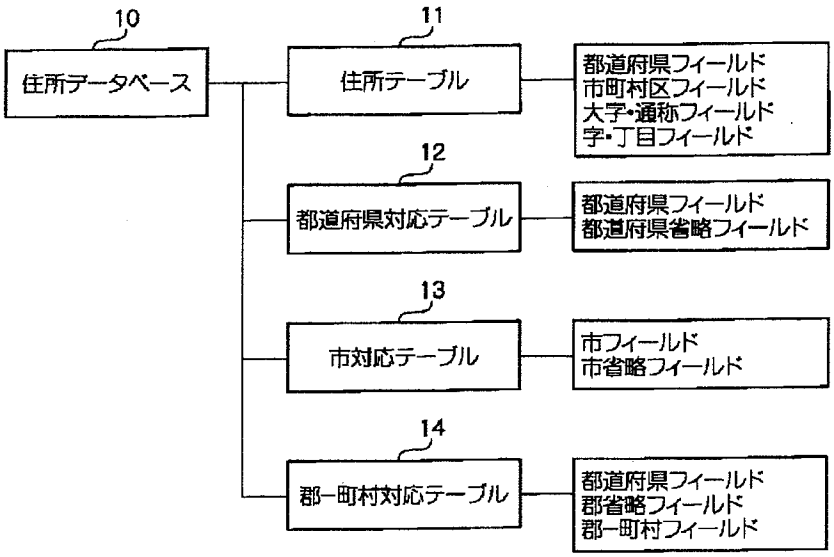
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

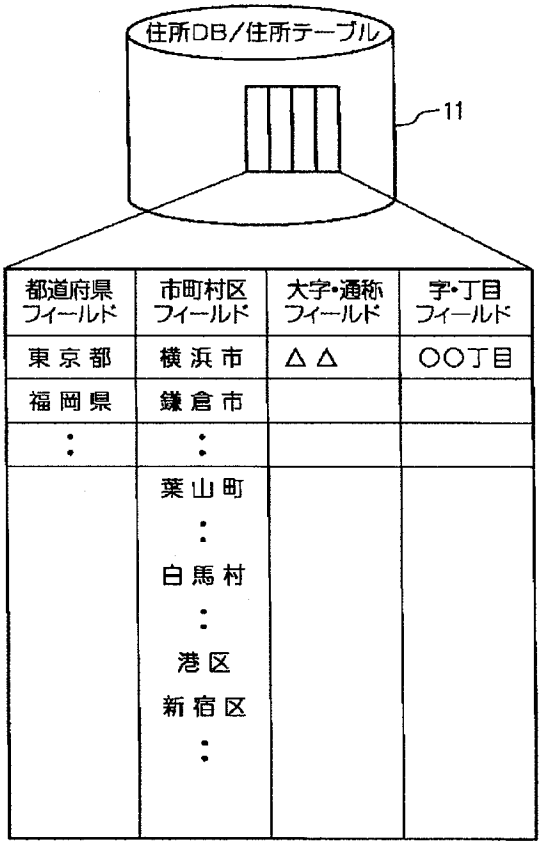
[Drawing 1]



[Drawing 2]



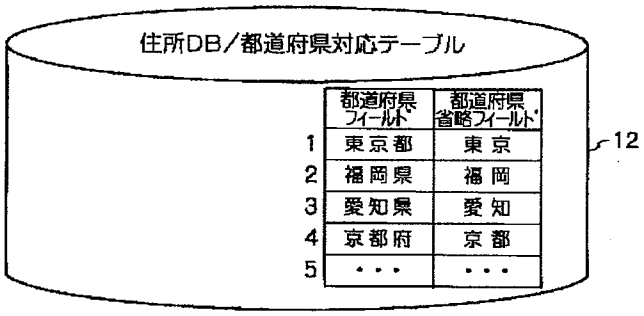
[Drawing 3]



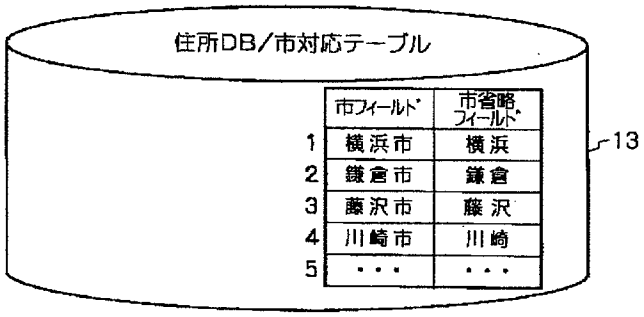
都道府県フィールド	市町村区フィールド	大字・通称フィールド	字・丁目フィールド
東京都	横浜市	△△	○○丁目
福岡県	鎌倉市		
：	：		
	葉山町		
	：		
	白馬村		
	：		
	港区		
	新宿区		
	：		

[Drawing 4]

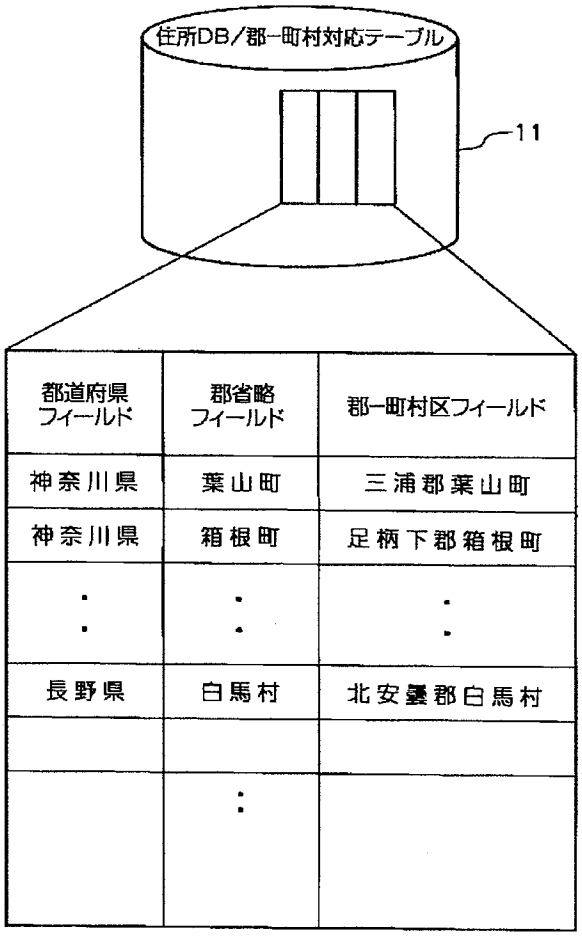
(a)


住所DB/都道府県対応テーブル
都道府県フィールド
都道府県省略フィールド
1 東京都 東京
2 福岡県 福岡
3 愛知県 愛知
4 京都府 京都
5 ・・・ ・・・
12

(b)

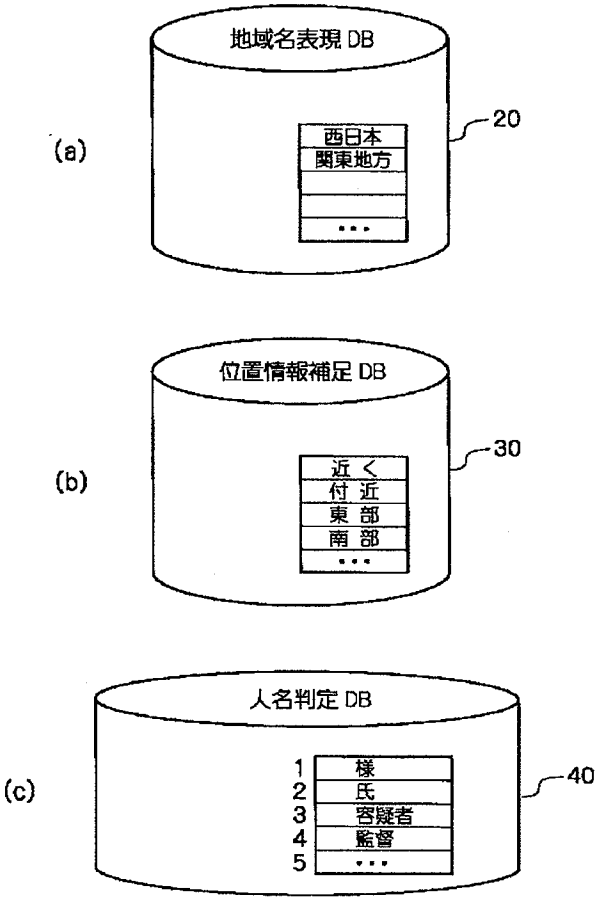

住所DB/市対応テーブル
市フィールド
市省略フィールド
1 横浜市 横浜
2 鎌倉市 鎌倉
3 藤沢市 藤沢
4 川崎市 川崎
5 ・・・ ・・・
13

[Drawing 5]

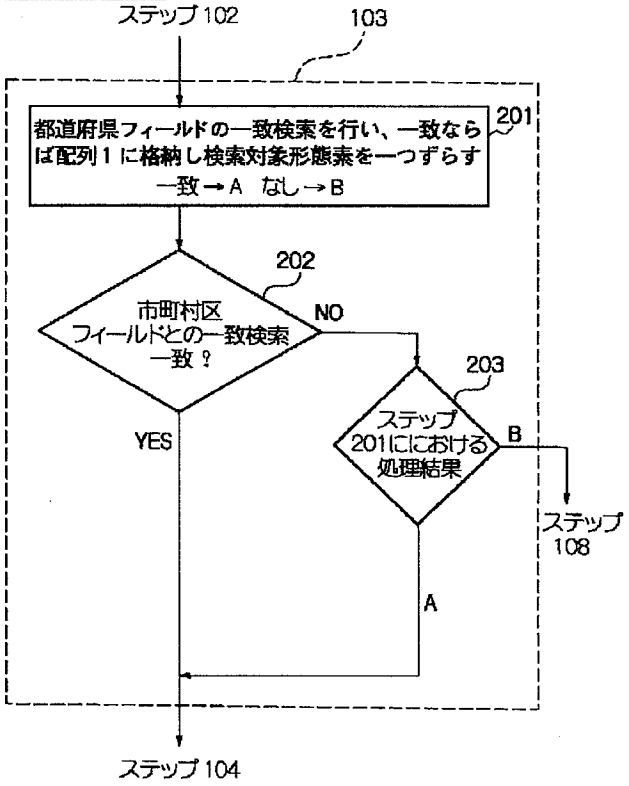


都道府県フィールド	郡省略フィールド	郡一町村区フィールド
神奈川県	葉山町	三浦郡葉山町
神奈川県	箱根町	足柄下郡箱根町
・ ・	・ ・	・ ・
長野県	白馬村	北安曇郡白馬村
	・	

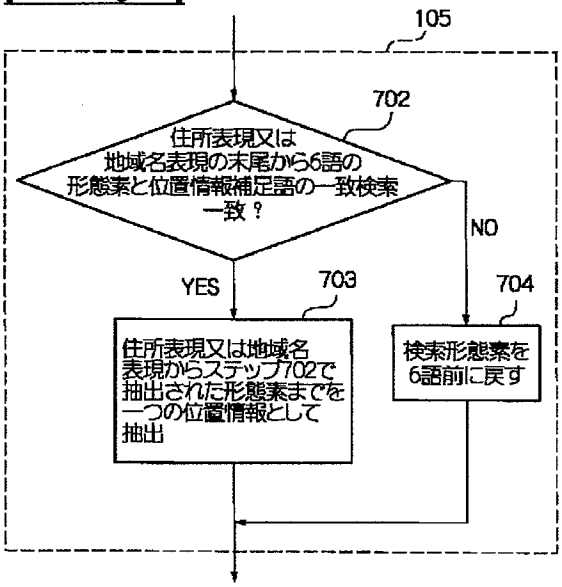
[Drawing 6]


(a)
地域名表現 DB
西日本
関東地方
...
20
(b)
位置情報補足 DB
近く
付近
東部
南部
...
30
(c)
人名判定 DB
1 様
2 氏
3 容疑者
4 監督
5 ...
40

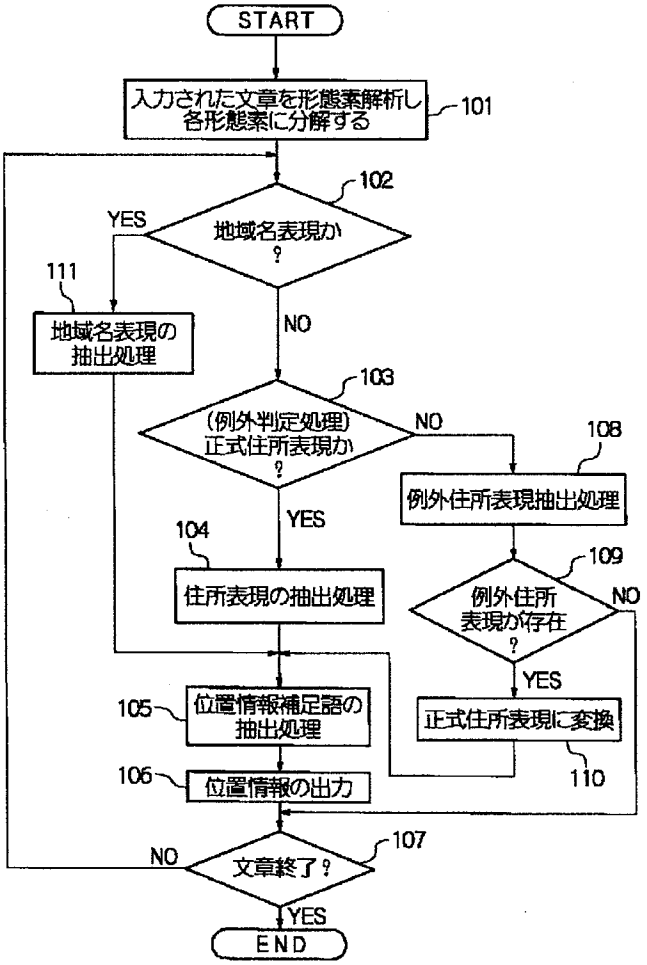
[Drawing 8]


ステップ 102
103
都道府県フィールドの一致検索を行い、一致ならば配列1に格納し検索対象形態素を一つずらす
一致→A なし→B
201
202
市町村区フィールドとの一致検索一致？
NO
YES
203
ステップ201ににおける処理結果
B
ステップ108
A
ステップ 104

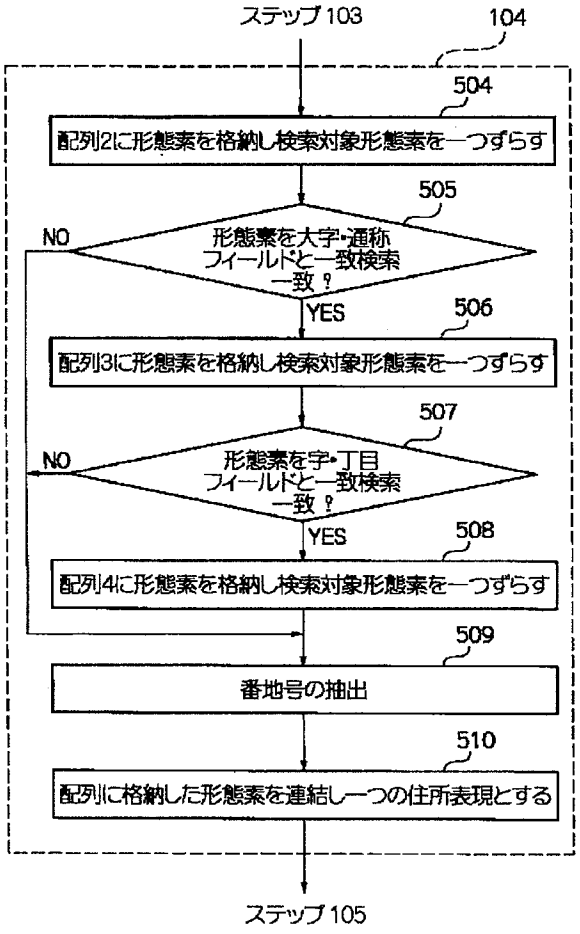
[Drawing 11]



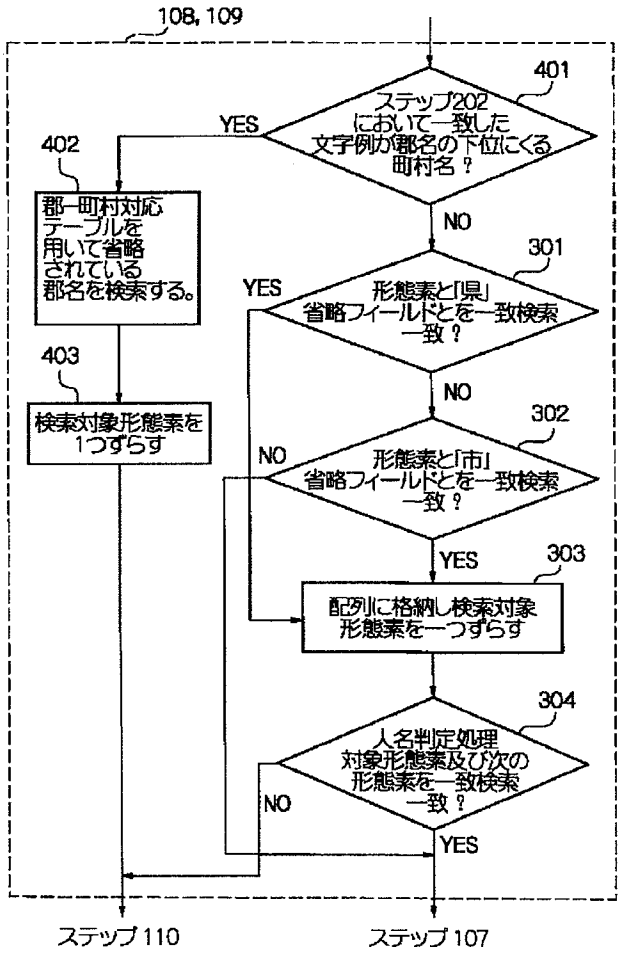
[Drawing 7]



[Drawing 9]


ステップ103
104
504
配列2に形態素を格納し検索対象形態素を一つずらす
505
NO
形態素を大字・通称フィールドと一致検索一致？
YES
506
配列3に形態素を格納し検索対象形態素を一つずらす
507
NO
形態素を字・丁目フィールドと一致検索一致？
YES
508
配列4に形態素を格納し検索対象形態素を一つずらす
509
番地号の抽出
510
配列に格納した形態素を連結し一つの住所表現とする
ステップ105

[Drawing 10]



[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2000－250931
（P2000－250931A）
(43)公開日　平成12年９月14日(2000.9.14)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/30		Ｃ０６Ｆ　15/401	３１０Ａ	５Ｂ００９
17/27		15/20	５５０Ｚ	５Ｂ０７５
		15/40	３７０Ａ	
		15/413	３１０Ｂ	

審査請求　未請求　請求項の数18　ＯＬ　（全 15 頁）

(21)出願番号　特願平11－53137

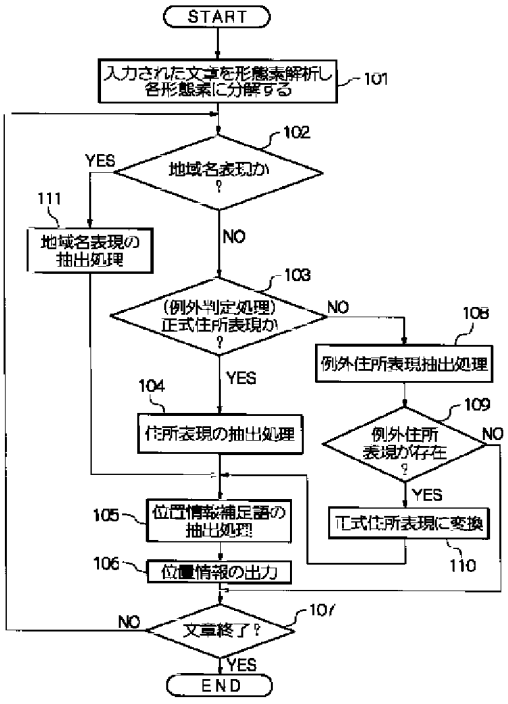
(22)出願日　平成11年３月１日(1999.3.1)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　杉浦　寛宣
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(72)発明者　土屋　秀幸
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(74)代理人　100088328
弁理士　金田　暢之

Ｆターム(参考)　5B009 QA12
5B075 ND03 NK02 NK32 NK49 NR06
QM02 UU06

(54)【発明の名称】　位置情報の自動抽出装置および自動抽出方法と記録媒体

(57)【要約】
【課題】　文章中に含まれている住所表現または地域名表現の位置情報を高い確率で自動抽出する。
【解決手段】　入力された文章を形態素に分解し（ステップ１０１）、各形態素と地域名表現と比較する（ステップ１０２）。各形態素を含む表現が、正式な住所表現であるかを判定する例外判定処理を行ない（ステップ１０３）、正式な住所表現の場合、形態素と日本全国の住所表現との比較を順次行ない住所表現を抽出する（ステップ１０４）。例外判定処理において正式な住所表現ではない場合に、各形態素と予め登録された例外住所表現との一致検索を行ない例外住所表現を抽出し、それらに「県」、「市」、郡名を追加して正式な住所表現に変換する（ステップ１０８～１１０）。抽出された位置情報から６語以内に位置情報補足語がある場合にはそれを含めて位置情報として出力する（ステップ１０６）。



【特許請求の範囲】
【請求項１】　入力された文章に含まれている位置情報を抽出するための位置情報の自動抽出方法であって、
入力された文章を１または複数の文字列単位で区切ることにより複数の形態素に分解する処理と、
前記各形態素のうちの１つの形態素を検索対象形態素とし、該検索対象形態素が、予め登録された地域名表現の文字列と一致するかどうかを判定する処理と、
前記検索対象形態素を含む表現が、正式な住所表現であるかとうかを判定する例外判定処理と、
前記例外判定処理において前記検索対象形態素を含む表現が正式な住所表現であると判定された場合に、前記検索対象形態素と予め登録された日本全国の住所表現の文字列との一致検索を順次行なうことにより前記文章から住所表現を抽出する住所抽出処理と、
前記例外判定処理において前記検索対象形態素を含む表現が正式な住所表現ではないと判定された場合に、前記検索対象形態素と予め登録された例外住所表現の文字列との一致検索を行うことにより例外住所表現を抽出し、抽出された該例外住所表現に省略された「県」または「市」の文字列若しくは省略された郡名を追加して正式な住所表現に変換する例外住所抽出処理と、
抽出された地域名表現および住所表現の末尾から一定数以内の各形態素に、予め登録された位置情報補足語と同一の文字列が存在するかどうか検索し、存在する場合には地域名表現または住所表現から該位置情報補足語までを位置情報とする位置情報補足語抽出処理とから構成される位置情報の自動抽出方法。
【請求項２】　前記例外判定処理が、
前記検索対象形態素と予め登録された都道府県名の文字列との一致検索を行う処理と、
前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名と一体した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、
前記検索対象形態素と予め登録された市町村区名の文字列との一致検索を行う処理と、
前記市町村区名との一致検索を行う処理において前記検索対象形態素が市町村区名の文字列と一致した場合に、前記検索対象形態素を含む表現は正式な住所表現であると判定する処理と、
前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名の文字列と一致せず、かつ前記市町村区名との一致検索を行う処理において検索対象形態素が市町村区名の文字列と一致しなかった場合に、前記検索対象形態素を含む表現は正式な住所表現ではないと判定する処理とから構成される請求項１記載の位置情報の自動抽出方法。
【請求項３】　前記住所抽出処理が、
前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、
前記検索対象形態素と、予め登録された大字・通称名の文字列との一致検索を行う処理と、
前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、
前記検索対象形態素と、予め登録された字・丁目の文字列との一致検索を行う処理と、
前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、
前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致しなかった場合、または前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致しなかった場合、または前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列とも一致し、かつ前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列とも一致した場合に、検索対象形態素の次の形態素が数字であるかどうかを判定し、数字である場合に該数字を住所表現における番地号であるとして抽出する番地号抽出処理と、
抽出された都道府県名の形態素、市町村名の形態素、大字・通称の形態素、字・丁目の形態素、番地号の形態素を連結して１つの住所表現とする処理とから構成される請求項２記載の位置情報の自動抽出方法。
【請求項４】　前記例外住所抽出処理が、
前記住所抽出処理において正式な住所表現でないと判定された表現が、正式な住所表現から郡名が省略された例外住所表現であるかどうかを判定する例外住所タイプ判定処理と、
前記例外住所タイプ判定処理において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現であると判定された場合に、前記検索対象形態素と、予め登録された、都道府県名から「県」が省略された文字列とを一致検索する処理と、
前記検索対象形態素と、予め登録された、市名から「市」が省略された文字列とを一致検索する処理と、
前記「市」が省略された文字列との一致検索において前記検索対象形態素が「市」が省略された文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、
前記検索対象形態素と、予め登録された人名を記述する際に使用される表現の文字列との一致検索する人名判定処理と、
前記人名判定処理において検索対象形態素が、人名を記述する際に使用される表現の文字列と一致した場合およ

び検査対象形態素が「県」が省略された文字列とも「市」が省略された文字列とも一致しなかった場合に、検査対象形態素を含む表現は、住所表現ではないと判定する処理とを有する請求項１から３のいずれか１項記載の位置情報の自動抽出方法。

【請求項５】　前記例外住所抽出処理が、

前記例外住所タイプ判定処理において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現でないと判定された場合に、都道府県名の文字列との一致検索において検索された都道府県名に基づいて、市町村区名との一致検索において一致した町村名と、予め登録された正式な郡－町村名から群名のみを省略した文字列との一致検索を行ない、一致した町村名の代わりにその町村名に対応する正式な郡－町村名の文字列を検索された住所表現とすることにより省略された郡名を補う処理と、

前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理とをさらに有する請求項４記載の位置情報の自動抽出方法。

【請求項６】　前記位置情報補足語抽出処理が、

住所表現又は地域名表現の末尾から一定の範囲内にある形態素と、予め登録された位置情報補足語との一致検索を行う処理と、

前記位置情報補足語との一致検索処理において一致した場合に、前記住所表現または地域名表現から一致した形態素までを１つの位置情報として抽出する処理と、

前記位置情報補足語との一致検索処理において一致しなかった場合に、検索対象形態素を６語前に戻す処理とから構成される請求項１から５のいずれか１項記載の位置情報の自動抽出方法。

【請求項７】　入力された文章に含まれている位置情報を抽出するための自動抽出処理をコンピュータに実行させるためのプログラムを記録した記録媒体であって、

入力された文章を１または複数の文字列単位で区切ることにより複数の形態素に分解する処理と、

前記各形態素のうちの１つの形態素を検索対象形態素とし、該検索対象形態素が、予め登録された地域名表現の文字列と一致するかどうかを判定する処理と、

前記検索対象形態素を含む表現が、正式な住所表現であるかとうかを判定する例外判定処理と、

前記例外判定処理において前記検索対象形態素を含む表現が正式な住所表現であると判定された場合に、前記検索対象形態素と予め登録された日本全国の住所表現の文字列との一致検索を順次行なうことにより前記文章から住所表現を抽出する住所抽出処理と、

前記例外判定処理において前記検索対象形態素を含む表現が正式な住所表現ではないと判定された場合に、前記検索対象形態素と予め登録された例外住所表現の文字列との一致検索を行うことにより例外住所表現を抽出し、抽出された該例外住所表現に省略された「県」または「市」の文字列若しくは省略された郡名を追加して正式な住所表現に変換する例外住所抽出処理と、

抽出された地域名表現および住所表現の末尾から一定数以内の各形態素に、予め登録された位置情報補足語と同一の文字列が存在するかどうか検索し、存在する場合には地域名表現または住所表現から該位置情報補足語までを位置情報とする位置情報補足語抽出処理とをコンピュータに実行させるためのプログラムを記録した記録媒体。

【請求項８】　前記例外判定処理が、

前記検索対象形態素と予め登録された都道府県名の文字列との一致検索を行う処理と、

前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名と一体した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、

前記検索対象形態素と予め登録された市町村区名の文字列との一致検索を行う処理と、

前記市町村区名との一致検索を行う処理において前記検索対象形態素が市町村区名の文字列と一致した場合に、前記検索対象形態素を含む表現は正式な住所表現であると判定する処理と、

前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名の文字列と一致せず、かつ前記市町村区名との一致検索を行う処理において検索対象形態素が市町村区名の文字列と一致しなかった場合に、前記検索対象形態素を含む表現は正式な住所表現ではないと判定する処理とから構成される請求項７記載の記録媒体。

【請求項９】　前記住所抽出処理が、

前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、

前記検索対象形態素と、予め登録された大字・通称名の文字列との一致検索を行う処理と、

前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、

前記検索対象形態素と、予め登録された字・丁目の文字列との一致検索を行う処理と、

前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、

前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致しなかった場合、または前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致しなかった場合、または前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列

とも一致し、かつ前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列とも一致した場合に、検索対象形態素の次の形態素が数字であるかどうかを判定し、数字である場合に該数字を住所表現における番地号であるとして抽出する番地号抽出処理と、
抽出された都道府県名の形態素、市町村名の形態素、大字・通称の形態素、字・丁目の形態素、番地号の形態素を連結して１つの住所表現とする処理とから構成される請求項８記載の記録媒体。
【請求項１０】　前記例外住所抽出処理が、
前記住所抽出処理において正式な住所表現でないと判定された表現が、正式な住所表現から郡名が省略された例外住所表現であるかどうかを判定する例外住所タイプ判定処理と、
前記例外住所タイプ判定処理において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現であると判定された場合に、前記検索対象形態素と、予め登録された、都道府県名から「県」が省略された文字列とを一致検索する処理と、
前記検索対象形態素と、予め登録された、市名から「市」が省略された文字列とを一致検索する処理と、
前記「市」が省略された文字列との一致検索において前記検索対象形態素が「市」が省略された文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、
前記検索対象形態素と、予め登録された人名を記述する際に使用される表現の文字列との一致検索する人名判定処理と、
前記人名判定処理において検索対象形態素が、人名を記述する際に使用される表現の文字列と一致した場合および検査対象形態素が「県」が省略された文字列とも「市」が省略された文字列とも一致しなかった場合に、検査対象形態素を含む表現は、住所表現ではないと判定する処理とを有する請求項７から９のいずれか１項記載の記録媒体。
【請求項１１】　前記例外住所抽出処理が、
前記例外住所タイプ判定処理において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現でないと判定された場合に、都道府県名の文字列との一致検索において検索された都道府県名に基づいて、市町村区名との一致検索において一致した町村名と、予め登録された正式な郡－町村名から群名のみを省略した文字列との一致検索を行ない、一致した町村名の代わりにその町村名に対応する正式な郡－町村名の文字列を検索された住所表現とすることにより省略された郡名を補う処理と、
前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理とをさらに有する請求項１０記載の記録媒体。
【請求項１２】　前記位置情報補足語抽出処理が、
住所表現又は地域名表現の末尾から一定の範囲内にある形態素と、予め登録された位置情報補足語との一致検索を行う処理と、
前記位置情報補足語との一致検索処理において一致した場合に、前記住所表現または地域名表現から一致した形態素までを１つの位置情報として抽出する処理と、
前記位置情報補足語との一致検索処理において一致しなかった場合に、検索対象形態素を６語前に戻す処理とから構成される請求項７から１１のいずれか１項記載の記録媒体。
【請求項１３】　入力された文章に含まれている位置情報を抽出するための自動抽出装置であって、
入力された文章を１または複数の文字列単位で区切ることにより複数の形態素に分解する形態素解析手段と、
前記各形態素のうちの１つの形態素を検索対象形態素とし、該検索対象形態素が、予め登録された地域名表現の文字列と一致するかどうかを判定する地域名表現抽出手段と、
前記検索対象形態素を含む表現が、正式な住所表現であるかとうかを判定し、前記検索対象形態素を含む表現が正式な住所表現であると判定された場合に、前記検索対象形態素と予め登録された日本全国の住所表現の文字列との一致検索を順次行なうことにより前記文章から住所表現を抽出する住所抽出手段と、
前記住所表現抽出手段において前記検索対象形態素を含む表現が正式な住所表現ではないと判定された場合に、前記検索対象形態素と予め登録された例外住所表現の文字列との一致検索を行うことにより例外住所表現を抽出し、抽出された該例外住所表現に省略された「県」または「市」の文字列若しくは省略された郡名を追加して正式な住所表現に変換する例外住所抽出手段と、
抽出された地域名表現および住所表現の末尾から一定数以内の各形態素に、予め登録された位置情報補足語と同一の文字列が存在するかどうか検索し、存在する場合には地域名表現または住所表現から該位置情報補足語までを位置情報とする位置情報補足語抽出手段とから構成される位置情報の自動抽出装置。
【請求項１４】　前記例外判定手段が、
前記検索対象形態素と予め登録された都道府県名の文字列との一致検索を行う手段と、
前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名と一体した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする手段と、
前記検索対象形態素と予め登録された市町村区名の文字列との一致検索を行う手段と、
前記市町村区名との一致検索を行う手段において前記検索対象形態素が市町村区名の文字列と一致した場合に、前記検索対象形態素を含む表現は正式な住所表現である

と判定する手段と、
前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名の文字列と一致せず、かつ前記市町村区名との一致検索を行う手段において検索対象形態素が市町村区名の文字列と一致しなかった場合に、前記検索対象形態素を含む表現は正式な住所表現ではないと判定する手段とから構成される請求項１３記載の位置情報の自動抽出装置。
【請求項１５】　前記住所抽出手段が、
前記検索対象形態素の次の形態素を新たな検索対象形態素とする手段と、
前記検索対象形態素と、予め登録された大字・通称名の文字列との一致検索を行う手段と、
前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする手段と、
前記検索対象形態素と、予め登録された字・丁目の文字列との一致検索を行う手段と、
前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする手段と、
前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致しなかった場合、または前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致しなかった場合、または前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列とも一致し、かつ前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列とも一致した場合に、検索対象形態素の次の形態素が数字であるかどうかを判定し、数字である場合に該数字を住所表現における番地号であるとして抽出する番地号抽出手段と、
抽出された都道府県名の形態素、市町村名の形態素、大字・通称の形態素、字・丁目の形態素、番地号の形態素を連結して１つの住所表現とする手段とから構成される請求項１４記載の位置情報の自動抽出装置。
【請求項１６】　前記例外住所抽出手段が、
前記住所抽出手段において正式な住所表現でないと判定された表現が、正式な住所表現から郡名が省略された例外住所表現であるかどうかを判定する例外住所タイプ判定手段と、
前記例外住所タイプ判定手段において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現であると判定された場合に、前記検索対象形態素と、予め登録された、都道府県名から「県」が省略された文字列とを一致検索する手段と、
前記検索対象形態素と、予め登録された、市名から「市」が省略された文字列とを一致検索する手段と、
前記「市」が省略された文字列との一致検索において前記検索対象形態素が「市」が省略された文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする手段と、
前記検索対象形態素と、予め登録された人名を記述する際に使用される表現の文字列との一致検索する人名判定手段と、
前記人名判定手段において検索対象形態素が、人名を記述する際に使用される表現の文字列と一致した場合および検査対象形態素が「県」が省略された文字列とも「市」が省略された文字列とも一致しなかった場合に、検査対象形態素を含む表現は、住所表現ではないと判定する手段とを有する請求項１３から１５のいずれか１項記載の位置情報の自動抽出装置。
【請求項１７】　前記例外住所抽出手段が、
前記例外住所タイプ判定手段において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現でないと判定された場合に、都道府県名の文字列との一致検索において検索された都道府県名に基づいて、市町村区名との一致検索において一致した町村名と、予め登録された正式な郡－町村名から群名のみを省略した文字列との一致検索を行ない、一致した町村名の代わりにその町村名に対応する正式な郡－町村名の文字列を検索された住所表現とすることにより省略された郡名を補う手段と、
前記検索対象形態素の次の形態素を新たな検索対象形態素とする手段とをさらに有する請求項１６記載の位置情報の自動抽出装置。
【請求項１８】　前記位置情報補足語抽出手段が、
住所表現又は地域名表現の末尾から一定の範囲内にある形態素と、予め登録された位置情報補足語との一致検索を行う手段と、
前記位置情報補足語との一致検索手段において一致した場合に、前記住所表現または地域名表現から一致した形態素までを１つの位置情報として抽出する手段と、
前記位置情報補足語との一致検索手段において一致しなかった場合に、検索対象形態素を６語前に戻す手段とから構成される請求項１３から１７のいずれか１項記載の位置情報の自動抽出装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】電子化された文字情報により構成された文章中に含まれている住所表現または地域名表現等の位置情報を自動的に抽出するための自動抽出方法および装置に関する。
【０００２】
【従来の技術】従来は、文章中に含まれている位置情報を検索する方法としては、利用者により指定された位置情報を示す文字列が文章中に含まれているかどうか文章

の全文を検索する方法があった。この方法では利用者が指定する位置情報を探すために文章の全文を検索することになる。しかしこの方法では、対象とする文章量の増加に伴い検索時間も大幅に増加してしまう。そのため、文章中に含まれている住所表現又は地域名表現等の位置情報を予め抽出しておき、検索する際には抽出された位置情報のみを調べることにより、検索時間を短縮することが求められている。

【０００３】しかし、位置情報には様々な記述方法があるため、文章中における位置情報と他の文字列との区別を自動的に行い、位置情報のみを抽出するのは容易ではない。例えば、文章中には「神奈川県鎌倉市」のような正式な住所表現である位置情報のみではなく、「鎌倉では、・・」のように県名や市名が省略されている場合や、「鎌倉太郎」等の住所表現の一部を含んだ人名である場合等がある。また、正式な住所表現が、「長野県北安曇郡白馬村」である場合でも、「長野県白馬村」のように郡名表現が省略される場合がある。さらに「東日本」、「関東地方」のような住所ではない地域名表現が位置情報として用いられるている場合もある。

【０００４】

【発明が解決しようとする課題】従来技術では、文章中に含まれている様々な表現の位置情報を抽出することができないという問題があった。

【０００５】本発明は、文章中に含まれている位置情報を高い確率で抽出することができる位置情報の自動抽出装置および方法を提供することを目的とする。

【０００６】

【課題を解決するための手段】上記目的を達成するために、本発明の位置情報の自動抽出方法は、入力された文章に含まれている位置情報を抽出するための位置情報の自動抽出方法であって、入力された文章を１または複数の文字列単位で区切ることにより複数の形態素に分解する処理と、前記各形態素のうちの１つの形態素を検索対象形態素とし、該検索対象形態素が、予め登録された地域名表現の文字列と一致するかどうかを判定する処理と、前記検索対象形態素を含む表現が、正式な住所表現であるかとうかを判定する例外判定処理と、前記例外判定処理において前記検索対象形態素を含む表現が正式な住所表現であると判定された場合に、前記検索対象形態素と予め登録された日本全国の住所表現の文字列との一致検索を順次行なうことにより前記文章から住所表現を抽出する住所抽出処理と、前記例外判定処理において前記検索対象形態素を含む表現が正式な住所表現ではないと判定された場合に、前記検索対象形態素と予め登録された例外住所表現の文字列との一致検索を行うことにより例外住所表現を抽出し、抽出された該例外住所表現に省略された「県」または「市」の文字列若しくは省略された郡名を追加して正式な住所表現に変換する例外住所抽出処理と、抽出された地域名表現および住所表現の末尾から一定数以内の各形態素に、予め登録された位置情報補足語と同一の文字列が存在するかどうか検索し、存在する場合には地域名表現または住所表現から該位置情報補足語までを位置情報とする位置情報補足語抽出処理とから構成される。

【０００７】本発明は、入力された文章を形態素に分解し、各形態素と予め登録された、地域名、または都道府県名、市町村区名、大字・通称名、字・丁目名からなる全国の住所名と一致検索することにより地域表現または住所表現を抽出する。そして、住所表現が、正式な住所表現から「県」または「市」の文字列が省略された例外住所表現または郡名が省略された例外住所表現である場合には、例外判定処理において正式な住所表現ではないと判定し、例外住所表現に「県」、「市」または郡名を補うことにより正式な住所表現に変換するようにしたものである。したがって、本発明の位置情報の自動抽出方法では、文章中に含まれている、正式な住所表現、正式でない住所表現および地域名表現等の位置情報を高い確率で自動的に抽出することができるとともに位置情報補足語を含めた位置情報を抽出することができる。

【０００８】また、本発明の位置情報の自動抽出方法は、前記例外判定処理が、前記検索対象形態素と予め登録された都道府県名の文字列との一致検索を行う処理と、前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名と一体した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、前記検索対象形態素と予め登録された市町村区名の文字列との一致検索を行う処理と、前記市町村区名との一致検索を行う処理において前記検索対象形態素が市町村区名の文字列と一致した場合に、前記検索対象形態素を含む表現は正式な住所表現であると判定する処理と、前記都道府県名の文字列との一致検索において前記検索対象形態素が都道府県名の文字列と一致せず、かつ前記市町村区名との一致検索を行う処理において検索対象形態素が市町村区名の文字列と一致しなかった場合に、前記検索対象形態素を含む表現は正式な住所表現ではないと判定する処理とから構成される。

【０００９】また、本発明の位置情報の自動抽出方法は、前記住所抽出処理が、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、前記検索対象形態素と、予め登録された大字・通称名の文字列との一致検索を行う処理と、前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、前記検索対象形態素と、予め登録された字・丁目の文字列との一致検索を行う処理と、前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、前記大字・通称名の文字

列との一致検索において前記検索対象形態素が大字・通称名の文字列と一致しなかった場合、または前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列と一致しなかった場合、または前記大字・通称名の文字列との一致検索において前記検索対象形態素が大字・通称名の文字列とも一致し、かつ前記字・丁目の文字列との一致検索において前記検索対象形態素が字・丁目の文字列とも一致した場合に、検索対象形態素の次の形態素が数字であるかどうかを判定し、数字である場合に該数字を住所表現における番地号であるとして抽出する番地号抽出処理と、抽出された都道府県名の形態素、市町村名の形態素、大字・通称の形態素、字・丁目の形態素、番地号の形態素を連結して１つの住所表現とする処理とから構成される。

【００１０】また、本発明の位置情報の自動抽出方法は、前記例外住所抽出処理が、前記住所抽出処理において正式な住所表現でないと判定された表現が、正式な住所表現から郡名が省略された例外住所表現であるかどうかを判定する例外住所タイプ判定処理と、前記例外住所タイプ判定処理において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現であると判定された場合に、前記検索対象形態素と、予め登録された、都道府県名から「県」が省略された文字列とを一致検索する処理と、前記検索対象形態素と、予め登録された、市名から「市」が省略された文字列とを一致検索する処理と、前記「市」が省略された文字列との一致検索において前記検索対象形態素が「市」が省略された文字列と一致した場合に、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理と、前記検索対象形態素と、予め登録された人名を記述する際に使用される表現の文字列との一致検索する人名判定処理と、前記人名判定処理において検索対象形態素が、人名を記述する際に使用される表現の文字列と一致した場合および検査対象形態素が「県」が省略された文字列とも「市」が省略された文字列とも一致しなかった場合に、検査対象形態素を含む表現は、住所表現ではないと判定する処理とを有する。

【００１１】また、本発明の位置情報の自動抽出方法は、前記例外住所抽出処理が、前記例外住所タイプ判定処理において検索対象形態素を含む表現が正式な住所表現から郡名が省略された例外住所表現でないと判定された場合に、都道府県名の文字列との一致検索において検索された都道府県名に基づいて、市町村区名との一致検索において一致した町村名と、予め登録された正式な郡－町村名から群名のみを省略した文字列との一致検索を行ない、一致した町村名の代わりにその町村名に対応する正式な郡－町村名の文字列を検索された住所表現とすることにより省略された郡名を補う処理と、前記検索対象形態素の次の形態素を新たな検索対象形態素とする処理とをさらに有する。

【００１２】また、本発明の位置情報の自動抽出方法は、前記位置情報補足語抽出処理が、住所表現又は地域名表現の末尾から一定の範囲内にある形態素と、予め登録された位置情報補足語との一致検索を行う処理と、前記位置情報補足語との一致検索処理において一致した場合に、前記住所表現または地域名表現から一致した形態素までを１つの位置情報として抽出する処理と、前記位置情報補足語との一致検索処理において一致しなかった場合に、検索対象形態素を６語前に戻す処理とから構成される。

【００１３】

【発明の実施の形態】次に、本発明の実施形態について図面を参照して詳細に説明する。

【００１４】図１は、本発明の一実施形態の位置情報の自動抽出装置の構成を示すブロック図である。

【００１５】本実施形態の位置情報の自動抽出装置は、住所データベース（ＤＢ）１０と、地域名表現データベース（ＤＢ）２０と、位置情報補足データベース（ＤＢ）３０と、人名判定データベース（ＤＢ）４０と、形態素解析部５０と、地域名表現抽出部６０と、住所表現抽出部７０と、例外住所表現抽出部８０と、位置情報補足語抽出部９０とから構成されている。

【００１６】住所データベース１０は、図２に示すように、住所テーブル１１と、都道府県対応テーブル１２と、市対応テーブル１３と、郡－町村対応テーブル１４とから構成されている。

【００１７】住所テーブル１１は、図３に示すように、「東京都」等の都道府県名が登録されている都道府県フィールドと、「横浜市」等の市名、「葉山町」等の町名、「白馬村」等の村名、「新宿区」等の東京２３区名が登録されている市町村区フィールドと、市町村区名の下位の住所表現となる大字・通称フィールドと、さらに下位の住所表現となる字・丁目フィールドとから構成されている。そして、これら４つのフィールドの組合せにより、日本全国の住所を全て表現することができるようになっている。

【００１８】都道府県対応テーブル１２は、図４（ａ）に示すように、都道府県フィールドと、都道府県名から「都」、「道」、「府」、「県」が省略された文字列が、都道府県フィールドと対応して登録されている都道府県省略フィールドとから構成されている。

【００１９】市対応テーブル１３は、図４（ｂ）に示すように、市フィールドと、市名から「市」が省略された文字列が、市フィールドと対応して登録されている市省略フィールドとから構成されている。

【００２０】郡－町村対応テーブル１４は、図５に示すように、都道府県フィールドと、郡名とその郡に属する町村名が登録された郡－町村フィールドと、郡－町村名フィールドから郡名が省略された町村名のみが登録された郡省略フィールドとから構成されている。そして、都

道府県フィールドと、郡－町村フィールドと郡省略フィールドはそれぞれ対応づけられている。

【００２１】地域名表現データベース２０は、図６（ａ）に示すように、「西日本」、「関東地方」等の住所表現ではない地域名表現が登録されている。

【００２２】位置情報補足データベース３０は、図６（ｂ）に示すように、「近く」、「付近」、「東部」、「南部」等の位置情報を補足するための表現である位置情報補足語が登録されている。

【００２３】人名判定データベース４０は、図６（ｃ）に示すように、「様」、「氏」、「容疑者」、「監督」等の人名を記述する際に使用される敬称等の表現が登録されている。

【００２４】形態素解析部５０は、入力された文章を、予め文字列が登録された辞書を使用して形態素解析することにより各形態素に分解している。ここで、形態素とは、文章を１つまたは数個の文字列毎に区切った単位であり、基本的に単語と同程度の文字列のことである。形態素解析において使用される辞書には、住所テーブル１１を構成している都道府県フィールド、市町村区フィールド、大字・通称フィールド、字・丁目フィールドに登録されている文字列も全て登録しておく。また、形態素解析の際には分解される形態素の大きさを決定する方法である最小値コスト法が用いられているが、都道府県、市町村区、大字・通称、字・丁目の各文字列の単位がこれ以上細分されないように、最小値コスト法における最小値を設定しておく。例えば、「神奈川県鎌倉市における県道で交通事故が発生しました。」という文章を形態素解析することにより、「神奈川県」、「鎌倉市」、「に」、「おける」、「県道」、「で」、「交通事故」、「が」、「発生」、「しました。」という複数の形態素が得られる。

【００２５】地域名表現抽出部６０は、形態素解析部５０により分解された各形態素と、地域名表現データベース２０に登録されている文字列とを比較することにより地域名表現検索を行なっている。

【００２６】住所表現抽出部７０は、地域名表現抽出部６０において、地域名とは一致しなかった各形態素が正式な住所表現であるかどうかの判定である例外判定処理を行ない、各形態素が正式な住所表現であると判定した場合には、各形態素と住所データベース１０の住所テーブル１１に登録されている文字列とを比較することにより文章から住所表現を抽出している。この例外判定処理においては、県名において「県」の文字列が含まれていない場合、市名において「市」の文字列が含まれていない場合には正式な住所表現ではないと判定される。また、正式な都道府県名が検索された場合でも、県名の次にくる文字列が郡名の下位の町村名の文字列である場合には郡名表示が省略された住所表現であると判定し、正式な住所表現ではないと判定する。

【００２７】例外住所表現抽出部８０は、住所表現抽出部７０において正式な住所表現ではないと判定された各形態素と、都道府県対応テーブル１２の都道府県省略フィールドまたは市対応テーブル１３の市省略フィールドに登録されている文字列とを検索することにより「県」または「市」が省略されいる例外住所表現を抽出し、「県」または「市」の文字列を補い正式な住所表現に変換する。

【００２８】また、例外住所表現抽出部８０は、地域名表現抽出部６０における例外判定処理において、郡名が省略されていると判定された場合には、例外判定処理において検索された都道府県名と町村名を元に、郡－町村対応テーブル１４を検索することにより省略された郡名を判定し、抽出した郡名が省略された例外住所表現に郡名を追加して正式な住所表現に変換する。

【００２９】さらに、例外住所表現抽出部８０は、各形態素の次の形態素または２語目の形態素が人名判定データベース４０に登録されている文字列の場合には、検索対象形態素は人名であると判定して住所表現として抽出しないようにする。このことにより、例外住所抽出処理において、例えば「鎌倉太郎」という人名の文字列から「鎌倉市」といる住所表現を誤って抽出することを避けることができる。

【００３０】位置情報補足語抽出部９０は、地域名表現抽出部６０、住所表現抽出部７０、例外住所表現抽出部８０において住所表現または地域名表現の位置情報が抽出された場合に、その位置情報の末尾から６語の範囲内にある形態素と、位置情報補足データベース３０に登録された文字列とを比較することにより位置情報補足語の検索処理を行なう。そして、位置情報補足語が抽出された場合には、位置情報からその検索された位置情報補足語までを新たに位置情報として出力する。

【００３１】次に、本実施形態の位置情報の自動抽出装置の動作を図７～図１１のフローチャートを参照して詳細に説明する。

【００３２】先ず、図７を参照して、本実施形態の位置情報の自動抽出装置の全体の動作について説明する。

【００３３】入力された文章は、形態素解析部５０において形態素解析されて各形態素に分解される。（ステップ１０１）。

【００３４】形態素解析部５０において分解された各形態素は、先ず、地域名表現抽出部６０において、地域名表現データベース２０に登録された文字列と比較検索される（ステップ１０２）。検索対象の形態素がステップ１０２にいて地域名表現であると判定された場合には、地域名表現抽出部６０はその形態素を地域名表現として抽出する（ステップ１１１）。

【００３５】ステップ１０２において、検索対象の形態素が地域名表現ではないと判定された場合には、正式な住所表現のみを取り出すために、住所表現抽出部７０に

おいて正式な住所表現であるかどうかが判定される（ステップ１０３）。

【００３６】ステップ１０３において正式な住所表現であると判定された場合には、住所表現抽出部７０において、住所表現の抽出処理が行われる（ステップ１０４）。

【００３７】ステップ１０３において正式な住所表現ではないと判定された場合には、例外住所表現抽出部８０において、例外住所表現抽出処理が行われ（ステップ１０８）、例外住所表現が存在する場合には（ステップ１０９）、正式な住所表現に変換される。例外住所抽出処理において、例外住所が存在しなかった場合には、例外住所表現抽出部８０は、検索対象形態素が位置情報に関するものではないと判定する。そして、全ての文章の処理が終了していないかが確認された後（ステップ１０７）、次の形態素が検索対象形態素として処理が行われる。

【００３８】ステップ１１１における地域名表現の抽出処理、ステップ１０４における住所表現の抽出処理、ステップ１１０における例外住所表現の正式な住所表現の変換処理が行われると、位置情報補足語抽出部９０は、抽出された位置情報の末尾から６語の範囲内にある形態素と、位置情報補足データベース３０に登録された文字列とを比較することにより位置情報補足語の検索処理を行なう。そして、位置情報補足語が抽出された場合には、位置情報からその検索された位置情報補足語までを新たな位置情報とし（ステップ１０５）、得られた位置情報を出力する（ステップ１０６）。

【００３９】そして、全ての文章の処理が終了していないかが確認された後（ステップ１０７）、次の形態素を検索対象形態素として処理が行なわれる。

【００４０】次に、図７中の例外判定処理（ステップ１０３）を図８のフローチャートを用いてさらに詳しく説明する。

【００４１】住所表現抽出処理部７０において、各形態素と住所テーブル１１の都道府県フィールドに登録されている文字列との一致検索が行なわれ、一致した場合にはその形態素を配列１に格納し、検索対象形態素を１つずらすとともにあるフラグを“Ａ”とし、一致しない場合には“Ｂ”とする（ステップ２０１）。

【００４２】次に住所表現抽出処理部７０は、各形態素と住所テーブル１１の市町村区フィールドに登録されている文字列との一致検索を行う（ステップ２０２）。ステップ２０２の一致検索処理において、検索対象形態素と市町村区フィールドに登録された文字列とが一致した場合には、正式な住所表現であると判定して処理をステップ１０４に移す。

【００４３】ステップ２０２の一致処理において、検索対象形態素と市町村区フィールドに登録された文字列とが一致しなかった場合には、ステップ２０１における処理結果に基づき次の処理を決定する（ステップ２０３）。具体的には、ステップ２０１において、都道府県フィールドとの一致検索において検索対象形態素と文字列が一致した場合（フラグが“Ａ”の場合）には、正式な住所表現であると判定して処理をステップ１０３に進め、一致しなかった場合（フラグが“Ｂ”の場合）には、例外住所表現であると判定して処理をステップ１０８に進める（ステップ２０３）。

【００４４】次に、図７中の住所表現の抽出処理（ステップ１０４）を図９のフローチャートを用いてさらに詳しく説明する。

【００４５】先ず、住所表現抽出部７０は、現在の検索対象形態素は正式な住所表現における市町村名であるため、その形態素を配列２に格納した後に検索対象形態素を１つづらす（ステップ５０４）。

【００４６】そして、検索対象形態素と、住所テーブル１１の大字・通称フィールドとの一致検索を行ない（ステップ５０５）、一致した場合には検索対象形態素を配列３に格納し、検索対象形態素を１つづらす（ステップ５０６）。同様にして、検索対象形態素と、住所テーブル１１の字・丁目フィールドとの一致検索を行ない（ステップ５０７）、一致した場合には検索対象形態素を配列４に格納し、検索対象形態素を１つづらす（ステップ５０８）。

【００４７】ステップ５０５またはステップ５０７のどちらかの一致検索において、一致しなかった場合、およびステップ５０８の処理の次に番地号の抽出処理が行われる（ステップ５０９）。番地号の抽出処理においては、住所表現の形態素の次の形態素が数字である場合に、その数字を番地号表現とする。住所表現の形態素が、数字のみ、数字－形態素－数字、又は数字－形態素－数字－形態素－数字、の並びはすべて番地号表現とする。

【００４８】最後に、住所表現抽出部７０は、配列１～４に格納されている各形態素および抽出された番地号に対する連結し１つの住所表現として住所表現抽出処理を終了する（ステップ５１０）。この住所表現抽出処理により住所表現の先頭となり得るのは、都道府県フィールド又は市町村区フィールドに格納されている文字列のみであり、他のフィールドに格納されている文字列は住所表現の先頭にはなり得ない。

【００４９】次に、図７中の例外住所表現抽出処理（ステップ１０８）および例外住所表現有無判定処理（ステップ１０９）を図１０のフローチャートを用いてさらに詳しく説明する。

【００５０】本実施形態では、「県」または「市」の文字が省略されている住所表現と、郡名が省略されている住所表現を例外住所表現として処理している。例外住所表現抽出部８０では、ステップ４０１において、例外住所表現と判定された住所表現がどちらのタイプかを判定

し、「県」または「市」の文字が省略されている例外住所表現である場合にはステップ３０１～３０４の処理を行ない、郡名が省略されている例外住所表現である場合にはステップ４０２、４０３の処理を行う。

【００５１】先ず最初に「県」または「市」の文字が省略されている例外住所表現の抽出処理について説明する。

【００５２】例外住所表現抽出部８０は、検索対象形態素と都道府県対応テーブル１２の都道府県省略フィールドとの一意検索を行う（ステップ３０１）。ステップ３０１において、一致しなかった場合には、検索対象形態素と市対応テーブル１３の市省略フィールドとの一意検索を行う（ステップ３０２）。

【００５３】ステップ３０１において一致した場合、およびステップ３０２において一致した場合には、その検索対象形態素を配列に格納し検索対象形態素を１つづらす（ステップ３０３）。そして、検索対象形態素および次の形態素と、人名判定データベース４０との一致検索を行う人名判定処理が行われる（ステップ３０４）。ステップ３０４における人名判定処理において、人名表現ではないと判定された場合には例外住所表現であると判定され処理をステップ１１０の進める。

【００５４】都道府県省略フィールド又市省略フィールドと正式住所表現は対応付けられており、上記検索により住所表現と判定されれば、抽出された例外住所表現は「県」又は「市」が付与されて正式住所に変換される（ステップ１１０）。

【００５５】ステップ３０４における人名判定処理において人名表現であると判定された場合およびステップ３０２において一致しなかった場合には、検索対象形態素により表された文字列は、正式住所表現でも例外住所表現でもないと判定され処理をステップ１０７のに進める。

【００５６】この処理により、例えば、「神奈川」や「横浜」という例外住所表現は、「神奈川県」や「横浜市」という正式な住所表現にそれぞれ変換される。

【００５７】次に、郡名が省略されている例外住所表現の抽出処理について説明する。

【００５８】先ず、例外住所抽出部８０は、郡－町村対応フィールド１４を用いて省略された郡名を補う処理を行う。この際に、例外住所抽出部８０は、ステップ２０１で検索された都道府県名情報より検索範囲を絞り、ステップ２０２において一致した町村名と群省略フィールドに登録された文字列との一致検索を行ない、一致した町村名の代わりにその町村名に対応する郡－町村フィールドの文字列を検索された住所表現とすることにより省略された郡名を補う処理を行う（ステップ４０２）。そして、例外住所抽出部８０は、検索対象形態素を１つずらす（ステップ４０３）。

【００５９】このステップ４０２における処理を、ステップ２０１において検索された都道府県名が「長野県」であり、ステップ２０２において検索された町村名が「白馬村」である場合を用いて具体的に説明する。先ず、例外住所抽出部８０は、都道府県フィールドが「長野県」である群省略フィールドに登録された文字列と「白馬村」との一致検索を行う。そして、「白馬村」の文字列を一致した郡省略フィールドに対応する郡－町村フィールドの文字列「北安曇郡白馬村」に置き換える。この処理により、例外住所表現である「長野県白馬村」は「長野県北安曇郡白馬村」に置き換えられる。

【００６０】最後に、図７中の位置情報補足語の抽出処理（ステップ１０５）を図１１のフローチャートを用いてさらに詳しく説明する。

【００６１】位置情報補足語抽出部９０は、住所表現又は地域名表現の末尾から６語以内の範囲にある形態素と、位置情報補足データベース３０との一致検索を行なう（ステップ７０２）。ステップ７０２において、一致した文字列が存在した場合には、抽出された住所表現又は地域名表現からステップ７０２において抽出された形態素までを１つの位置情報として抽出する（ステップ７０３）。ステップ７０２において、一致した文字列が存在しない場合には、位置情報補足語抽出部９０は、検索対象形態素を６語前に戻し位置情報補足語抽出処理を終了する。

【００６２】本実施形態における自動抽出装置を用いて、新聞記事（１０００）記事およびインターネットにおけるホームページ（３００ページ）中に含まれる位置情報の自動抽出を行なったところ、新聞記事では９５．２％、ホームページでは、８０．１％の自動抽出率を得ることができた、ホームページにおける自動抽出率が新聞記事よりも低下したのは、ホームページでは、文字が文字情報ではなく画像情報として与えられている場合があるためである。

【００６３】このように、本実施形態の位置情報の自動抽出装置では、文章中に含まれている、正式な住所表現、正式でない住所表現および地域名表現等の位置情報を高い確率で自動的に抽出することができるとともに位置情報補足語を含めた位置情報を抽出することができる。

【００６４】また、図には示されていないが、本実施形態の自動抽出装置は、データ処理装置（コンピュータ）と、記憶装置と、入出力処理装置と、自動抽出方法を実行するためのプログラムを記録した記録媒体とによっても構成することができる。この記録媒体は磁気ディスク、半導体メモリまたはその他の記録媒体であってもよい。このプログラムは、記録媒体からデータ処理装置に読み込まれ、データ処理装置の動作を制御し、図１における形態素解析部５０、地域名表現抽出部６０、住所表現抽出部７０、例外住所表現抽出部８０、位置情報補足語抽出部９０によって行われる処理を実行する。そし

て、記憶装置は、住所データベース１０、地域名表現データベース２０、位置情報補足データベース３０、人名判定データベース４０により構成され、入出力装置は、位置情報を抽出するための文章情報の入力および文章から抽出された位置情報の出力を行う。を備えている。

【００６５】

【発明の効果】以上説明したように、本発明は、文章中に記述されている住所表現又は地域名表現を自動的に抽出することが可能となることにより、文章中における位置情報を検索する時間が大幅に短縮されるという効果を有する。

【００６６】また、地理情報システムに対して本発明を適用した場合には、新聞記事などの情報中から位置情報を自動で抽出することにより、地理情報システムに自動的に情報を貼り付けることが可能となるという効果を有する。

【図面の簡単な説明】

【図１】本発明の一実施形態の位置情報の自動抽出装置の構成を示すブロック図である。

【図２】図１中の住所データベース１０のデータ構造を示す図である。

【図３】図２中の住所テーブル１１のデータ構造を示す図である。

【図４】図２中の都道府県対応テーブル１２のデータ構造を示す図（図４（ａ））および市対応テーブル１３のデータ構造を示す図（図４（ｂ））である。

【図５】図２中の郡－町村対応テーブル１４のデータ構造を示す図である。

【図６】図１中の地域名表現データベース２０のデータ構造を示す図（図６（ａ））、位置情報補足データベース３０のデータ構造を示す図（図６（ｂ））および人名判定データベース４０の構造を示す図（図６（ｃ））である。

【図７】図１の位置情報の自動抽出装置の動作を示すフローチャートである。

【図８】図７中の例外判定処理（ステップ１０３）をさらに詳しく示したフローチャートである。

【図９】図７中の住所表現抽出処理（ステップ１０４）をさらに詳しく示したフローチャートである。

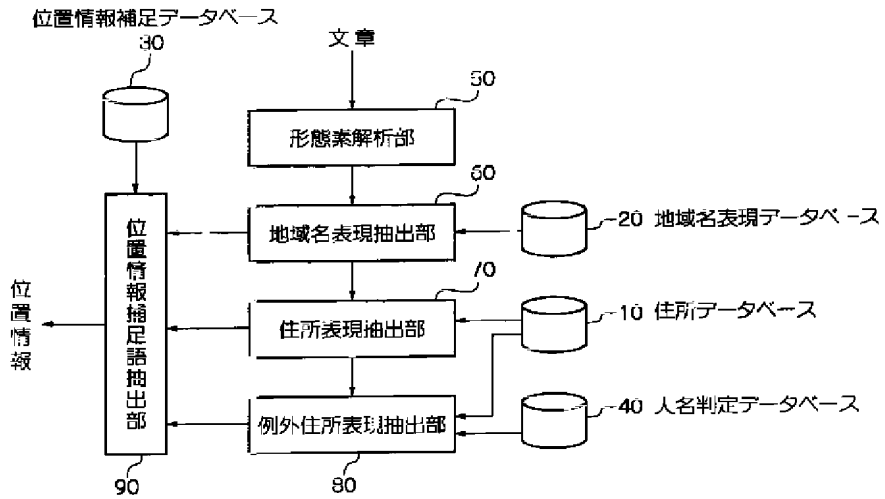
【図１０】図７中の例外住所表現抽出処理（ステップ１０８）および例外住所表現有無判定処理（ステップ１０９）をさらに詳しく示したフローチャートである。

【図１１】図７中の位置情報補足語の抽出処理（ステップ１０５）をらに詳しく示したフローチャートである。

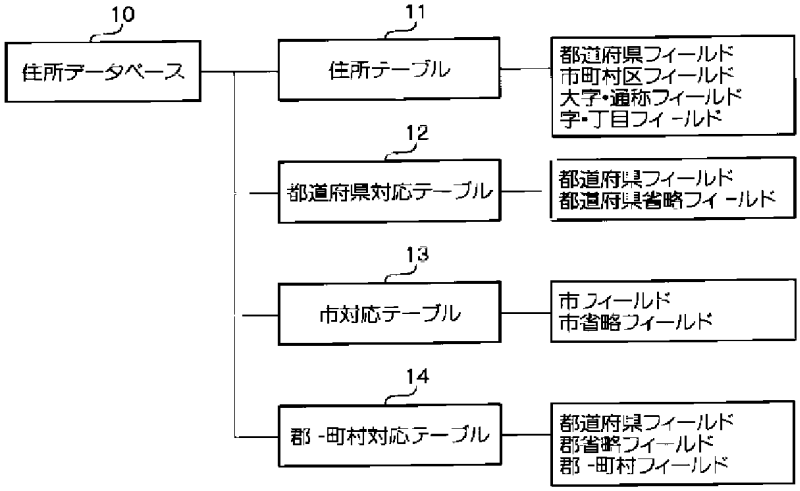
【符号の説明】

１０　住所データベース（ＤＢ）
１１　住所テーブル
１２　都道府県対応テーブル
１３　市対応テーブル
１４　群－町村対応テーブル
２０　地域名表現データベース（ＤＢ）
３０　位置情報補足データベース（ＤＢ）
４０　人名判定データベース（ＤＢ）
５０　形態素解析部
６０　地域名表現抽出部
７０　住所表現抽出部
８０　例外住所表現抽出部
９０　位置情報補足語抽出部
１０１～１１１　ステップ
２０１～２０３　ステップ
３０１～３０４　ステップ
４０１～４０３　ステップ
５０４～５１０　ステップ
７０２～７０４　ステップ

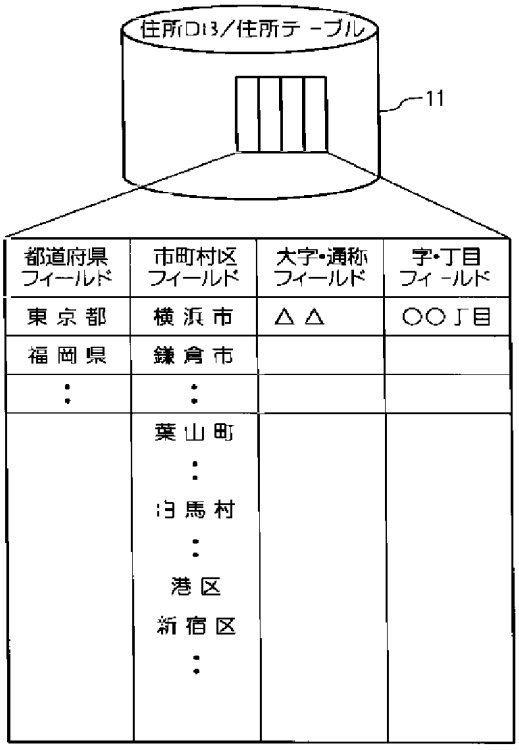
【図１】



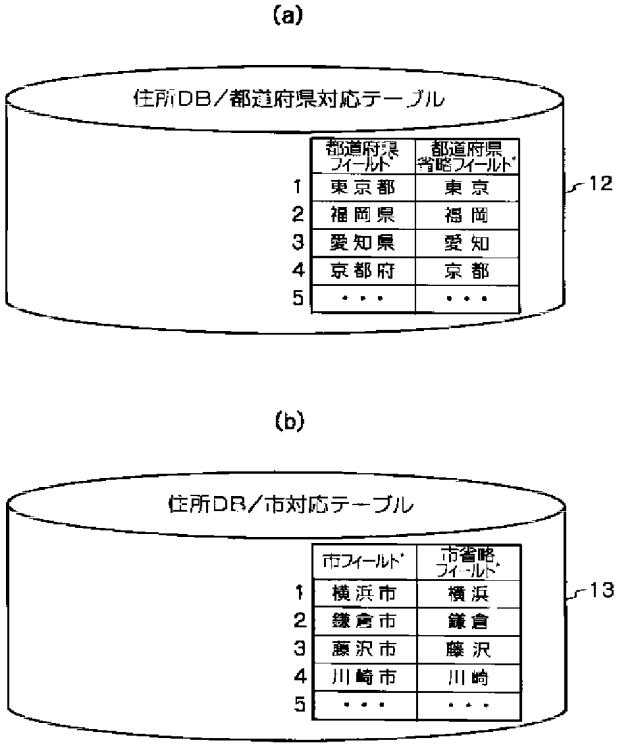
【図2】


10
住所データベース
11
住所テーブル
都道府県フィールド
市町村区フィールド
大字・通称フィールド
字・丁目フィールド
12
都道府県対応テーブル
都道府県フィールド
都道府県省略フィールド
13
市対応テーブル
市フィールド
市省略フィールド
14
郡・町村対応テーブル
都道府県フィールド
郡省略フィールド
郡・町村フィールド

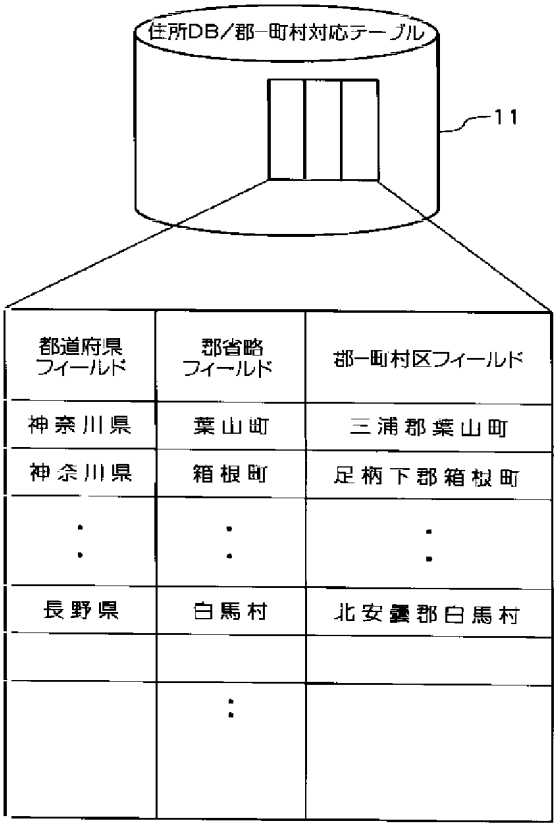
【図3】


住所DB/住所テーブル
11
都道府県フィールド
市町村区フィールド
大字・通称フィールド
字・丁目フィールド
東京都
横浜市
△△
○○丁目
福岡県
鎌倉市
葉山町
:
羽馬村
:
港区
新宿区
:

【図4】

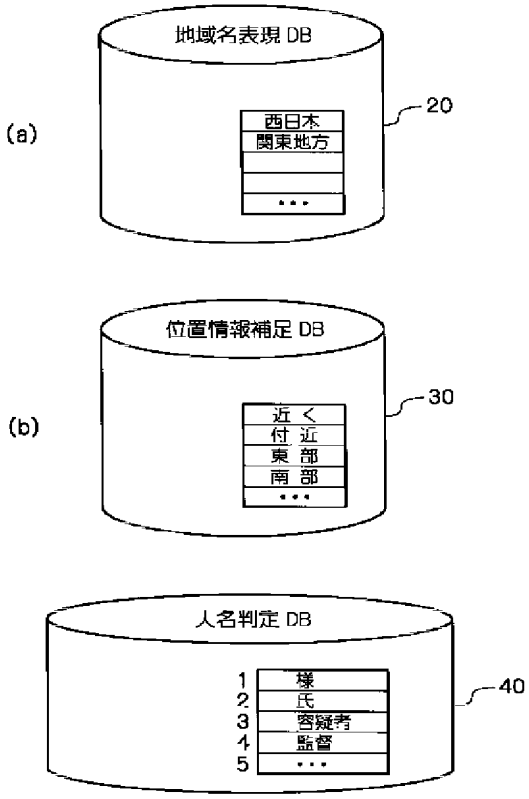

(a)
住所DB/都道府県対応テーブル
12
都道府県フィールド
都道府県省略フィールド
1 東京都 東京
2 福岡県 福岡
3 愛知県 愛知
4 京都府 京都
5 ・・・ ・・・
(b)
住所DB/市対応テーブル
13
市フィールド
市省略フィールド
1 横浜市 横浜
2 鎌倉市 鎌倉
3 藤沢市 藤沢
4 川崎市 川崎
5 ・・・ ・・・

【図5】

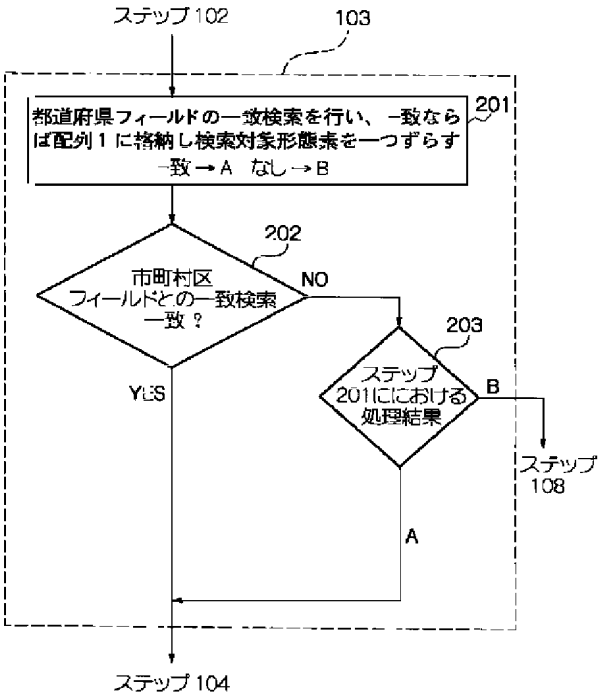


都道府県フィールド	郡省略フィールド	郡−町村区フィールド
神奈川県	葉山町	三浦郡葉山町
神奈川県	箱根町	足柄下郡箱根町
・	・	・
・	・	・
長野県	白馬村	北安曇郡白馬村
	：	

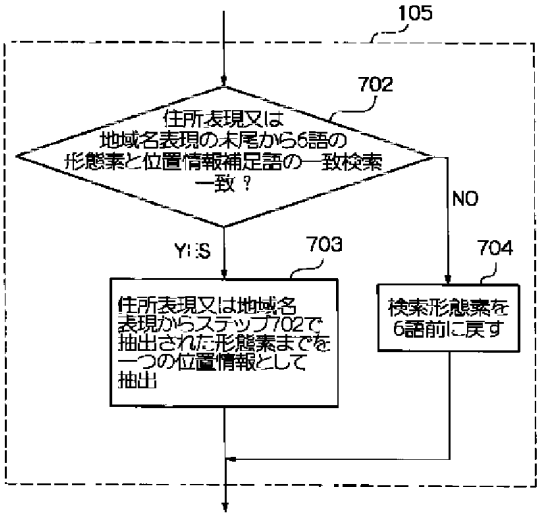
【図6】



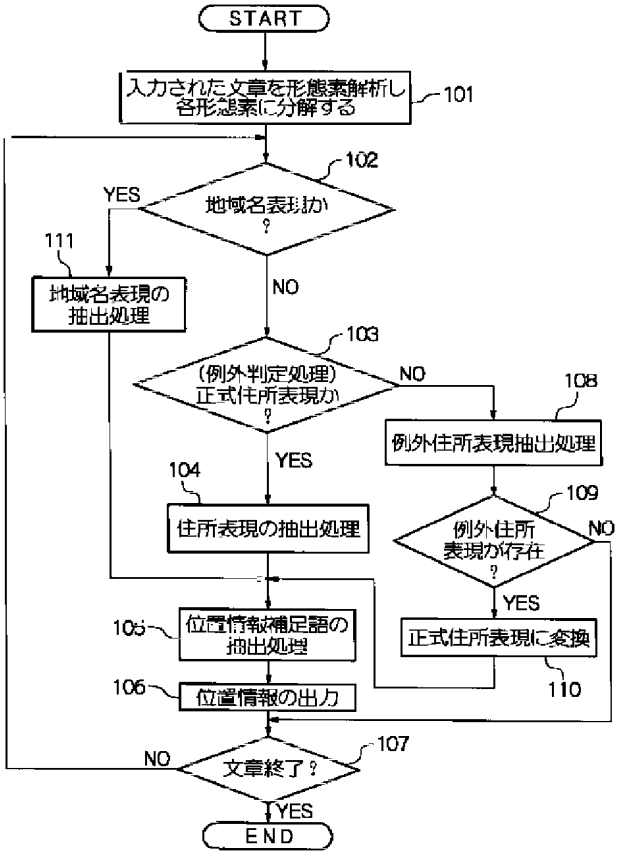
【図8】



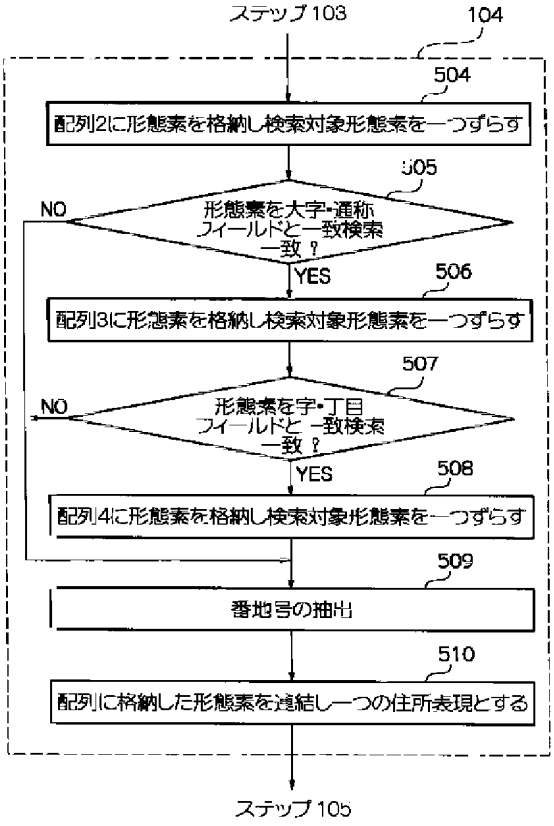
【図11】



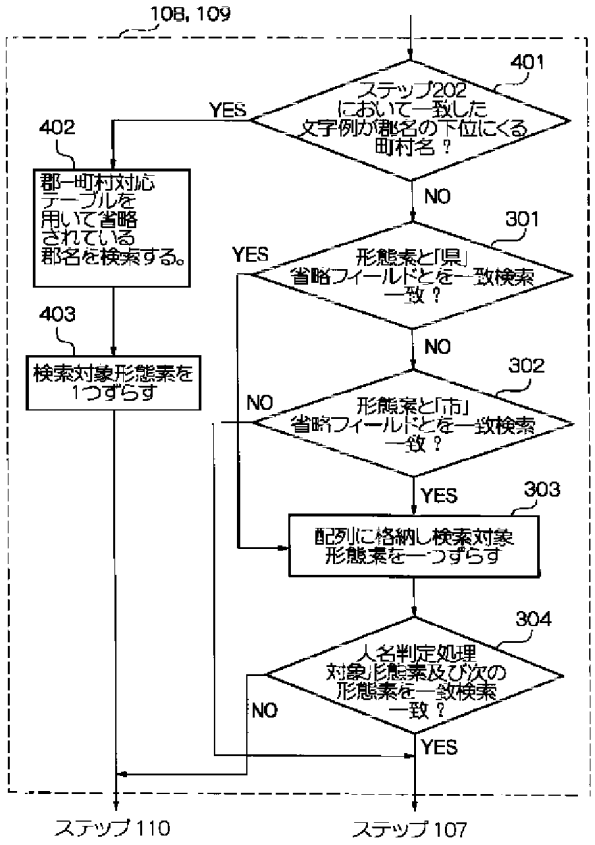
【図７】


START
入力された文章を形態素解析し各形態素に分解する
101
102
地域名表現か？
YES
NO
111
地域名表現の抽出処理
103
（例外判定処理）正式住所表現か？
NO
108
例外住所表現抽出処理
YES
104
住所表現の抽出処理
109
例外住所表現が存在？
NO
YES
105
位置情報補足語の抽出処理
正式住所表現に変換
110
106
位置情報の出力
NO
文章終了？
107
YES
END

【図９】


ステップ103
104
504
配列2に形態素を格納し検索対象形態素を一つずらす
505
NO
形態素を大字・通称フィールドと一致検索一致？
YES
506
配列3に形態素を格納し検索対象形態素を一つずらす
507
NO
形態素を字・丁目フィールドと一致検索一致？
YES
508
配列4に形態素を格納し検索対象形態素を一つずらす
509
番地号の抽出
510
配列に格納した形態素を連結し一つの住所表現とする
ステップ105

【図１０】


108, 109
401
ステップ202
において一致した
文字例が郡名の下位にくる
町村名？
YES
402
郡ー町村対応
テーブルを
用いて省略
されている
郡名を検索する。
NO
301
形態素と「県」
省略フィールドとを一致検索
一致？
YES
403
検索対象形態素を
1つずらす
NO
302
形態素と「市」
省略フィールドとを一致検索
一致？
NO
YES
303
配列に格納し検索対象
形態素を一つずらす
304
人名判定処理
対象形態素及び次の
形態素を一致検索
一致？
NO
YES
ステップ110
ステップ107

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-348041
(43)Date of publication of application : 15.12.2000

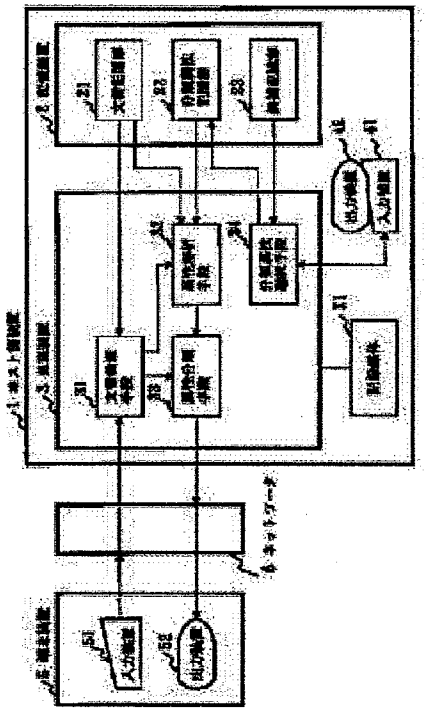
(51)Int.Cl. G06F 17/30
G06F 17/21

(21)Application number : 11-156423
(22)Date of filing : 03.06.1999

(71)Applicant : NEC CORP
(72)Inventor : IKEDA TAKAHIRO
SATO KENJI
OKUMURA AKITOSHI

(54) DOCUMENT RETRIEVAL METHOD, DEVICE THEREFOR AND MECHANICALLY READABLE RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To easily select a document required by a user by classifying many retrieved documents from a point of view corresponding to a field to which a document to be retrieved belongs in a document retrieving device.
SOLUTION: A document storing part 21 stores plural documents belonging to some field. The manager of a host-side device 1 selects the kind of an attribute suited to classifying the document of the field stored in the part 21 from in the kind of the attribute which is stored in a candidate storing part 23 and can be made to be a classifying key, and stores it in a sorting attribute storing part 22. An attribute analyzing means 32 analyzes which attribute kind among attribute kinds stored in the part 22 an attribute element existing in each document retrieved by a document retrieving means 31 belongs to. The means 33 classifies each retrieved document by each kind of attribute element included in the document in accordance with the analyzing result of the means 32.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]Out of a kind of two or more attributes usable when searching a document which suits specified conditions from two or more documents and classifying a document. Choose a kind of attribute used when classifying a document, and it stores in a classifying attribute storage parts store, A document retrieval method analyzing whether an attribute element belonging to a kind of which attribute of the kinds of attribute stored in said classifying attribute storage parts store exists in said each searched document, and classifying said each searched document for every kind of attribute element which the document contains based on the analysis result.
[Claim 2]A document retrieval method classifying into an independent category a document which does not contain an attribute element belonging to a kind of attribute stored in said classifying attribute storage parts store in the document retrieval method according to claim 1 when classifying said each searched document.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates to the document-retrieval art which classifies search results automatically and outputs them especially from a lot of documents about the document-retrieval art of searching what suits the specified conditions.
[0002]
[Description of the Prior Art]Before, methods of searching the target document out of a lot of documents include all the inputted keywords or the method of searching a document including the part. Such a search method is realized as software which searches the service which searches various kinds of documents currently exhibited by the Internet, a commercial online service, etc., and the document placed by the hard disk. However, in the criteria specification by a keyword, though it was difficult to narrow down only to the document which a user needs and it suited conditions in search results from a lot of documents, there was a problem that many unnecessary documents will mix. Although search results can be narrowed down now one by one by adding a keyword in the service which searches the document on the Internet, an unnecessary document still cannot be eliminated thoroughly.
[0003]In order to solve this problem, search results are not extracted on the conditions by a keyword, but the method of classifying search results from other viewpoints and showing a user exists. For example, the method of classifying search results into JP,H8-235160,A and JP,H9-231238,A is indicated.
[0004]When the number of search results is larger than a predetermined value, with attribute data beforehand given to each document, such as a document name and a document registration date, search results are classified according to a document retrieval method and a device given in JP,H8-235160,A, and a user is shown with them.
[0005]Search results are classified and expressed as the text browsing result display method

and a device given in JP,H9-231238,A by analyzing the theme of each text of search results and dividing the theme into two or more groups.

[0006]The method of extracting a keyword with 5W1H attribute from each document to JP,H10-320411,A, and on the other hand, classifying a document according to the extracted keyword with 5W1H attribute on a two-dimensional matrix as a method of classifying two or more documents, is indicated.

[0007]

[Problem(s) to be Solved by the Invention]In the document retrieval method of a conventional example, there was a case where narrowing down of a suitable document or a suitable classification could not be provided to a user.

[0008]For example, suppose that the document containing the keyword "X hotel" was searched with the purpose of acquiring the information which the user who thinks that he would like to stay at a certain hotel "X hotel" needs in order to stay at "X hotel." In this case, information required for a user is a contact of "X hotel", an address of "X hotel", etc., and a document required for a user is a document in which those information is described. However, it cannot narrow down only to the document in which the contact and address of "X hotel" are described out of a lot of documents only on the conditions that the keyword "X hotel" is included. For example, in this case, although the Web document of the diary style of the news which reports that the exhibition of the new product was held in X hotel, and the contents of having had a meal at the restaurant of X hotel is unnecessary for a user, it is contained in search results. Since it cannot express by a keyword, the conditions that the information about a contact or an address is described in a document are adding a keyword further, cannot narrow down search results and cannot eliminate an unnecessary document.

[0009]In a document retrieval method and a device given in JP,H8-235160,A, although search results can be classified according to the attribute attached to the document, the attribute required for a classification must be beforehand given to the document. For this reason, search results cannot be classified into the document in which those information exists, and the document not existing unless it is beforehand recorded as an attribute of a document whether the information about a contact or an address exists. It is difficult to correspond to the Web document currently especially exhibited on the Internet.

[0010]Although search results can be classified into JP,H9-231238,A according to the text browsing result display method and device of a description according to the contents of the document, the standard of a classification is the theme of each document. Therefore, the information about a contact or an address can divide and classify the document currently written as the theme, and the document which is not written as the theme. However, the information about a contact or an address may be written also in the document in which the information about a contact or an address is not written as the theme. For example, the contact

and address of X hotel may be written to the news which has told as the theme that X hotel extended the new building. Therefore, it cannot necessarily be said that a document required for a user and an unnecessary document can be divided according to this classification.

[0011]Although a document can be classified according to the recording medium which recorded the document sorting device, method, and document group program of the description on JP,H10-320411,A by the keyword with 5W1H attribute extracted out of the document, Whenever the kind of 5W1H used as the key of a classification classifies, a user has to specify it. In order to classify according to the unit of 5W1H, a classification in a fine unit like an address, a nearby station, a telephone number, and an E-mail address cannot be performed.

[0012]Then, the purpose of this invention is enabling it to sort out the document which a user needs easily by classifying many searched documents according to the viewpoint according to the field to which a retrieval object document belongs.

[0013]

[Means for Solving the Problem]According to whether an attribute element showing concrete contents about a specific attribute (concept) is contained in a document, a document of search results is classified and the 1st document retrieval system of this invention is classified for every kind of the above-mentioned specific attribute about a document containing an attribute element about the above-mentioned specific attribute. An attribute element specifically expresses in a document an element which expresses concretely the contents of the specific attributes, such as an address, a telephone number, a nearby station, a price, a date, time, an E-mail address, URL, a company name, a product name, and a part number. For example, as an attribute element showing the attribute of an address, there is "Chiyoda-ku, Tokyo" etc. and there is "\12,000" etc. as an attribute element showing the attribute of a price.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL FIELD

[Field of the Invention]This invention relates to the document-retrieval art which classifies search results automatically and outputs them especially from a lot of documents about the document-retrieval art of searching what suits the specified conditions.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art]Before, methods of searching the target document out of a lot of documents include all the inputted keywords or the method of searching a document including the part. Such a search method is realized as software which searches the service which searches various kinds of documents currently exhibited by the Internet, a commercial online service, etc., and the document placed by the hard disk. However, in the criteria specification by a keyword, though it was difficult to narrow down only to the document which a user needs and it suited conditions in search results from a lot of documents, there was a problem that many unnecessary documents will mix. Although search results can be narrowed down now one by one by adding a keyword in the service which searches the document on the Internet, an unnecessary document still cannot be eliminated thoroughly.
[0003]In order to solve this problem, search results are not extracted on the conditions by a keyword, but the method of classifying search results from other viewpoints and showing a user exists. For example, the method of classifying search results into JP,H8-235160,A and JP,H9-231238,A is indicated.
[0004]When the number of search results is larger than a predetermined value, with attribute data beforehand given to each document, such as a document name and a document registration date, search results are classified according to a document retrieval method and a device given in JP,H8-235160,A, and a user is shown with them.
[0005]Search results are classified and expressed as the text browsing result display method and a device given in JP,H9-231238,A by analyzing the theme of each text of search results and dividing the theme into two or more groups.
[0006]The method of extracting a keyword with 5W1H attribute from each document to JP,H10-320411,A, and on the other hand, classifying a document according to the extracted keyword with 5W1H attribute on a two-dimensional matrix as a method of classifying two or more documents, is indicated.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

EFFECT OF THE INVENTION

[Effect of the Invention]The 1st effect of this invention is the point of enabling the user of retrieval-required origin to choose easily the document which the attribute element of the needed kind contains out of many searched documents.
[0102]The Reason An address, a telephone number, a nearby station, a price, a date, time, an E-mail address, Out of the kind of attribute which exist [part number / URL, a company name, a product name,] and which can serve as a sort key. It is because he is trying to classify the document searched only using the kind of attribute which chooses the kind of actually used attribute, stores in the classifying attribute storage parts store, and is stored in this classifying attribute storage parts store when classifying the searched document. That is, since it becomes a thing to which the document made into a retrieval object belongs and which differs in an effective sort key (viewpoint of a classification) for every field, if it classifies by fixing a sort key to 5W1H like conventional technology, a classification may not be performed in the form which a user tends to sort out, but. According to this invention, since the kind of attribute according to the field to which the document made into a retrieval object belongs can be chosen from the kinds of many attributes and it can be made a sort key, it can classify according to the form which a user tends to sort out.
[0103]The 2nd effect of this invention is being able to divide search results into the document in which the attribute element to which its attention should be paid is contained, and the document which is not contained. As a result, when the document in which the attribute to which its attention should be paid is not described is unnecessary, that unnecessary document can be easily excepted from search results.
[0104]An attribute analysis means the Reason about the attribute of several kinds which are memorized by the classifying attribute storage parts store. It analyzes what kind of attribute element is contained in each document of search results, and an attribute classification means is because the document in which the attribute element of the kind memorized by the

classifying attribute storage parts store is not contained is classified into the independent category.

[0105]The 3rd effect is being able to classify the document of search results according to the attribute element of the specific kind in a document. As a result, the user who needs the document in which a certain specific kind of attribute is described can obtain now the search results classified according to those concrete contents, i.e., the contents corresponding to a matter required for themselves. Thereby, it becomes easy to narrow down search results further.

[0106]This is because the documents in which an attribute element extraction means extracts the attribute element of the kind specified by a user out of each document of search results, and an attribute element sorting means contains the same attribute element classify search results so that it may become the same category.

[0107]The 4th effect is packing into one category the document which contains a near attribute element semantically, and being able to classify it so that the Type of the category at the time of classifying search results may not become detailed too much. As a result, the user can obtain the classification result of the detailed degree to need by specifying the level to collect.

[0108]The Reason holds the word to which a thesaurus storage parts store hits the generic concept of each word.

An attribute element thesaurus sorting means is because search results are classified so that the documents which ask for the word of the generic concept of the level specified by a user and to which the called-for word becomes the same may serve as the same category from the attribute element extracted from each document.

[0109]The 5th effect is that it is possible to stop the number of categories, as the category at the time of classifying search results does not increase too much.

[0110]The Reason is the same as the Reason of the 4th effect. That is, it is because the number of categories can be reduced by packing into one category the document which contains a near attribute element semantically, and classifying it with reference to a thesaurus.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]In the document retrieval method of a conventional example, there was a case where narrowing down of a suitable document or a suitable classification could not be provided to a user.
[0008]For example, suppose that the document containing the keyword "X hotel" was searched with the purpose of acquiring the information which the user who thinks that he would like to stay at a certain hotel "X hotel" needs in order to stay at "X hotel." In this case, information required for a user is a contact of "X hotel", an address of "X hotel", etc., and a document required for a user is a document in which those information is described. However, it cannot narrow down only to the document in which the contact and address of "X hotel" are described out of a lot of documents only on the conditions that the keyword "X hotel" is included. For example, in this case, although the Web document of the diary style of the news which reports that the exhibition of the new product was held in X hotel, and the contents of having had a meal at the restaurant of X hotel is unnecessary for a user, it is contained in search results. Since it cannot express by a keyword, the conditions that the information about a contact or an address is described in a document are adding a keyword further, cannot narrow down search results and cannot eliminate an unnecessary document.
[0009]In a document retrieval method and a device given in JP,H8-235160,A, although search results can be classified according to the attribute attached to the document, the attribute required for a classification must be beforehand given to the document. For this reason, search results cannot be classified into the document in which those information exists, and the document not existing unless it is beforehand recorded as an attribute of a document whether the information about a contact or an address exists. It is difficult to correspond to the Web document currently especially exhibited on the Internet.
[0010]Although search results can be classified into JP,H9-231238,A according to the text browsing result display method and device of a description according to the contents of the

document, the standard of a classification is the theme of each document. Therefore, the information about a contact or an address can divide and classify the document currently written as the theme, and the document which is not written as the theme. However, the information about a contact or an address may be written also in the document in which the information about a contact or an address is not written as the theme. For example, the contact and address of X hotel may be written to the news which has told as the theme that X hotel extended the new building. Therefore, it cannot necessarily be said that a document required for a user and an unnecessary document can be divided according to this classification.

[0011]Although a document can be classified according to the recording medium which recorded the document sorting device, method, and document group program of the description on JP,H10-320411,A by the keyword with 5W1H attribute extracted out of the document, Whenever the kind of 5W1H used as the key of a classification classifies, a user has to specify it. In order to classify according to the unit of 5W1H, a classification in a fine unit like an address, a nearby station, a telephone number, and an E-mail address cannot be performed.

[0012]Then, the purpose of this invention is enabling it to sort out the document which a user needs easily by classifying many searched documents according to the viewpoint according to the field to which a retrieval object document belongs.

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－348041
（P2000－348041A）
(43)公開日　平成12年12月15日(2000.12.15)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ	５Ｂ００９
17/21		15/20	５７０Ｎ	５Ｂ０７５
			５７０Ｒ	

審査請求　有　　請求項の数18　ＯＬ　（全 16 頁）

(21)出願番号　特願平11－156423

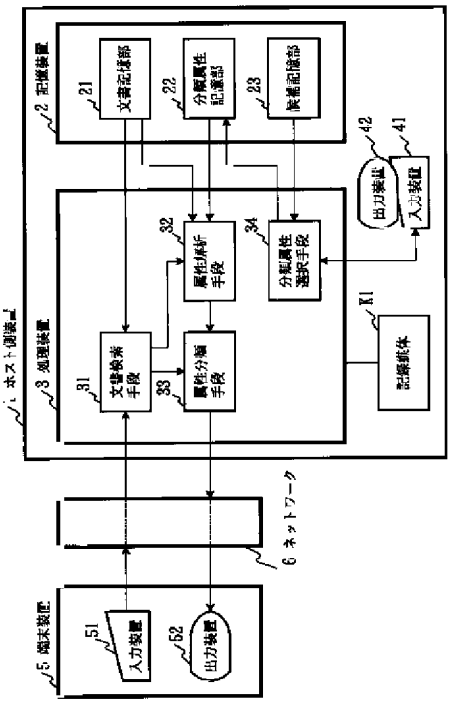
(22)出願日　平成11年６月３日(1999.6.3)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　池田　崇博
東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　佐藤　研治
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100088959
弁理士　境　廣巳

最終頁に続く

(54)【発明の名称】　文書検索方法及びその装置並びにプログラムを記録した機械読み取り可能な記録媒体

(57)【要約】
【課題】　文書検索装置において、検索された多数の文書を、検索対象文書が属する分野に応じた観点で分類することにより、ユーザが必要とする文書を容易に選別できるようにする。
【解決手段】　文書記憶部21には、或る分野に属する複数の文書が格納されている。ホスト側装置１の管理者は、分類属性選択手段34を用いて、候補記憶部23に格納されている、分類キーとすることができる属性の種類の中から、文書記憶部21に格納されている分野の文書を分類するのに適した属性の種類を選択して分類属性記憶部22に格納する。属性解析手段32は、文書検索手段31で検索された各文書に、分類属性記憶部22に格納されている属性の種類の内の、どの属性の種類に属する属性要素が存在するのかを解析し、属性分類手段33は、検索された各文書を、属性解析手段32の解析結果に従って、その文書が含む属性要素の種類毎に分類する。



【特許請求の範囲】
【請求項１】　指定された条件に適合する文書を複数の文書から検索し、
文書を分類する際に使用可能な複数の属性の種類の中から、文書を分類する際に使用する属性の種類を選択して分類属性記憶部に格納し、
前記検索した各文書に、前記分類属性記憶部に格納されている属性の種類の内の、どの属性の種類に属する属性要素が存在するのかを解析し、
その解析結果に基づいて、前記検索した各文書を、その文書が含有する属性要素の種類毎に分類することを特徴とする文書検索方法。
【請求項２】　請求項１記載の文書検索方法において、前記検索した各文書を分類する際、前記分類属性記憶部に格納されている属性の種類に属する属性要素を含有しない文書を、独立した範疇に分類することを特徴とする文書検索方法。
【請求項３】　指定された条件に適合する文書を複数の文書から検索し、
文書を分類する際に使用可能な複数の属性の種類の中から、文書を分類する際に使用する属性の種類を選択して分類属性記憶部に格納し、
前記検索した各文書から、前記分類属性記憶部に格納されている属性の種類の内の、検索要求元のユーザによって指定された属性の種類に属する属性要素を抽出し、
その抽出結果に基づいて、前記検索した各文書を、その文書が含有する属性要素毎に分類することを特徴とする文書検索方法。
【請求項４】　請求項３記載の文書検索方法において、前記検索した各文書を分類する際、前記ユーザによって指定された属性の種類に属する属性要素を含有しない文書を、独立した範疇に分類することを特徴とする文書検索方法。
【請求項５】　指定された条件に適合する文書を複数の文書から検索し、
文書を分類する際に使用可能な複数の属性の種類の中から、文書を分類する際に使用する属性の種類を選択して分類属性記憶部に格納し、
前記検索した各文書から、前記分類属性記憶部に格納されている属性の種類の内の、検索要求元のユーザによって指定された属性の種類に属する属性要素を抽出し、
前記検索した各文書を、その文書が含有する前記抽出した属性要素に対する上位概念に当たる単語であって、各単語の上位概念に当たる単語を階層的に保持するシソーラスにおいて検索要求元のユーザによって指定された階層レベルに存在する単語毎に分類することを特徴とする文書検索方法。
【請求項６】　請求項５記載の文書検索方法において、前記検索した文書を分類する際、前記ユーザによって指定された属性の種類に属する属性要素を含有しない文書を、独立した範疇に分類することを特徴とする文書検索方法。
【請求項７】　複数の文書が格納された文書記憶部と、
指定された条件に適合する文書を前記文書記憶部から検索する文書検索手段と、
文書を分類する際に使用可能な複数の属性の種類の内の、指定された属性の種類のみが格納された分類属性記憶部と、
前記文書検索手段で検索された各文書に、前記分類属性記憶部に格納されている属性の種類の内の、どの属性の種類に属する属性要素が存在するのかを解析する属性解析手段と、
前記文書検索手段で検索された各文書を、前記属性解析手段の解析結果に基づいて、その文書が含有する属性要素の種類毎に分類する属性分類手段とを備えたことを特徴とする文書検索装置。
【請求項８】　請求項７記載の文書検索装置において、
文書を分類する際に使用可能な複数の属性の種類が格納された候補記憶部と、
該候補記憶部に格納されている属性の種類の内の、文書検索装置の管理者によって指定された属性の種類のみを前記分類属性記憶部に格納する分類属性選択手段とを備えたことを特徴とする文書検索装置。
【請求項９】　請求項８記載の文書検索装置において、
前記属性分類手段は、前記分類属性記憶部に格納されている属性の種類に属する属性要素を含有しない文書を、独立した範疇に分類する構成を有することを特徴とする文書検索装置。
【請求項１０】　複数の文書が格納された文書記憶部と、
指定された条件に適合する文書を前記文書記憶部から検索する文書検索手段と、
文書を分類する際に使用可能な複数の属性の種類の内の、指定された属性の種類のみが格納された分類属性記憶部と、
前記文書検索手段が検索した各文書から、前記分類属性記憶部に格納されている属性の種類の内の、検索要求元のユーザによって指定された属性の種類に属する属性要素を抽出する属性要素抽出手段と、
前記文書検索手段が検索した各文書を、前記属性要素抽出手段の抽出結果に基づいて、文書が含有する属性要素毎に分類する属性要素分類手段とを備えたことを特徴とする文書検索装置。
【請求項１１】　請求項１０記載の文書検索装置において、
文書を分類する際に使用可能な複数の属性の種類が格納された候補記憶部と、
該候補記憶部に格納されている属性の種類の内の、文書検索装置の管理者によって指定された属性の種類のみを前記分類属性記憶部に格納する分類属性選択手段とを備

えたことを特徴とする文書検索装置。
【請求項１２】　請求項１１記載の文書検索装置において、
前記属性要素分類手段は、前記ユーザによって指定された属性の種類に属する属性要素を含有しない文書を、独立した範疇に分類する構成を有することを特徴とする文書検索装置。
【請求項１３】　複数の文書が格納された文書記憶部と、
指定された条件に適合する文書を前記文書記憶部から検索する文書検索手段と、
文書を分類する際に使用可能な複数の属性の種類の内の、指定された属性の種類のみが格納された分類属性記憶部と、
前記文書検索手段が検索した各文書から、前記分類属性記憶部に格納されている属性の種類の内の、検索要求元のユーザによって指定された属性の種類に属する属性要素を抽出する属性要素抽出手段と、
各単語の上位概念に当たる単語を階層的に保持するシソーラス記憶部と、
前記文書検索手段で検索された各文書を、その文書が含有する前記属性要素抽出手段で抽出した属性要素に対する上位概念に当たる単語であって、前記シソーラス記憶部において検索要求の要求元によって指定された階層レベルに存在する単語毎に分類する属性要素シソーラス分類手段とを備えたことを特徴とする文書検索装置。
【請求項１４】　請求項１３記載の文書検索装置において、
文書を分類する際に使用可能な複数の属性の種類が格納された候補記憶部と、
該候補記憶部に格納されている属性の種類の内の、文書検索装置の管理者によって指定された属性の種類のみを前記分類属性記憶部に格納する分類属性選択手段とを備えたことを特徴とする文書検索装置。
【請求項１５】　請求項１４記載の文書検索装置において、
前記属性要素シソーラス分類手段は、前記ユーザが指定した属性の種類に属する属性要素を含有しない文書を、独立した範疇に分類することを特徴とする文書検索装置。
【請求項１６】　コンピュータに、
指定された条件に適合する文書を複数の文書から検索する文書検索処理と、
文書を分類する際に使用可能な複数の属性の種類の中から、文書を分類する際に使用する属性の種類を選択して分類属性記憶部に格納する分類属性選択処理と、
前記検索した各文書に、前記分類属性記憶部に格納されている属性の種類の内の、どの属性の種類に属する属性要素が存在するのかを解析する属性解析処理と、
その解析結果に基づいて、前記検索した各文を、その文書が含有する属性要素の種類毎に分類する属性分類処理とを実行させるためのプログラムを記録した機械読み取り可能な記録媒体。
【請求項１７】　コンピュータに、
指定された条件に適合する文書を複数の文書から検索する文書検索処理と、
文書を分類する際に使用可能な複数の属性の種類の中から、文書を分類する際に使用する属性の種類を選択して分類属性記憶部に格納する分類属性選択処理と、
前記検索した各文書から、前記分類属性記憶部に格納されている属性の種類の内の、検索要求元のユーザによって指定された属性の種類に属する属性要素を抽出する属性要素抽出処理と、
その抽出結果に基づいて、前記検索した各文書を、その文書が含有する属性要素毎に分類する属性要素分類処理とを実行させるためのプログラムを記録した機械読み取り可能な記録媒体。
【請求項１８】　コンピュータに、
指定された条件に適合する文書を複数の文書から検索する文書検索処理と、
文書を分類する際に使用可能な複数の属性の種類の中から、文書を分類する際に使用する属性の種類を選択して分類属性記憶部に格納する分類属性選択処理と、
前記検索した各文書から、前記分類属性記憶部に格納されている属性の種類の内の、検索要求元のユーザによって指定された属性の種類に属する属性要素を抽出する属性要素抽出処理と、
前記検索した各文書を、その文書が含有する前記抽出した属性要素に対する上位概念に当たる単語であって、各単語の上位概念に当たる単語を階層的に保持するシソーラスにおいて検索要求元のユーザによって指定された階層レベルに存在する単語毎に分類する属性要素シソーラス分類処理とを実行させるためのプログラムを記録した機械読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、大量の文書から、指定した条件に適合するものを検索する文書検索技術に関し、特に、検索結果を自動的に分類して出力する文書検索技術に関する。
【０００２】
【従来の技術】従来より、大量の文書の中から目的の文書を検索する方法として、入力されたキーワードすべて、あるいは、その一部を含む文書を検索する方法がある。このような検索方法は、インターネットやパソコン通信サービスなどで公開されている各種の文書を検索するサービスや、ハードディスクに置かれた文書を検索するソフトウェアとして実現されている。しかしながら、キーワードによる条件指定では、大量の文書から、ユーザが必要とする文書だけに絞り込むことが難しく、検索

結果の中に、条件に適合しながらも不要な文書が多く混入してしまうという問題があった。インターネット上の文書を検索するサービスなどでは、キーワードを追加することで、検索結果を順次絞り込むことができるようになっているが、それでも不要な文書を完全に排除することはできない。

【０００３】この問題を解決するために、検索結果をキーワードによる条件で絞るのではなく、検索結果を他の観点から分類してユーザに提示する方法が存在する。例えば、特開平８－２３５１６０号公報や、特開平９－２３１２３８号公報に、検索結果を分類する方法が開示されている。

【０００４】特開平８－２３５１６０号公報に記載の文書検索方法及び装置では、検索結果の件数が所定値より大きい場合に、各文書に予め付与されている、文書名，文書登録日などの属性データによって検索結果を分類してユーザに提示する。

【０００５】特開平９－２３１２３８号公報に記載のテキスト検索結果表示方法及び装置では、検索結果の各テキストの主題を分析し、主題を複数のグループに分割することにより、検索結果を分類して表示する。

【０００６】一方、複数の文書を分類する方法としては、特開平１０－３２０４１１号公報に、各文書から５Ｗ１Ｈ属性付きのキーワードを抽出し、抽出された５Ｗ１Ｈ属性付きのキーワードによって文書を２次元マトリックス上に分類する方法が開示されている。

【０００７】

【発明が解決しようとする課題】従来例の文書検索方法では、ユーザに対して適切な文書の絞り込み、あるいは、適切な分類を提供できていない場合があった。

【０００８】例えば、あるホテル「Ｘホテル」に宿泊したいと考えているユーザが、「Ｘホテル」に宿泊するために必要な情報を得る目的で、「Ｘホテル」というキーワードを含む文書を検索したとする。この場合、ユーザにとって必要な情報は、「Ｘホテル」の連絡先や、「Ｘホテル」の所在地などであり、ユーザにとって必要な文書とは、それらの情報が記述されている文書である。しかしながら、「Ｘホテル」というキーワードを含むという条件だけでは、大量の文書の中から、「Ｘホテル」の連絡先や所在地が記述されている文書だけに絞り込むことができない。例えば、Ｘホテルで新製品の発表会が行われたことを報じるニュースや、Ｘホテルのレストランで食事をしたという内容の日記風のＷｅｂ文書は、この場合、ユーザにとって不要であるにもかかわらず、検索結果に含まれる。文書内に連絡先や所在地についての情報が記述されているという条件は、キーワードで表現することができないため、キーワードをさらに追加することで、検索結果を絞り込み、不要な文書を排除することはできない。

【０００９】特開平８－２３５１６０号公報に記載の文書検索方法及び装置では、文書に付けられた属性によって、検索結果を分類することができるが、分類のために必要な属性が、文書に予め付与されていなければならない。このため、予め、連絡先や所在地についての情報が存在するかどうかを文書の属性として記録しておかない限り、検索結果を、それらの情報が存在する文書と存在しない文書に分類することができない。特に、インターネットで公開されているＷｅｂ文書に対しては、対応することが困難である。

【００１０】特開平９－２３１２３８号公報に記載のテキスト検索結果表示方法及び装置では、文書の内容に応じて検索結果を分類できるが、分類の基準は、各文書の主題である。したがって、連絡先や所在地についての情報が主題として書かれている文書と、主題として書かれていない文書を分けて分類することができる。しかしながら、連絡先や所在地についての情報が主題として書かれていない文書の中にも、連絡先や所在地についての情報が書かれている場合がある。例えば、Ｘホテルが新館を増築したということを主題として伝えているニュースに、Ｘホテルの連絡先や所在地が書かれていることがありうる。したがって、この分類により、必ずしも、ユーザにとって必要な文書と不要な文書を分けられるとはいえない。

【００１１】特開平１０－３２０４１１号公報に記載の文書分類装置、方法及び文書分類プログラムを記録した記録媒体では、文書中から抽出した５Ｗ１Ｈ属性付きのキーワードによって文書を分類できるが、分類のキーとする５Ｗ１Ｈの種類は、分類を行う度にユーザが指定しなければならない。また、５Ｗ１Ｈの単位で分類するために、住所や最寄り駅、電話番号、Ｅ－ｍａｉｌアドレスのような細かい単位での分類を行うことができない。

【００１２】そこで、本発明の目的は、検索された多数の文書を、検索対象文書が属する分野に応じた観点で分類することにより、ユーザが必要とする文書を容易に選別できるようにすることにある。

【００１３】

【課題を解決するための手段】本発明の第１の文書検索装置は、特定の属性（概念）に関する具体的な内容を表す属性要素が文書中に含まれているかどうかによって、検索結果の文書を分類すると共に、上記特定の属性についての属性要素を含む文書については、上記特定の属性の種類毎に分類する。属性要素とは、具体的には文書中で、住所、電話番号、最寄り駅、価格、日付、時間、Ｅ－ｍａｉｌアドレス、ＵＲＬ、会社名、製品名、型番などの特定の属性の内容を具体的に表している要素を表す。例えば住所の属性を表す属性要素としては、「東京都千代田区」などがあり、価格の属性を表す属性要素としては、「￥１２，０００」などがある。

【００１４】本発明の第１の文書検索装置は、より具体的には、文書を分類する際に使用可能な複数の属性の種

類の内の、指定された属性の種類のみが格納された分類属性記憶部（図１の２２）と、分類属性記憶部（図１の２２）に記憶されている各属性の種類に対応する属性要素が検索結果の文書中に含まれているかどうかを解析する属性解析手段（図１の３２）と、同一種類の属性要素を含む文書どうしが、同一の範疇となり、且つ属性要素を含まない文書が独立した範疇となるように、検索結果を分類する属性分類手段（図１の３３）を有する。

【００１５】属性解析手段（図１の３２）は、検索結果の各文書を解析し、どの文書が分類属性記憶部（図１の２２）に記憶されているどの属性の種類に対応する属性要素を含むかという情報を、属性分類手段に送る（図１の３３）。属性分類手段（図１の３３）は、属性解析手段（図１の３２）より送られる情報に基づき、検索結果の各文書について、それが分類属性記憶部（図１の２２）に記憶されている何れかの種類の属性要素を含むかどうか判定し、含む場合には、その属性要素の種類に対応する範疇にその文書を分類する。分類属性記憶部（図１の２２）に記憶されている何れの種類の属性要素も含まない文書は、そのような文書に対応する１つの範疇に分類する。

【００１６】本発明の第２の文書検索装置は、特定の種類の属性要素に関して、各文書が同一の属性要素を含むかどうかによって、検索結果の文書を分類する。より具体的には、文書を分類する際に使用可能な複数の属性の種類の内の、指定された属性の種類のみが格納された分類属性記憶部（図３の２２）と、分類属性記憶部（図３の２２）に記憶されている属性の種類の内の、検索要求元のユーザによって指定された種類の属性要素を検索結果の文書から抽出する属性要素抽出手段（図３の３５）と、同一の属性要素を含む文書どうしが、同一の範疇となるように検索結果を分類する属性要素分類手段（図３の３６）を有する。

【００１７】属性要素抽出手段（図３の３５）は、分類属性記憶部（図３の２２）に記憶されている属性の種類のうち、検索要求元のユーザが指定した種類の属性要素を、検索結果の各文書から抽出し、どの文書にどの属性要素が含まれているかという情報を、属性要素分類手段（図３の３６）に送る。属性要素分類手段（図３の３６）は、属性要素抽出手段（図３の３５）より送られる情報に基づき、検索結果の各文書について、それがユーザが指定した種類の属性要素を含むかどうか判定し、含む場合には、その属性要素に対応する範疇にその文書を分類する。ユーザが指定した種類の属性要素を含まない文書は、そのような文書に対応する１つの範疇に分類する。

【００１８】本発明の第３の文書検索装置は、特定の種類の属性要素に関して、意味的に近い属性要素を含む文書を１つの範疇にまとめるように、検索結果を分類する。より具体的には、文書を分類する際に使用可能な複数の属性の種類の内の、指定された属性の種類のみが格納された分類属性記憶部（図５の２２）と、各単語の上位概念に当たる単語を保持するシソーラス記憶部（図５の２４）と、分類属性記憶部（図５の２２）に記憶されている属性の種類の内の、検索要求元のユーザが指定した種類の属性要素を検索結果の文書から抽出する属性要素抽出手段（図５の３５）と、抽出された属性要素の、ユーザが指定したレベルの上位概念に当たる単語が同一である文書どうしが、同一の範疇となるように検索結果を分類する属性要素シソーラス分類手段（図５の３７）を有する。

【００１９】シソーラス記憶部（図５の２４）は、単語間の上位概念、下位概念の関係によって、単語が階層構造をなしており、各階層に対して、絶対的なレベルが付与されている。属性要素抽出手段（図５の３５）は、分類属性記憶部（図５の２２）に記憶されている属性の種類のうち、ユーザが指定した種類の属性要素を、検索結果の各文書から抽出し、どの文書にどの属性要素が含まれているかという情報を、属性要素シソーラス分類手段（図５の３７）に送る。属性要素シソーラス分類手段（図５の３７）は、属性要素抽出手段（図５の３５）より送られる情報に基づき、検索結果の各文書について、それがユーザが指定した種類の属性要素を含むかどうか判定し、含む場合には、シソーラス記憶部（図５の２４）を参照して、その属性要素のユーザが指定したレベルの上位概念の語を求め、その上位概念の語に対応する範疇にその文書を分類する。ユーザが指定した種類の属性要素を含まない文書は、そのような文書に対応する１つの範疇に分類する。

【００２０】

【発明の実施の形態】次に本発明の実施の形態について図面を参照して詳細に説明する。

【００２１】図１を参照すると、本発明の第１の実施の形態は、ホスト側装置１と、端末装置５と、両者を接続するネットワーク６とを含む。

【００２２】端末装置５は、キーボード，マウス等の入力装置５１と、ディスプレイ装置等の出力装置５２とを含む。

【００２３】ホスト側装置１は、記憶装置２と、処理装置３と、キーボード，マウス等の入力装置４１と、ディスプレイ装置等の出力装置４２と、記録媒体Ｋ１とを含む。

【００２４】記憶装置２は、文書記憶部２１と、分類属性記憶部２２と、候補記憶部２３とを備える。

【００２５】文書記憶部２１には、或る分野に属する、検索対象となる複数の文書が格納される。候補記憶部２３には、文書を分類する際に使用可能な複数の属性の種類が格納される。分類属性記憶部２２には、検索結果の文書を分類するときに分類キーとして実際に用いる属性の種類が格納される。

【００２６】属性要素とは、住所、電話番号、最寄り駅、価格、日付、時間、Ｅ－ｍａｉｌアドレス、ＵＲＬ、会社名、製品名、型番の具体的な内容のように、文書中で特定の概念の内容を具体的に表している要素を表す。このときの、住所、電話番号などの概念の種類が属性の種類である。なお、以下では、住所の属性を表す属性要素を住所要素などと記述することにする。例えば、「Ａ社は、標準価格２，０００円で商品Ｘを発売する。」という文においては、「Ａ社」が会社名要素、「２，０００円」が価格要素、「商品Ｘ」が製品名要素である。

【００２７】分類属性記憶部２２には、候補記憶部２３に格納されている複数の属性の種類の内、分類のキーとして実際に用いるものが格納される。ここに格納する属性の種類を、文書記憶部２１に格納されている文書が属する分野に応じたものにしておくことにより、有効な分類キーだけによる分類結果を端末装置５のユーザに提示できる。例えば、飲食店情報を検索するユーザにとって、製品名、型番等が文書に含まれているかどうかは検索結果を選別する上での基準になりえない。ホスト側装置１が、文書記憶部２１に飲食店情報に関する文書を記憶するものである場合、分類属性記憶部２２に、属性の種類として住所、電話番号、最寄り駅、価格を設定しておくことで、住所、電話番号、最寄り駅、価格の観点から検索結果を分類することができる。

【００２８】処理装置３は、文書検索手段３１と、属性解析手段３２と、属性分類手段３３と、分類属性選択手段３４とを備える。

【００２９】文書検索手段３１は、端末装置５のユーザが入力装置５１を用いて入力した検索条件をネットワーク６を介して受け取り、その条件に適合する文書を文書記憶部２１から検索し、検索文書の文書名，文書番号等の識別子を属性解析手段３２と属性分類手段３３に送る。検索条件としては、例えば、１個以上のキーワードを受け取り、そのすべてのキーワード含む文書を検索するようにする。

【００３０】属性解析手段３２は、分類属性記憶部２２を参照して、分類に用いる属性の種類を読み込み、文書記憶部２１を参照して、文書検索手段３１より送られてくる識別子によって示される文書それぞれについて、分類属性記憶部２２に記憶されている属性の種類に対応する属性要素が含まれているかどうかを解析し、どの文書にどの種類の属性要素が出現したかという情報を属性分類手段３３に送る。

【００３１】属性分類手段３３は、文書検索手段３１から受け取った検索結果の文書識別子を、属性解析手段３２によって解析された、各文書の属性要素の含有状況に関する情報に従って分類し、出力装置５２に出力する。すなわち、ある種類の属性要素を含む文書は、その属性要素の種類に対応する範疇に分類し、どの属性要素も含まない文書は、そのような文書に対応する１つの範疇に分類する。２種類以上の属性要素を含む文書は、２つ以上の範疇に分類されることになる。分類の結果としては、各範疇に分類された文書のリストを出力してもよいし、各範疇に分類された文書の件数を出力してもよい。

【００３２】分類属性選択手段３４は、候補記憶部２３に格納されている属性の種類を出力装置４２に表示し、この表示を見たホスト管理者が入力装置４１を用いて選択した属性の種類のみを分類属性記憶部２２に格納する。

【００３３】なお、本実施の形態では、文書検索手段３１、属性解析手段３２、属性分類手段３３、分類属性選択手段３４は、処理装置３に備え付けられている必要はなく、コンピュータからなる処理装置３を制御するためのプログラムとして、ＣＤ－ＲＯＭやフロッピーディスクなどの記録媒体Ｋ１に格納して提供され、処理装置３に読み込まれて実行されるものとしてもよい。

【００３４】次に、図１および図２を参照して、本発明の第１の実施の形態の動作について説明する。

【００３５】ホスト管理者は、ホスト側装置１を文書検索装置として運用する際、その運用開始に先立って、分類属性選択手段３４を用いて分類属性記憶部２２に、分類キーとして用いる属性の種類を格納しておく。このときの分類属性選択手段３４の動作は、次のようになる。分類属性選択手段３４は、ホスト管理者によって起動されると、候補記憶部２３に格納されている全ての属性の種類を出力装置４２に表示する。そして、ホスト管理者は、表示された属性の種類の中から、ユーザが検索結果を選別する上で有効な分類キーになり得る属性の種類を入力装置４１を用いて選択する。例えば、文書記憶部２１に、飲食店情報に関する文書が格納されている場合には、多数表示される属性の種類の中から、住所、電話番号、最寄り駅、価格等を選択する。分類属性選択手段３４は、ホスト管理者によって選択された属性の種類を分類記憶部２２に格納する。

【００３６】次に、文書検索時の動作について説明する。文書検索手段３１は、まず、端末装置５のユーザが入力装置５１を用いて入力した検索条件を読み込む（図２，ステップＡ１）。次に、文書記憶部２１より、検索条件に適合する文書を検索し、検索条件に適合する文書すべての文書識別子を属性解析手段３２、および、属性分類手段３３に送る（ステップＡ２）。

【００３７】続いて、属性解析手段３２が、分類属性記憶部２２を参照し、分類キーとして使用する属性の種類を読み込む（ステップＡ３）。次に、属性解析手段３２は、文書記憶部２１を参照し、文書検索手段３１より文書識別子が送られてきている各文書それぞれについて、ステップＡ３で読み込んだ属性の種類の内の、どの属性の種類に属する属性要素を含んでいるかを解析し、その結果を属性分類手段３３に送る（ステップＡ４）。尚、

ステップＡ３で読み込んだ属性の種類に属する属性要素が存在しない文書については、そのことを示す解析結果を属性分類手段３３に送る。

【００３８】文書中に含まれる属性要素は、例えば、文書に対して形態素解析処理を行い、ある特定のパターンに適合する単語等の条件により検出あるいは抽出することができる。例えば、「～県～市」等のパターンに適合する単語を住所要素に、「～月～日」等のパターンに適合する単語を日付要素に、「ｈｔｔｐ：／／～」等のパターンに適合する単語をＵＲＬ要素に、「（株）～」等のパターンに適合する単語を会社名要素にすることができる。このほかにも、会社名や製品名を表す単語等を予め収集しておき、各単語とそれらを照合して、一致するものをそれぞれ会社名要素、製品名要素とする方法等もある。

【００３９】次に、属性分類手段３３が、文書検索手段３１より文書識別子が送られてきた１つの文書について、その文書が分類属性記憶部２２に格納されている属性の種類に対応する属性要素を含むかどうかを、属性解析手段３２から送られた情報に従って判定する（ステップＡ５）。そして、その文書が何れかの属性要素を含んでいる場合には、その属性要素の種類に対応する範疇にその文書を分類する。複数の属性要素を含む場合には、複数の範疇に分類する（ステップＡ６）。これに対して、その文書が何れの属性要素も含んでいない場合には、属性要素を含まない文書の範疇にその文書を分類する（ステップＡ７）。

【００４０】属性分類手段３３は、文書検索手段３１で検索されたすべての文書について分類が完了したかどうかを確認し（ステップＡ８）、完了していれば、分類の結果を出力して処理を終了する（ステップＡ９）。完了していなければ、ステップＡ５に戻って処理を繰り返す。

【００４１】次に、本発明の第１の実施の形態の効果について説明する。

【００４２】本発明の第１の実施の形態は、検索結果の文書を、それに含まれる属性要素の種類に従って分類する。また、属性要素を含まない文書を、１つの独立した範疇に分類する。これにより、検索結果の中から、住所、電話番号、最寄り駅、価格、日付、時間、Ｅ－ｍａｉｌアドレス、ＵＲＬ、会社名、製品名、型番などの属性が記述されている文書だけを選択することができるようになる。

【００４３】また、本発明の第１の実施の形態は、候補記憶部２３に格納されている属性の種類の内のいくつかを分類属性選択手段３４を用いて分類属性記憶部２２に格納しておき、分類属性記憶部２２に格納されている属性の種類だけを用いて検索結果の文書の分類を行う。これにより、分類属性記憶部２２に設定する属性の種類を検索対象の文書に合ったものにすることができるため、検索対象の文書に合わせた分類の観点で検索結果を分類することができるようになる。検索対象の文書に合わせて、分類に用いる属性の種類を設定すれば、ユーザは、有効な分類キーのみにより分類された結果を得ることができる。

【００４４】次に、本発明の第２の実施の形態について、図面を参照して詳細に説明する。

【００４５】図３を参照すると、本発明の第２の実施の形態は、ホスト側装置１ａの処理装置３ａが、図１に示された第１の実施の形態の処理装置３の構成における属性解析手段３２に代わり属性要素抽出手段３５を有し、属性分類手段３３に代わり属性要素分類手段３６を有する点で異なる。

【００４６】属性要素抽出手段３５は、分類属性記憶部２２を参照して、分類に用いる属性の種類を読み込み、その中からユーザが指定する属性の種類を、入力装置５１，ネットワーク６を通して受け取り、文書記憶部２１を参照して、文書検索手段３１より送られてくる検索結果の各文書から、その文書に含まれる属性要素のうち、ユーザによって指定された種類のものを抽出する。さらに、どの文書からどの属性要素が抽出されたかという情報を属性要素分類手段３６に送る。

【００４７】属性要素分類手段３６は、文書検索手段３１から受け取った検索結果の文書を、属性要素抽出手段３５によって抽出された、各文書中の特定の種類の属性要素に従って、同一の属性要素を含むものが同一の範疇になるように分類し、出力装置５２に出力する。なお、属性要素抽出手段３５によってユーザが指定した種類の属性要素が抽出されなかった文書は、そのような文書に対応する１つの範疇に分類する。２つ以上の異なる属性要素を含む文書は、２つ以上の範疇に分類されることになる。第１の実施の形態と同様に、分類の結果としては、各範疇に分類された文書のリストを出力してもよいし、各範疇に分類された文書の件数を出力してもよい。

【００４８】なお、本実施の形態では、文書検索手段３１、分類属性選択手段３４、属性要素抽出手段３５、属性要素分類手段３６は、処理装置３ａに備え付けられている必要はなく、コンピュータからなる処理装置３ａを制御するためのプログラムとして、ＣＤ－ＲＯＭやフロッピーディスクなどの記録媒体Ｋ２に格納して提供され、処理装置３ａに読み込まれて実行されるものとしてもよい。

【００４９】次に、図３および図４を参照して、本発明の第２の実施の形態の文書検索時の動作について説明する。

【００５０】図４のステップＡ１、Ａ２で示される、第２の実施の形態における文書検索手段３１の動作は、第１の実施の形態における文書検索手段３１の動作と同一のため、説明は省略する。

【００５１】文書検索手段３１でステップＡ１、Ａ２の

処理が行われた後、属性要素抽出手段３５は、分類属性記憶部２２に格納されている属性の種類を全て読み出し、それを検索要求元の端末装置５へ送る（ステップＢ１）。

【００５２】端末装置５の出力装置５２は、送られてきた属性の種類を表示し、その表示を見た端末装置５のユーザは、入力装置５１を用いて、表示されている属性の種類の中から、文書中に含まれている必要がある属性の種類を選択する。選択された属性の種類は、ネットワーク６を介してホスト側装置１ａに送られる。

【００５３】属性要素抽出手段３５は、ネットワーク６を介して送られてくるユーザが選択した属性の種類を読み込むと（ステップＢ２）、文書記憶部２１を参照し、文書検索手段３１から文書識別子が送られてきている各文書について、その文書中に含まれている、ステップＢ２で指定された種類の属性要素を抽出し、その結果を属性要素分類手段３６に送る（ステップＢ３）。

【００５４】次に、属性要素分類手段３６が、文書検索手段３１から文書識別子が送られてきている１つの文書について、その文書からユーザが指定した種類の属性要素が抽出されたかどうかを、属性要素抽出手段３５から送られた情報に従って判定する（ステップＢ４）。そして、その文書がユーザが指定した種類の属性要素を含んでいる場合には、その属性要素に対応する範疇にその文書を分類する。複数の属性要素を含む場合には、複数の範疇に分類する（ステップＢ５）。これに対して、その文書がユーザが指定した種類の属性要素を１つも含んでいない場合には、属性要素を含まない文書の範疇にその文書を分類する（ステップＢ６）。

【００５５】属性要素分類手段３６は、すべての文書について分類が完了したかどうかを確認し（ステップＢ７）、完了していれば、分類の結果を出力して処理を終了する（ステップＢ８）。完了していなければ、ステップＢ４に戻って処理を繰り返す。

【００５６】次に、本発明の第２の実施の形態の効果について説明する。

【００５７】本発明の第２の実施の形態は、文書中に出現する、ユーザが指定した種類の属性要素ごとに、検索結果の文書を分類する。検索結果が、文書中に記述されている、ユーザーが指定した種類の属性要素に従って分類されるため、ある特定の種類の属性が記述されている文書が必要なユーザは、自分が必要としている種類の属性に関する文書中の内容によって分類された検索結果を得ることができ、検索結果を絞り込むことが容易になる。

【００５８】次に、本発明の第３の実施の形態について、図面を参照して詳細に説明する。

【００５９】図５を参照すると、本発明の第３の実施の形態は、ホスト側装置１ｂの記憶装置２ｂが、図３に示された第２の実施の形態の記憶装置２の構成に加え、シソーラス記憶部２４を備え、処理装置３ｂが、図３に示された第２の実施の形態の処理装置３ａの構成における属性要素分類手段３６に代わり、属性要素シソーラス分類手段３７を有する点で異なる。

【００６０】シソーラス記憶部２４は、各単語に対して、その上位概念に相当する単語にリンクを張ったシソーラスを記憶している。例えば、Ａ社、Ｂ社、Ｃ社が、何れも電機メーカーである場合、シソーラスでは、「Ａ社」、「Ｂ社」、「Ｃ社」という語から、それらの共通の上位概念に相当する「電機メーカー」という単語にリンクを張るようにすることができる。

【００６１】ある単語の上位概念に当たる単語に対して、さらにその上位概念に当たる単語が再帰的に存在しうるため、単語とその上位概念に当たる単語との関係は、階層構造を成している。シソーラス記憶部２４では、その各階層に対し、絶対的なレベルが付与されている。例えば、「港区」の上位概念として「東京都」が、「東京都」の上位概念として「日本」があるとき、「港区」のレベルを２、「東京都」のレベルを１、「日本」のレベルを０などとする。

【００６２】属性要素シソーラス分類手段３７は、入力装置５１を通してユーザが指定するレベルを読み取り、文書検索手段３１から受け取った検索結果の文書を、属性要素抽出手段３５によって抽出された、各文書中の特定の種類の属性要素に従い、その属性要素に対する指定したレベルの上位概念の単語が同一のものが同一の範疇になるように分類し、出力装置５２に出力する。例えば、「ＰＣ－ＡＢＣ」と「ＸＹＺ－ＰＣ」の上位概念が共に「パソコン」であり、「ＰＣ－ＡＢＣ」、「ＸＹＺ－ＰＣ」のレベルが１、「パソコン」のレベルが０であるとき、ユーザによって指定されたレベルが１であれば、「ＰＣ－ＡＢＣ」を含む文書と「ＸＹＺ－ＰＣ」を含む文書が異なる範疇に分類されるが、ユーザによって指定されたレベルが０であれば、「ＰＣ－ＡＢＣ」を含む文書と「ＸＹＺ－ＰＣ」を含む文書は同一の範疇に分類される。

【００６３】属性要素シソーラス分類手段３７は、属性要素抽出手段３５によってユーザが指定した種類の属性要素が何も抽出されなかった文書は、そのような文書に対応する１つの範疇に分類する。一方、２つ以上の異なる属性要素を含む文書は、２つ以上の範疇に分類されることもある。分類の結果としては、第２の実施の形態と同様に、各範疇に分類された文書のリストを出力してもよいし、各範疇に分類された文書の件数を出力してもよい。

【００６４】属性要素シソーラス分類手段３７が、各単語の指定されたレベルの上位概念の単語を求める際には、シソーラス記憶部２４を参照する。ただし、ユーザが指定した属性要素の種類が、日付、時間、価格等の値を表す属性の場合には、上位概念として一定の範囲の値

を採用してもよい。例えば、１９９９年６月２０日の上位概念として１９９９年の６月１カ月間を採用し、さらにその上位概念として１９９９年１年間を採用することができる。どの範囲の値をどのレベルの階層にするかは予め定めておけばよい。このように上位概念を定義する場合は、シソーラス記憶部２２を参照する必要はない。

【００６５】なお、本実施の形態では、文書検索手段３１、分類属性選択手段３４、属性要素抽出手段３５、属性要素シソーラス分類手段３７は、処理装置３ｂに備え付けられている必要はなく、コンピュータからなる処理装置３ｂを制御するためのプログラムとして、ＣＤ－ＲＯＭやフロッピーディスクなどの記録媒体Ｋ３に格納して提供され、処理装置３ｂに読み込まれて実行されるものとしてもよい。

【００６６】次に、図５および図６を参照して、本発明の第３の実施の形態の動作について説明する。

【００６７】図６のステップＡ１、Ａ２、Ｂ１、Ｂ２、Ｂ３で示される、第３の実施の形態における文書検索手段３１、および、属性要素抽出手段３５の動作は、第２の実施の形態における文書検索手段３１、および、属性要素抽出手段３５の動作と同一のため、説明は省略する。

【００６８】ステップＡ１、Ａ２、Ｂ１、Ｂ２、Ｂ３の処理の後、属性要素シソーラス分類手段３７が、入力装置５１を通して、ユーザが指定する、分類時のシソーラス中での概念のレベルを読み込む（ステップＣ１）。続いて、属性要素シソーラス分類手段３７は、文書検索手段３１から文書識別子が送られてきている１つの文書について、その文書からユーザが指定した種類の属性要素が抽出されたかどうかを、属性要素抽出手段３５から送られてきた情報に従って判定する（ステップＣ２）。

【００６９】もし、その文書がユーザが指定した種類の属性要素を含んでいる場合には、その文書が含むすべての属性要素について、シソーラス記憶部２４を参照し、ユーザが指定するレベルの上位概念に当たる単語を求める（ステップＣ３）。そして、求めた上位概念に対応する範疇にその文書を分類する。その文書が複数の属性要素を含む場合には、求められた上位概念が複数になることがあるが、その場合には、文書を複数の範疇に分類する（ステップＣ４）。

【００７０】これに対して、その文書がユーザが指定した種類の属性要素を１つも含んでいない場合には、属性要素を含まない文書の範疇にその文書を分類する（ステップＣ５）。

【００７１】属性要素シソーラス分類手段３７は、すべての文書について分類が完了したかどうかを確認し（ステップＣ６）、完了していれば、分類の結果を出力装置５２出力して処理を終了する（ステップＣ７）。完了していなければ、ステップＣ２に戻って処理を繰り返す。

【００７２】次に、本発明の第３の実施の形態の効果について説明する。

【００７３】本発明の第３の実施の形態は、文書中に出現する、ユーザが指定した種類の属性要素に応じて検索結果の文書を分類する。特に、このとき、シソーラス中のユーザが指定したレベルにおいて同一の概念に対応する属性要素を含む文書どうしが、同一の範疇に分類されるようにする。このため、検索結果が多く、検索結果の文書に含まれる、ユーザが指定した種類の属性要素が多岐に渡る場合でも、ユーザが適当なシソーラスのレベルを指定することにより、範疇の数をより少なくすることができる。また、ユーザがシソーラスのレベルを自由に設定できるため、ユーザが必要とするレベルでの分類を提供することが可能になる。

【００７４】なお、本発明の第２、第３の実施の形態において、属性要素ごと、あるいは、その上位概念の単語ごとに範疇を設けることにより、範疇の数が多くなりすぎる場合には、代表的な属性要素、あるいは、上位概念の単語についてのみ範疇を設け、それに対応しない属性要素を有する文書は「その他」の範疇に分類し、ユーザから指定された場合に、「その他」に分類された文書を対象に再帰的に分類を行うようにしてもよい。

【００７５】また、本発明の第３の実施の形態において、ユーザが指定したシソーラスのレベルでの分類結果に対して、ユーザから指定された場合に、その範疇の１つに対して、その範疇に分類された文書を、現在とは異なるシソーラスのレベルで再帰的に分類するようにしてもよい。このとき、再帰的に分類を行う際のシソーラスのレベルは、再度ユーザが指定するようにしてもよいし、前回よりも１つ上、あるいは、１つ下のレベルを採用することにしてもよい。

【００７６】〔実施例〕本発明の第一の実施の形態の一実施例の動作を詳細に説明する。

【００７７】例えば、本発明の文書検索装置を用いて「焼肉」というキーワードを含む文書を検索するものとする。

【００７８】なお、この装置が分類に用いる属性の種類（分類属性記憶部２２に格納されている属性の種類）は、「住所」と「価格」のみであるものとして動作を説明する。

【００７９】まず、文書検索手段３１が、入力装置５１から入力されたキーワード「焼肉」を読み込む（ステップＡ１）。そして、文書記憶部２１から、キーワード「焼肉」を含む文書を検索する（ステップＡ２）。これにより、図７に示す文書識別子＃１～＃１０の１０件の文書が検索結果として得られたとする。

【００８０】属性解析手段３２は、まず、分類属性記憶部２２を参照して、分類に用いる属性の種類「住所」，「価格」を読み込む（ステップＡ３）。属性解析手段３２は、次に、検索結果の文書＃１～＃１０のすべてに対して、ステップＡ３で読み込んだ各種類の属性要素がそ

れぞれの文書中に含まれているかを解析する（ステップＡ４）。今、ステップＡ３では、「住所」，「価格」の２種類の属性が読み込まれているため、属性解析手段３２は、各文書中に住所要素と価格要素が含まれているかどうかを検査する。ここで、属性解析手段３２が、「～都」、「～道」、「～府」、「～県」というパターンに当てはまる単語を住所要素とみなし、「～円」というパターンに当てはまる単語を価格要素とみなすものとする。文書＃１～＃１０に含まれるこれらのパターンに当てはまる単語を表にすると図８のようになっている。従って、属性解析手段３２は、文書＃１、＃４、＃５、＃６、＃７、＃１０の６つが住所要素を含み、文書＃１、＃３、＃４の３つが価格要素を含み、文書＃２、＃８、＃９の３つはステップＡ３で読み込んだ種類の属性要素を含まないと判断する。

【００８１】属性分類手段３３は、この結果を受け取り、検索結果の文書＃１～＃１０を分類する。まず、文書＃１について、それが分類属性記憶部２２に記憶されている種類の属性要素を含むかどうかを判定する（ステップＡ５）。この例では、文書＃１は、住所要素と価格要素を含んでいるため、属性要素を含むと判定し、文書＃１を住所要素を含む文書の範疇と価格要素を含む文書の範疇の２つの範疇に分類する（ステップＡ６）。

【００８２】次に、属性分類手段３３は、すべての文書について分類が完了したかどうかを判定する（ステップＡ８）。この例では、まだ９文書残っているので、完了していないと判定し、ステップＡ５に戻って、次の文書の処理を行う。

【００８３】文書＃２は、分類属性記憶部２２に記憶されている種類のどの属性要素も含んでいないため、属性分類手段３３は、文書＃２を属性要素を含まない文書と判定し（ステップＡ５）、属性要素を含まない文書の範疇に文書＃２を分類する（ステップＡ７）。

【００８４】以下、文書＃３～＃１０について、ステップＡ５、Ａ６、もしくは、ステップＡ５、Ａ７の処理を繰り返し、最終的な分類結果を出力して、処理を終了する（ステップＡ９）。出力結果は、例えば、図９のようになる。

【００８５】次に、本発明の第２の実施の形態の一実施例の動作を詳細に説明する。

【００８６】この例でも、本発明の文書検索装置を用いて「焼肉」というキーワードを含む文書を検索するものとする。また、分類属性記憶部２２には、属性の種類として「住所」，「価格」が格納されているとする。本実施の形態の文書検索手段３１のステップＡ１、Ａ２の動作は、第１の実施の形態の文書検索手段３１の動作と同一のため、説明は省略する。また、ステップＡ２における文書の検索で、図７に示す文書識別子＃１～＃１０の１０件の文書が検索結果として得られたとする。

【００８７】文書検索手段３１でステップＡ１、Ａ２の処理が行われた後、属性要素抽出手段３５は、分類属性記憶部２２に格納されている全ての属性の種類「住所」，「価格」を読み出し、それを検索要求元の端末装置５へ送る（ステップＢ１）。端末装置５の出力装置５２は、属性の種類「住所」，「価格」が送られてくると、それらを表示する。この表示を見た端末装置５のユーザは、入力装置５１を用いて、表示されている属性の種類「住所」，「価格」の中から、文書中に含まれている必要がある属性の種類を指定する。今、例えば、「住所」を指定したとすると、それがネットワーク６を介してホスト側装置１ａに送られる。

【００８８】ホスト側装置１ａ内の属性要素抽出手段３５は、端末装置５から送られてくるユーザが指定した属性の種類「住所」を読み込むと（ステップＢ２）、検索結果の文書＃１～＃１０のすべてに対して、それぞれの文書中に含まれる住所要素を抽出する（ステップＢ３）。今、属性要素抽出手段３５が、第一の実施の形態の実施例における属性解析手段３２と同様の基準で、住所要素の判定を行うとすると、図１０に示す住所要素のみがステップＢ３で抽出される。

【００８９】属性要素分類手段３６は、この結果を受け取り、検索結果の文書＃１～＃１０を分類する。まず、文書＃１について、その文書から住所要素が抽出されたかどうかを判定する（ステップＢ４）。この例では、文書＃１からは、「東京都」という住所要素が抽出されているため、住所要素が抽出されたと判定し、文書＃１を「東京都」という範疇に分類する（ステップＢ５）。

【００９０】次に、属性要素分類手段３６は、すべての文書＃１～＃１０について分類が完了したかどうかを判定する（ステップＢ７）。この例では、まだ９文書残っているので、完了していないと判定し、ステップＢ４に戻って、次の文書の処理を行う。

【００９１】文書＃２は、住所要素を含んでいないため、属性要素分類手段３６は、文書＃２からは住所要素が抽出されなかったと判定し（ステップＢ４）、住所要素を含まない文書の範疇に文書＃２を分類する（ステップＢ６）。

【００９２】以下、文書＃３～＃１０について、ステップＢ４，Ｂ５、もしくは、ステップＢ４，Ｂ６の処理を繰り返し、最終的な分類結果を出力して、処理を終了する（ステップＢ８）。出力結果は、例えば、図１１のようになる。

【００９３】次に、本発明の第３の実施の形態の一実施例の動作を詳細に説明する。

【００９４】この例で、シソーラス記憶部２４は、地名に関するシソーラスとして、図１２に示すような階層関係を保持しているとする。図１２の例は、例えば、「東京都」、「神奈川県」、「千葉県」、「埼玉県」等の上位概念を表す単語が「関東地方」であり、「関東地方」、「近畿地方」等の上位概念を表す単語が「日本」

であることを示している。また、「東京都」、「神奈川県」、「千葉県」、「埼玉県」等のレベルが２、「関東地方」、「近畿地方」等のレベルが１、「日本」のレベルが０となっている。

【００９５】このとき、本発明の文書検索装置を用いて「焼肉」というキーワードを含む文書を検索するものとする。本実施の形態の文書検索手段３１のステップＡ１、Ａ２の動作、および、属性要素抽出手段３５のステップＢ１、Ｂ２、Ｂ３の動作は、第２の実施の形態の文書検索手段３１、および、属性要素抽出手段３５の動作と同一のため、説明は省略する。ここでは、ステップＡ１、Ａ２、Ｂ１、Ｂ２、Ｂ３において、第２の実施の形態の一実施例の説明と同様に処理が進んだものとし、その続きの処理について説明する。すなわち、ステップＡ１の処理で、図７に示す１０件の文書＃１～＃１０が検索結果として得られ、ステップＢ２の処理で、ユーザが指定する属性要素の種類として「住所」を読み込み、ステップＢ３において、図１０に示す住所要素が各文書から抽出されている。

【００９６】属性要素シソーラス分類手段３７は、まず、ユーザが入力装置５１を用いて指定したシソーラスのレベルを読み込む（ステップＣ１）。ここでは、ユーザがシソーラスのレベルとして「１」を指定したとする。

【００９７】続いて、属性要素シソーラス分類手段３７は、属性要素抽出手段３５による属性要素の抽出結果を受け取り、検索結果の文書＃１～＃１０を分類する。まず、文書＃１について、その文書から住所要素が抽出されたかどうかを判定する（ステップＣ２）。この例では、文書＃１からは、「東京都」という住所要素が抽出されているため、住所要素が抽出されたと判定し、シソーラス記憶部２４を参照して、「東京都」の上位概念にあたる単語で、ユーザより指定されたレベル１にあるものを求める（ステップＣ３）。この例では、ステップＣ３において、「関東地方」という単語が求められ、属性要素シソーラス分類手段３７は、「関東地方」という範疇に文書＃１を分類する（ステップＣ４）。

【００９８】次に、属性要素シソーラス分類手段３７は、すべての文書＃１～＃１０について分類が完了したかどうかを判定する（ステップＣ６）。この例では、まだ9文書残っているので、完了していないと判定し、ステップＣ２に戻って、次の文書の処理を行う。

【００９９】文書＃２は、住所要素を含んでいないため、属性要素シソーラス分類手段３７は、文書＃２からは住所要素が抽出されなかったと判定し（ステップＣ２）、住所要素を含まない文書の範疇に文書＃２を分類する（ステップＣ５）。

【０１００】以下、文書＃３～＃１０について、ステップＣ２，Ｃ３，Ｃ４もしくは、ステップＣ２，Ｃ５の処理を繰り返し、最終的な分類結果を出力して、処理を終了する（ステップＣ７）。出力結果は、例えば、図１３のようになる。

【０１０１】

【発明の効果】本発明の第１の効果は、検索要求元のユーザが、検索された多数の文書の中から、必要としている種類の属性要素が含有されている文書を、容易に選択することが可能になるという点である。

【０１０２】その理由は、住所、電話番号、最寄り駅、価格、日付、時間、Ｅ－ｍａｉｌアドレス、ＵＲＬ、会社名、製品名、型番など多数存在する、分類キーとなり得る属性の種類の中から、検索された文書を分類する際に実際に使用する属性の種類を選択して分類属性記憶部に格納しておき、この分類属性記憶部に格納されている属性の種類のみを用いて検索された文書を分類するようにしているからである。つまり、検索対象とする文書が属する分野毎に、有効な分類キー（分類の観点）が異なるものとなるので、従来技術のように、分類キーを５Ｗ１Ｈに固定して分類を行うと、ユーザが選別しやすい形で分類が行われない場合があるが、本発明によれば、多数の属性の種類の中から、検索対象とする文書が属する分野に応じた属性の種類を選択して分類キーにすることができるので、ユーザが選別しやすい形で分類を行うことができる。

【０１０３】本発明の第２の効果は、検索結果を、着目すべき属性要素が含まれる文書と含まれない文書に分けることができることである。この結果、着目すべき属性が記述されていない文書が不要な場合には、その不要な文書を検索結果から容易に除外することができるようになる。

【０１０４】その理由は、属性解析手段が、分類属性記憶部に記憶されている各種類の属性について、検索結果の各文書にどの種類の属性要素が含まれているかを解析し、属性分類手段が、分類属性記憶部に記憶されている種類の属性要素が含まれていない文書を独立した範疇に分類するからである。

【０１０５】第３の効果は、文書中の特定の種類の属性要素に従って、検索結果の文書を分類できることである。この結果、ある特定の種類の属性が記述されている文書が必要なユーザは、その具体的な内容、すなわち、自分にとって必要な事項に対応する内容によって分類された検索結果を得ることができるようになる。これにより、検索結果をさらに絞り込むことが容易になる。

【０１０６】その理由は、属性要素抽出手段が、ユーザが指定した種類の属性要素を検索結果の各文書中から抽出し、属性要素分類手段が、同一の属性要素を含む文書どうしが、同一の範疇となるように検索結果を分類するからである。

【０１０７】第４の効果は、検索結果を分類する際の範疇の区分が詳細になりすぎないように、意味的に近い属性要素を含む文書を１つの範疇にまとめて分類できるこ

とである。この結果、まとめ上げるレベルを指定することで、ユーザは、必要とする詳細度の分類結果を得ることができる。

【０１０８】その理由は、シソーラス記憶部が、各単語の上位概念に当たる単語を保持しており、属性要素シソーラス分類手段が、各文書から抽出された属性要素から、ユーザが指定したレベルの上位概念の単語を求め、求められた単語が同一になる文書どうしが、同一の範疇となるように検索結果を分類するからである。

【０１０９】第５の効果は、検索結果を分類する際の範疇が多くなりすぎないように、範疇の数を抑えることが可能なことである。

【０１１０】その理由は、第４の効果の理由と同一である。すなわち、シソーラスを参照して、意味的に近い属性要素を含む文書を１つの範疇にまとめて分類することで、範疇の数を減らすことができるからである。

【図面の簡単な説明】

【図１】本発明の第１の実施の形態の構成例を表すブロック図である。

【図２】本発明の第１の実施の形態の動作例を表すフローチャートである。

【図３】本発明の第２の実施の形態の構成例を表すブロック図である。

【図４】本発明の第２の実施の形態の動作例を表すフローチャートである。

【図５】本発明の第３の実施の形態の構成例を表すブロック図である。

【図６】本発明の第３の実施の形態の動作例を表すフローチャートである。

【図７】本発明の実施の形態の実施例において、文書検索手段３１によって検索された検索結果の文書を示す図である。

【図８】本発明の第１の実施の形態の実施例において、検索結果の各文書中に含まれている属性要素を示す図である。

【図９】本発明の第１の実施の形態の実施例において、出力される結果を示す図である。

【図１０】本発明の第２、および、第３の実施の形態の実施例において、検索結果の各文書中から抽出された住所要素を示す図である。

【図１１】本発明の第２の実施の形態の実施例において、出力される結果を示す図である。

【図１２】本発明の第３の実施の形態の実施例におけるシソーラス記憶部２４の内容を示す図である。

【図１３】本発明の第３の実施の形態の実施例において、出力される結果を示す図である。

【符号の説明】

１，１ａ．１ｂ…ホスト側装置
２，２ｂ…記憶装置
２１…文書記憶部
２２…分類属性記憶部
２３…候補記憶部
２４…シソーラス記憶部
３，３ａ，３ｂ…処理装置
３１…文書検索手段
３２…属性解析手段
３３…属性分類手段
３４…分類属性選択手段
３５…属性要素抽出手段
３６…属性要素分類手段
３７…属性要素シソーラス分類手段
４１…入力装置
４２…出力装置
５…端末装置
５１…入力装置
５２…出力装置
６…ネットワーク
Ｋ１，Ｋ２，Ｋ３…記録媒体

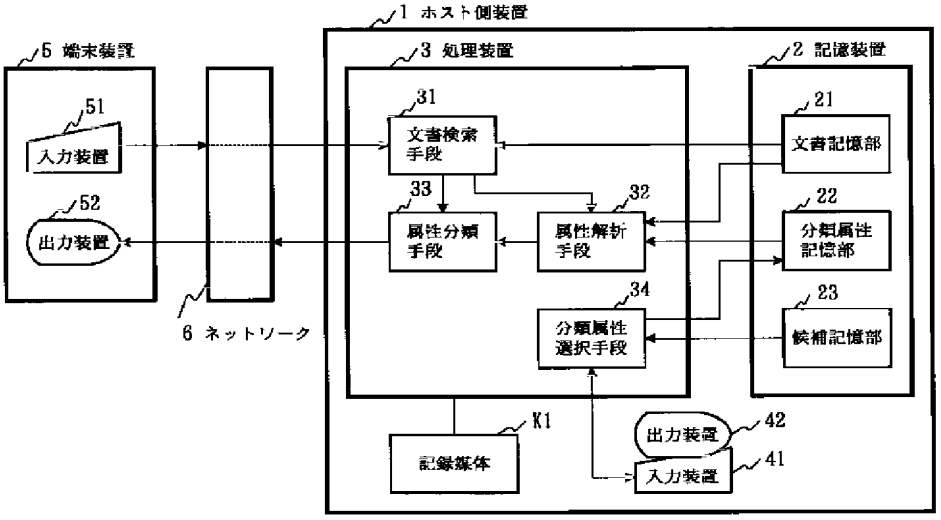
【図９】

住所要素を含む文書	6件	＃1，＃4，＃5，＃6，＃7，＃10
価格要素を含む文書	3件	＃1，＃3，＃4
属性要素を含まない文書	3件	＃2，＃8，＃9

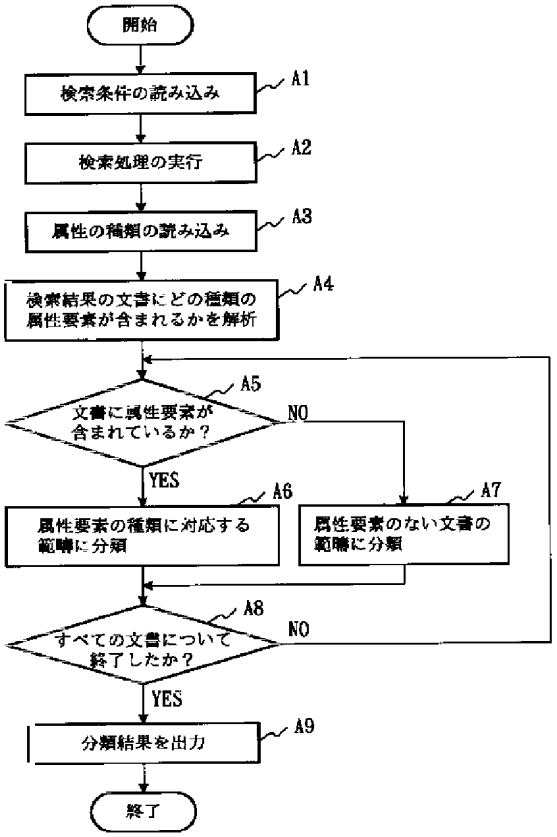
【図１１】

東京都	2件	＃1，＃10
大阪府	2件	＃4，＃6
神奈川県	2件	＃5，＃10
京都府	1件	＃7
属性要素なし	3件	＃2，＃8，＃9

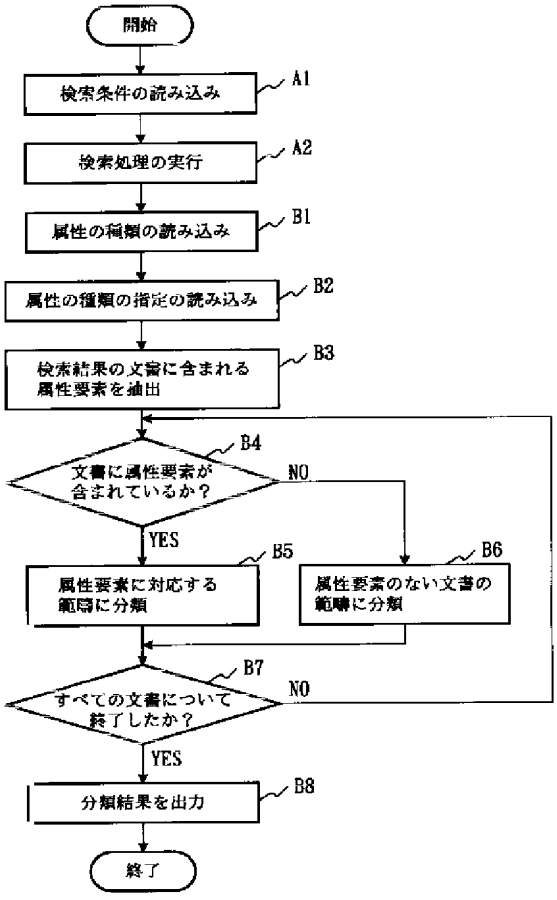
【図1】


1 ホスト側装置
5 端末装置
3 処理装置
2 記憶装置
51
入力装置
52
出力装置
31
文書検索手段
21
文書記憶部
33
属性分類手段
32
属性解析手段
22
分類属性記憶部
34
分類属性選択手段
23
候補記憶部
6 ネットソーク
K1
記録媒体
出力装置
42
入力装置
41

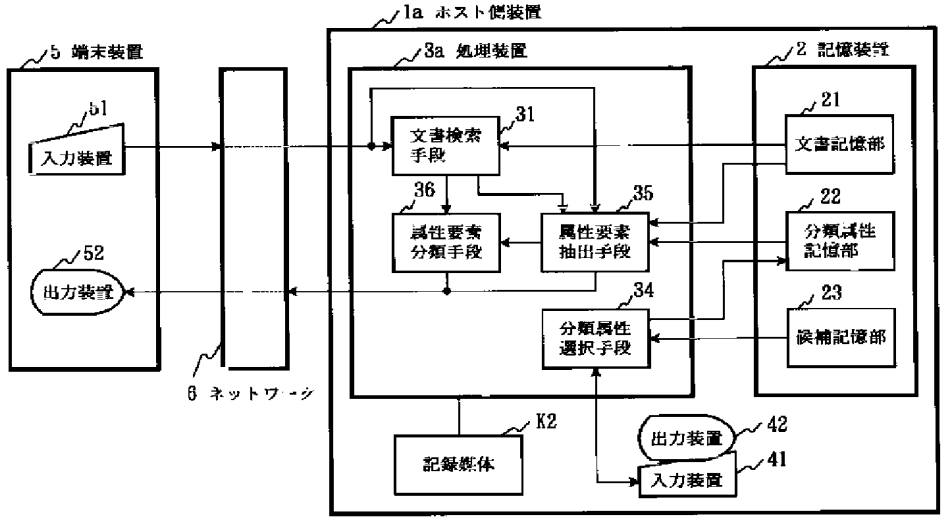
【図2】


開始
検索条件の読み込み
A1
検索処理の実行
A2
属性の種類の読み込み
A3
検索結果の文書にどの種類の属性要素が含まれるかを解析
A4
文書に属性要素が含まれているか？
A5
NO
YES
属性要素の種類に対応する範疇に分類
A6
属性要素のない文書の範疇に分類
A7
すべての文書について終了したか？
A8
NO
YES
分類結果を出力
A9
終了

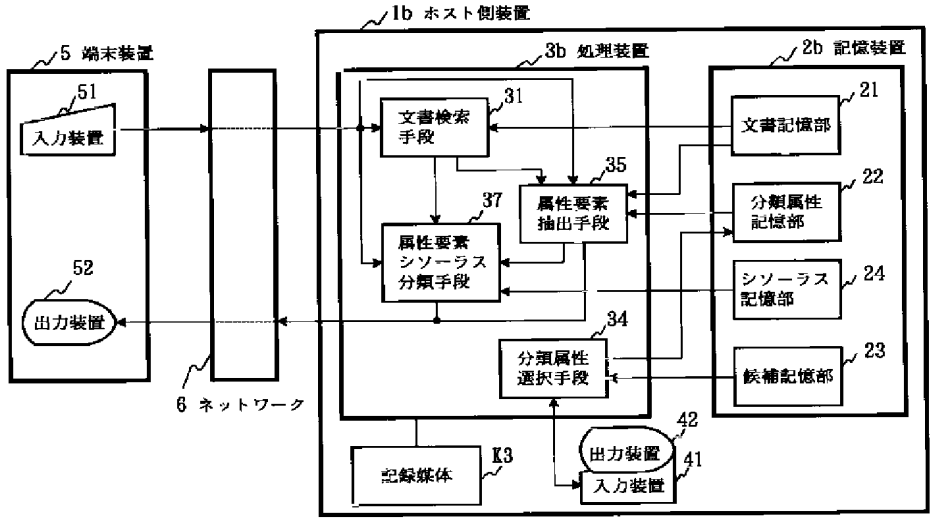
【図4】


開始
検索条件の読み込み
A1
検索処理の実行
A2
属性の種類の読み込み
B1
属性の種類の指定の読み込み
B2
検索結果の文書に含まれる属性要素を抽出
B3
文書に属性要素が含まれているか？
B4
NO
YES
属性要素に対応する範疇に分類
B5
属性要素のない文書の範疇に分類
B6
すべての文書について終了したか？
B7
NO
YES
分類結果を出力
B8
終了

【図３】



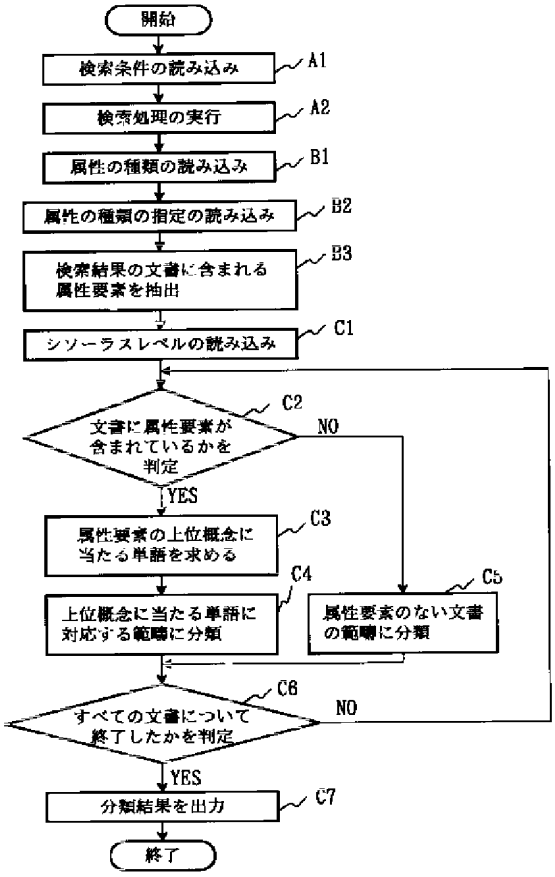
【図５】



【図１３】

関東地方	3件	＃1，＃5，＃10
近畿地方	3件	＃4，＃6，＃7
属性要素なし	3件	＃2，＃8，＃9

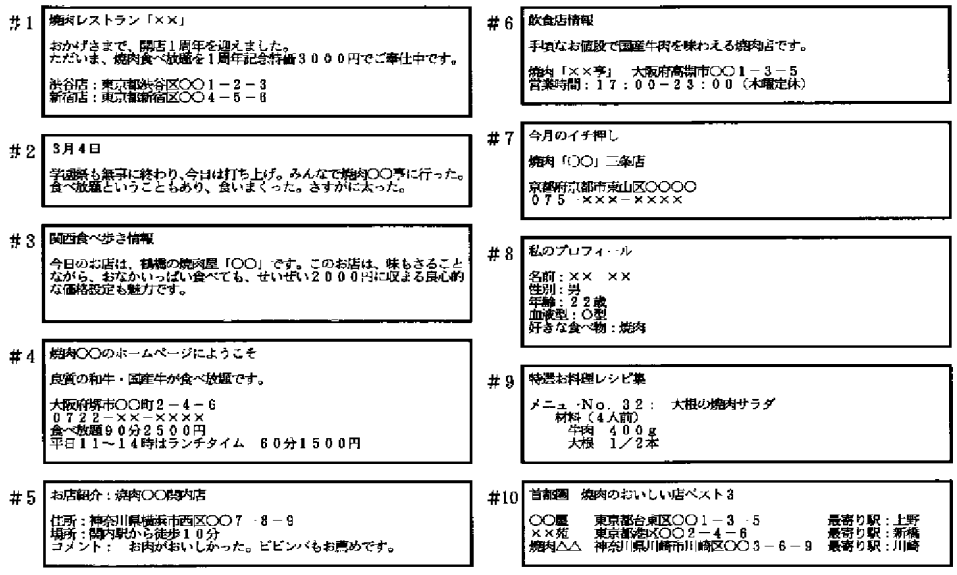
【図6】



【図8】

文書番号	属性要素	属性の種類
＃1	３０００円 東京都 東京都	価格 住所 住所
＃3	２０００円	価格
＃4	大阪府 ２５００円 １５００円	住所 価格 価格
＃5	神奈川県	住所
＃6	大阪府	住所
＃7	京都府	住所
＃10	東京都 東京都 神奈川県	住所 住所 住所

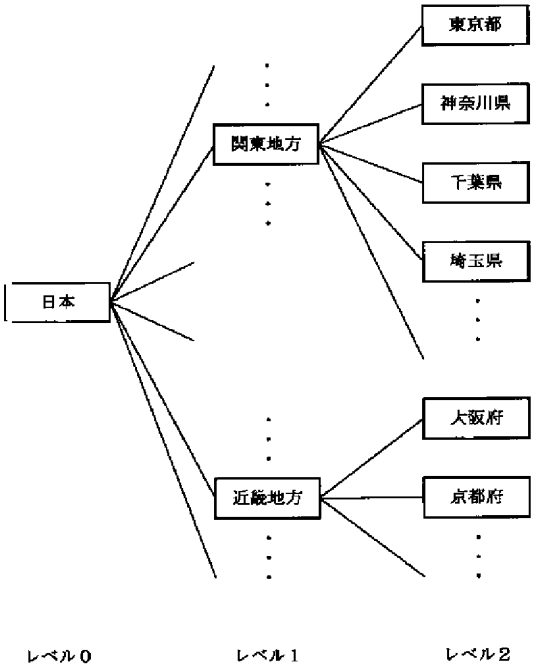
【図7】



【図１０】

文書番号	住所要素
＃１	東京都 東京都
＃４	大阪府
＃５	神奈川県
＃６	大阪府
＃７	京都府
＃１０	東京都 東京都 神奈川県

【図１２】



フロントページの続き

(72)発明者　奥村　明俊
東京都港区芝五丁目７番１号　日本電気株式会社内

Ｆターム(参考)　5B009 SA00 SA12 SA14 VA02 VA09
5B075 KK07 KK40 ND20 NK02 NR02
NR12 PP02 PP03 PP12 PP22
PQ02 PQ20 QS20 UU40

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-067419

(43)Date of publication of application : 07.03.2003

(51)Int.Cl. G06F 17/30
G06F 12/00

(21)Application number : 2001-254772

(22)Date of filing : 24.08.2001

(71)Applicant : TOSHIBA CORP

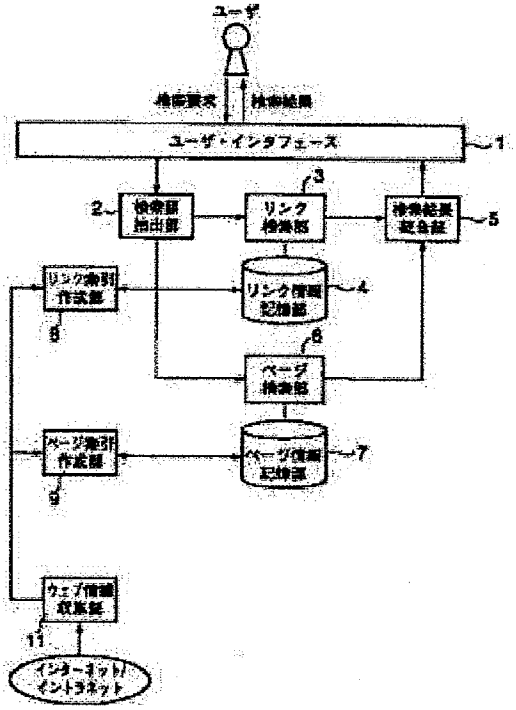
(72)Inventor : GOTO KAZUYUKI

(54) INFORMATION RETRIEVING METHOD AND INFORMATION RETRIEVAL SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide an information retrieving method that can easily retrieve documents which conform to complicated retrieval conditions comprising natural sentences or words from documents which are connected through a hyperlink to obtain retrieval results with high precision, and also provide an information retrieval system using the method.

SOLUTION: From not only a group of documents which are in referring relations through a one-stage hyperlink but also a group of documents which are in referring relations through a plural-stage hyperlink, for the hyperlink representing the referring relations, words included in a label given to each hyperlink are extracted and become respective indexes (indexing words) for searching each of documents to which searching is made. The words included in the searching conditions are compared with the above indexes, and the adaptability of each of the documents to the retrieval conditions is computed. Based on the adaptability, the order in displaying of the documents to be displayed as the results of the retrieval is decided, and the documents as the results of the retrieval are displayed in this order.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]It is an information retrieval method for searching a document which suits a search condition which consists of a natural sentence or two or more words which were inputted by user from two or more documents, It is linked by one step of hyperlink between [of said two or more documents] two arbitrary documents, (a) From a document group which is in reference relation about each of two or more of said documents through the document, said one step of hyperlink, and two or more steps of hyperlinks. extracting a word included in a label which was alike, respectively and was attached about said hyperlink showing said reference relation -- (b) -- with two or more words included in said search condition. Compare an extracted word about each of two or more of said documents, and about each of two or more of said documents. An information retrieval method computing goodness of fit with said search condition, determining ranking of a document displayed as search results based on the (c) aforementioned goodness of fit, and displaying a document as said search results according to this ranking.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates to the information retrieval system for retrieving the information for which a user asks, for example, a web page, on the large-scale network represented in the Internet or intranet.
[0002]
[Description of the Prior Art]By the spread of the Internet, the information which everyone wants to disseminate to the world can be freely released now in the form of a web page. On the other hand, it became possible to retrieve the information for which he asks from a huge number of pages by progress of information retrieval technique and the improved efficiency of a computer.
[0003]However, by the time a user can retrieve only information very needed efficiently, it will not have resulted. For example, in the full text retrieval system of a conventional type a user, As a search condition expressing the information for which it asks, the logical formula of a search term (a keyword and a phrase) is inputted, and a search system outputs the page which suits a search condition, i.e., the page which contains a search term so that a logical formula may be filled, as search results. Ranking of search results is mainly performed by the frequency where a search term appears in a page, and the position. However, it is impossible to find out information worthy for a user in such a simple retrieval system out of the web page to which billions are said.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL FIELD

[Field of the Invention]This invention relates to the information retrieval system for retrieving the information for which a user asks, for example, a web page, on the large-scale network represented in the Internet or intranet.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art]By the spread of the Internet, the information which everyone wants to disseminate to the world can be freely released now in the form of a web page. On the other hand, it became possible to retrieve the information for which he asks from a huge number of pages by progress of information retrieval technique and the improved efficiency of a computer.

[0003]However, by the time a user can retrieve only information very needed efficiently, it will not have resulted. For example, in the full text retrieval system of a conventional type a user, As a search condition expressing the information for which it asks, the logical formula of a search term (a keyword and a phrase) is inputted, and a search system outputs the page which suits a search condition, i.e., the page which contains a search term so that a logical formula may be filled, as search results. Ranking of search results is mainly performed by the frequency where a search term appears in a page, and the position. However, it is impossible to find out information worthy for a user in such a simple retrieval system out of the web page to which billions are said.

[0004]From such reflection, a worthy website is first looked for by human being's handicraft, and service which provides a user with this came to be performed. There is service provided in the form where the website collected with the help is arranged to directory structure, and it is easy to use it for one of them. For example, the group of the name of organizations, such as a company, and the place (URL) of the website which the organization is managing is put in a database, and the service etc. which present the website equivalent to the company name which the user inputted are employed. However, the work which arranges comprehensively splenium and the information updated every day by a help is impossible, and also requires a labor in or dramatically.

[0005]On the other hand, it asks for a page worthy for a user automatically, and some methods of showing this preferentially in search results are considered. For example, in search system

Google (http://www.google.com/) of U.S. Google. With the importance which was searched for based on the hypothesis that the page linked to many pages is an important page, and the page linked to the important page is still more important and which is called PageRank. Search results. The method of carrying out ranking is taken (document 1:Sergey.). Brinand LawrencePage. The Anatomy of a Large-Scale Hypertextual Web Search Engine.The 7 th International WorldWide WebConference, 1998.

[0006]According to this method, for example to the retrieval required "T company", a page with many linking numbers among web pages including the word of "T company" is ranked more as a higher rank. However, in this method, since the importance of a page is searched for regardless of the word of a user's retrieval required, i.e., "T company", there is a possibility that the page of the contents which are unrelated to retrieval required may be ranked as the higher rank of search results.

[0007]On the other hand, not only in a linking number, Expression (in the case of an HTML document, called an anchor text) of the label of a link is taken into consideration. The method of carrying out ranking of the search results is tried (document 2: the Kazuhiro Kazama, Masanori Harada, and Shin-ya Sato . "way method of hyperlink, information retrieval [using an anchor text], and ranking" Information Processing Society of Japan report of research, SIGDD, Vol.24-2000). This method is based on the hypothesis that the label of a link shows the contents of the page of a link destination well. According to this method, the page mostly referred to from other pages by the link to which the label "T company" was given is important for a user, and retrieval required is called for as it is a page which suits well.

[0008]The same method is devised also with the "hypertext retrieval device" (document 3: Patent Gazette No. 3108015). Also by this patent, in addition to the goodness of fit of the contents of a page themselves, the goodness of fit of the anchor text of the reference origin of a page was also taken into consideration, and the method of asking for the goodness of fit of a page is taken.

[0009]By unifying the pages in reference relation and asking for goodness of fit with retrieval required, For example, when there are a page including the word of "Aomori" and a page including the word of an "apple" and the latter is referred by the link from the former, it considers that these two pages are one document, and suppose that it is this a document which suits the retrieval required the "Aomori apple." It is supposed that a user can grasp the reference relation of these pages easily by showing search results in the form where the page including "Aomori" and the page containing an "apple" were combined.

[0010]As the same invention as document 3, there is other "storages which recorded the text retrieval device, the method, and the document retrieval program and in which computer reading is possible" (document 4: JP,2000-259648,A).

[0011]According to the method of above-mentioned document 1 and document 2, the top page

of the website which organizations, such as what is called a formal website, i.e., a company etc., are managing officially is reported to be able to search almost correctly by making the corporate name into retrieval required. To the retrieval required "T company", the top page of the website which the company "T company" is employing officially is actually ranked as the higher rank of search results. The top page of the official site of "T company" has more linking numbers than other pages included the expression "T company", and this is because it is referred by the link to which the label included the expression "T company" was given in many cases.

[0012]However, a user's demand is not only finding an official site. For example, more detailed and complicated information [say / "liking to fix the notebook computer of T company"] is required in many cases. And the page which suits well cannot be searched with the method of document 1 and 2 to such retrieval required. At least 3000 pages or more of pages containing all the search terms "T company", a "note", a "personal computer", and "repair" are on the Internet. Among these, the page about the method or procedure of repair of a notebook computer which the information for which a user asks, i.e., "T company" which is the manufacturer, exhibits officially on the website is not necessarily ranked as a higher rank by the method of document 1. It is because it is rare that extremely many pages other than the top page of a website are linked from other pages, so it is hard to come to the importance (PageRank of document 1) of a page out of a significant difference.

[0013]Similarly, it can refer correctly also to the method of document 2, and twists and is afraid by it. It is because many pages for which a user asks are not necessarily linked with a label including four words of "T company", a "note", a "personal computer", and "repair." For example, the page for which it asks may be opened to the place which followed the link in order in the label a "personal computer" and "PC customer center" from the top page of "T company" site, and may be linked to this page itself only with the label "PC customer center." On the contrary, the page which were linked with the label containing the three words a "note", a "personal computer", and "repair" and which is unrelated to "T company" may exist.

[0014]It is supposed that it can refer to the method currently devised by document 3 or document 4 on the other hand as a page which unified two or more hypertexts based on the reference relation by a link. However, when there are a huge number of pages in the Internet and the number of the link between pages is also averaged, it is also several times - about ten times of the number of a page. The web page included at least one of the four words of "T company", a "note", a "personal computer", and "repair" is actually impossible for unifying the each by a link in detail, and asking for goodness of fit on the Internet, about those with 2 million pages or more, and these huge pages. If what unified two or more documents is made into a retrieval object, although the recall of search improves, generally it is known that precision will fall. It is necessary to search the page which suits a user's complicated retrieval required by

the method often controlled more efficiently.

[0015]The structure of an actual link is not complicated and is not necessarily arranged by hierarchical structure. There are many commutative links which are not understood which [the links aiming at helping the ease of carrying out of not only the thing showing the relation on the contents the page's but a user's browsing and either] the link of a web page has quoted, links which became a loop, etc.

[0016]By document 4, there is a possibility of unifying the page which does not almost have the contents relation of what it is being supposed that the link which became a loop can be eliminated too, much. It is complicated and it difficult for a user to arrange and show a legible form link structure with much number on search results by the method of document 3 or document 4.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

EFFECT OF THE INVENTION

[Effect of the Invention]As explained above, while being able to search easily the document which suits the complicated search condition which consists of two or more words out of a lot of hypertext format documents according to this invention, the search results of high accuracy are obtained.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]As explained above, when searching conventionally based on the reference relation by the link between web pages, Since a certain web page used only one step of link which carries out the direct reference of other web pages, there was a problem that the web page which suits the complicated search condition which consists of two or more words could not be searched easily.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

MEANS

[Means for Solving the Problem]This invention is for searching a document which suits a search condition which consists of a natural sentence or two or more words which were inputted by user from two or more documents, It is linked by one step of hyperlink between [of said two or more documents] two arbitrary documents, (a) From a document group which is in reference relation about each of two or more of said documents through the document, said one step of hyperlink, and two or more steps of hyperlinks. extracting a word included in a label which was alike, respectively and was attached about said hyperlink showing said reference relation -- (b) -- with two or more words included in said search condition. Compare an extracted word about each of two or more of said documents, and about each of two or more of said documents. Goodness of fit with said search condition is computed, ranking of a document displayed as search results is determined based on the (c) aforementioned goodness of fit, and a document as said search results is displayed according to this ranking.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]The figure showing the example of composition of the information retrieval system concerning a 1st embodiment of this invention.
[Drawing 2]The identifier given to each of two or more pages of a retrieval object, and the figure showing the example of memory of URL of each page corresponding to each identifier.
[Drawing 3]The figure showing the example of memory of the link information in a link information storage part.
[Drawing 4]The figure showing the example of memory of the page information in a page information storage parts store.
[Drawing 5]The figure showing an example of the reference relation by the hyperlink of two or more pages of a retrieval object.
[Drawing 6]The flow chart for explaining the processing for creating the vector of the link of n stage about each page.
[Drawing 7]The flow chart for explaining the processing for creating the vector of page contents about each page.
[Drawing 8]The flow chart for explaining retrieval processing operation of the information retrieval system of drawing 1.
[Drawing 9]The figure showing an example of the input screen which inputs retrieval required with the figure showing the example of a screen display of the user interface of the information retrieval system of drawing 1.
[Drawing 10]The figure showing the display example of search results with the figure showing the example of a screen display of the user interface of the information retrieval system of drawing 1.
[Drawing 11]The figure showing the display example of search results with the figure showing the example of a screen display of the user interface of the information retrieval system of

drawing 1.

[Drawing 12]The figure showing the example of composition of the information retrieval system concerning a 2nd embodiment of this invention.

[Drawing 13]The figure showing an example of the reference relation by two or more pages of a retrieval object, and document groups's hyperlink.

[Drawing 14]The flow chart for explaining retrieval processing operation of the information retrieval system of drawing 12.

[Drawing 15]The figure showing the display example of search results with the figure showing the example of a screen display of the user interface of the information retrieval system of drawing 12.

[Drawing 16]The figure showing other display examples of search results with the figure showing the example of a screen display of the user interface of the information retrieval system of drawing 12.

[Description of Notations]

1 -- User interface
2 -- Search term extraction part
3 -- Link retrieval part
4 -- Link information storage part
5 -- Search-results integration part
6 -- Page retrieval part
7 -- Page information storage parts store
8 -- Link index build part
9 -- Page reference preparing part
10 -- The document groups link index build part
11 -- Web information gathering part
50 -- Retrieval part in document groups
51 -- The document groups link retrieval part
52 -- Link information storage part between document groups

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

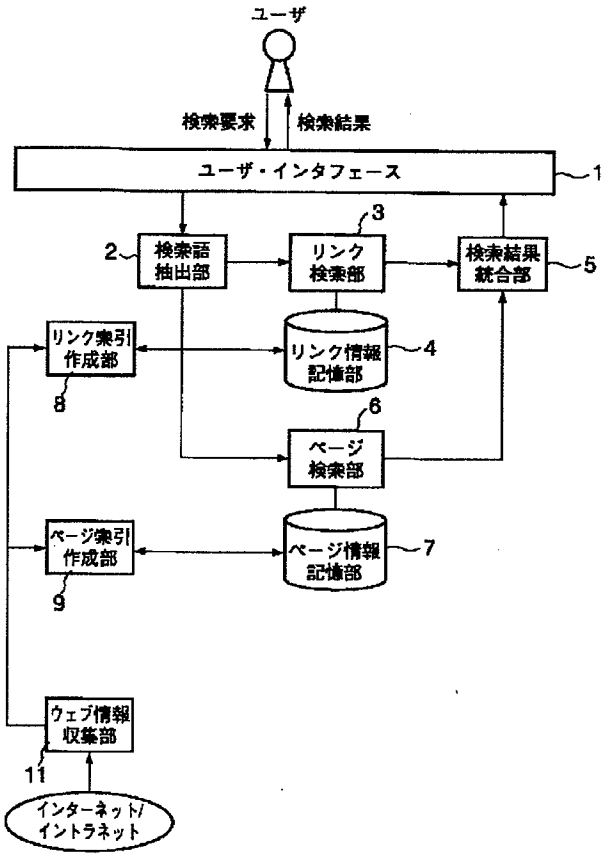
[Drawing 2]

ID	URL
1-1	http://www.foo.co.jp/
1-2	http://www.foo.co.jp/products/
1-3	http://www.foo.co.jp/products/pc.html
2-1	http://www.bar.com/
2-2	http://www.bar.com/about/

[Drawing 3]

リンク元ID	リンク先ID	ラベル
8-16	5-1	株式会社T社
27-368	5-1	㈱T社のホームページ
32-59	5-1	T社へのリンク
5-4	5-1	T社トップ

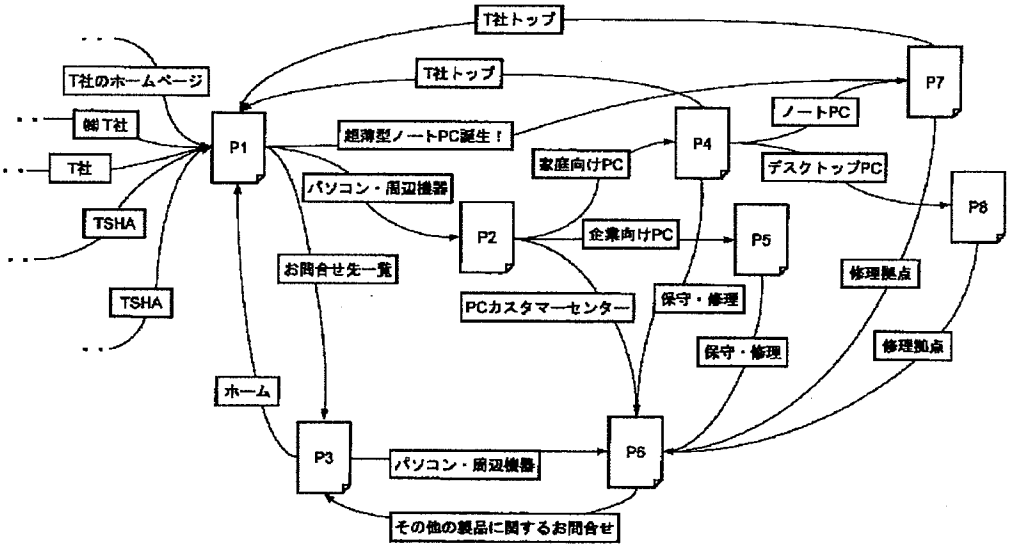
[Drawing 1]



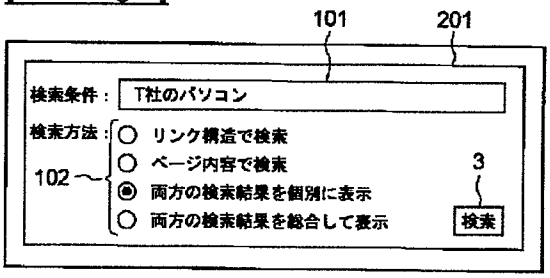
[Drawing 4]

ID	タイトル	本 文
5-1	㈱T社	T社のウェブサイトへようこそ！ 製品紹介 企業情報 IR情報 採用情報 お問合せ・・
5-2	T社PCウェブ	T社PCウェブはT社製パソコン製品の総合サイトです。更新日 2001年4月7日最新情報・・
5-6	T社PCカスタマーセンター	T社PCカスタマーセンター無償修理 アップデート 修理拠点 お知らせ・・

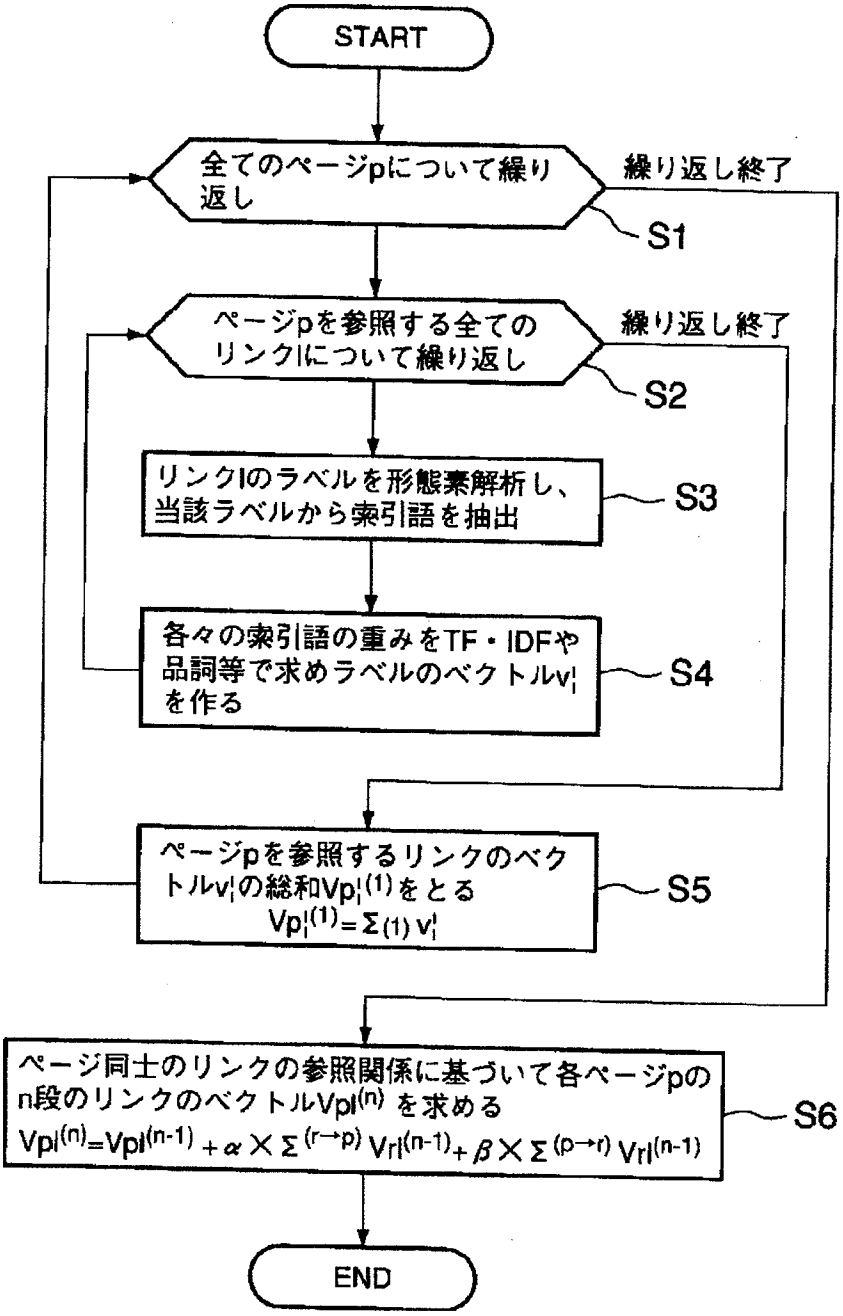
[Drawing 5]



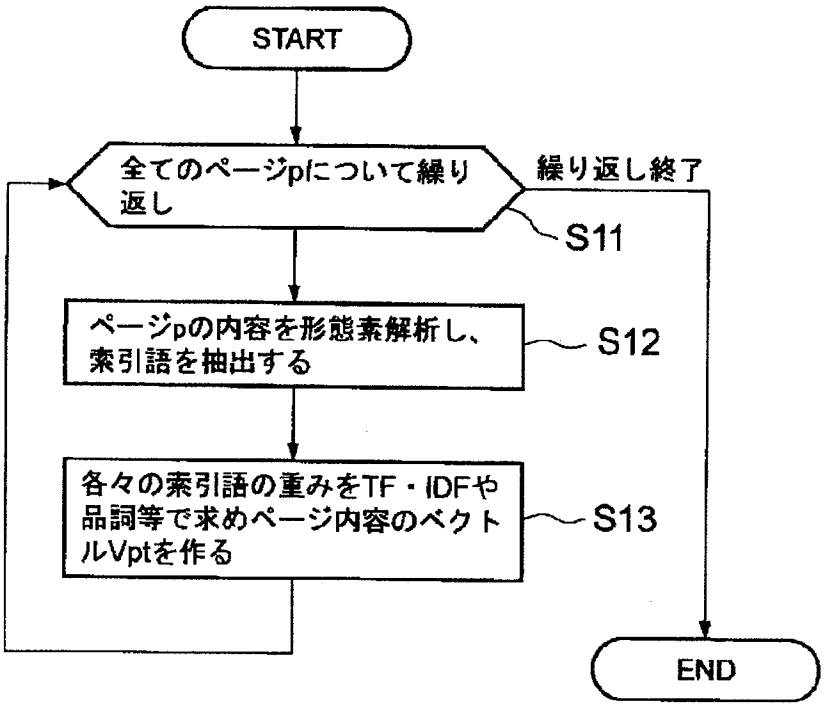
[Drawing 9]



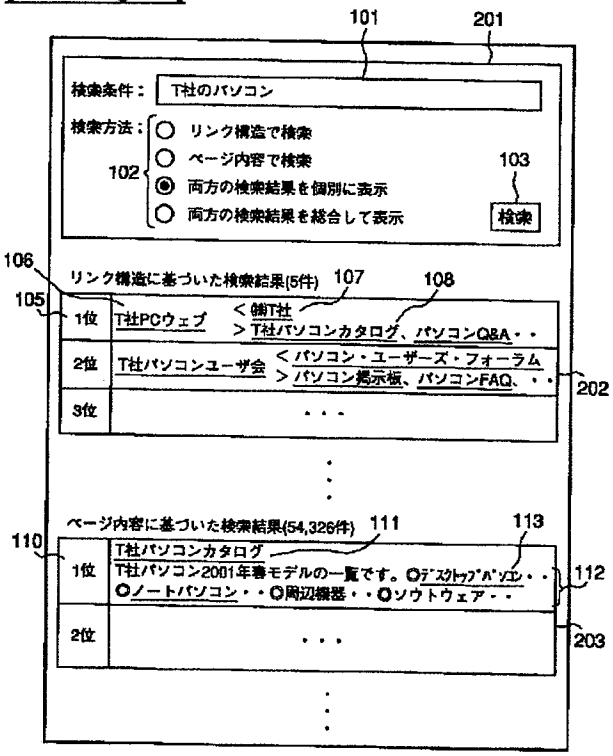
[Drawing 6]



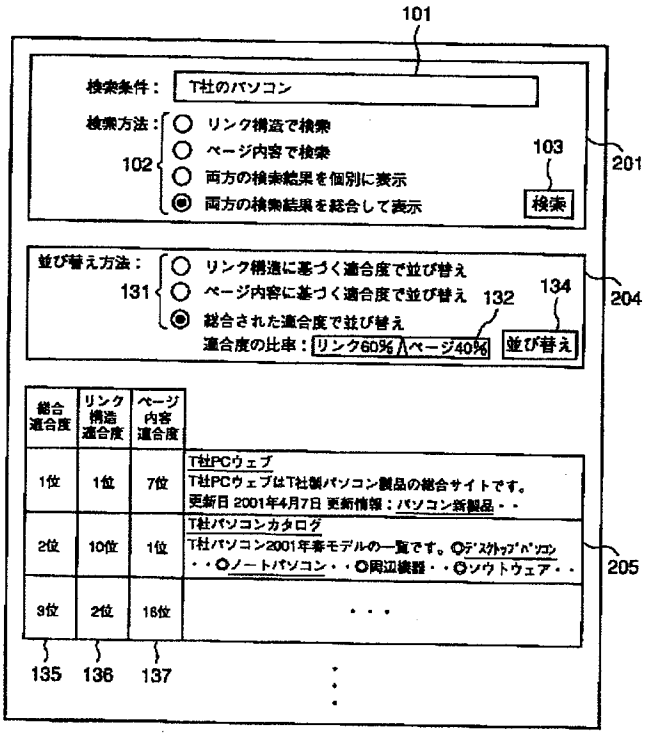
[Drawing 7]



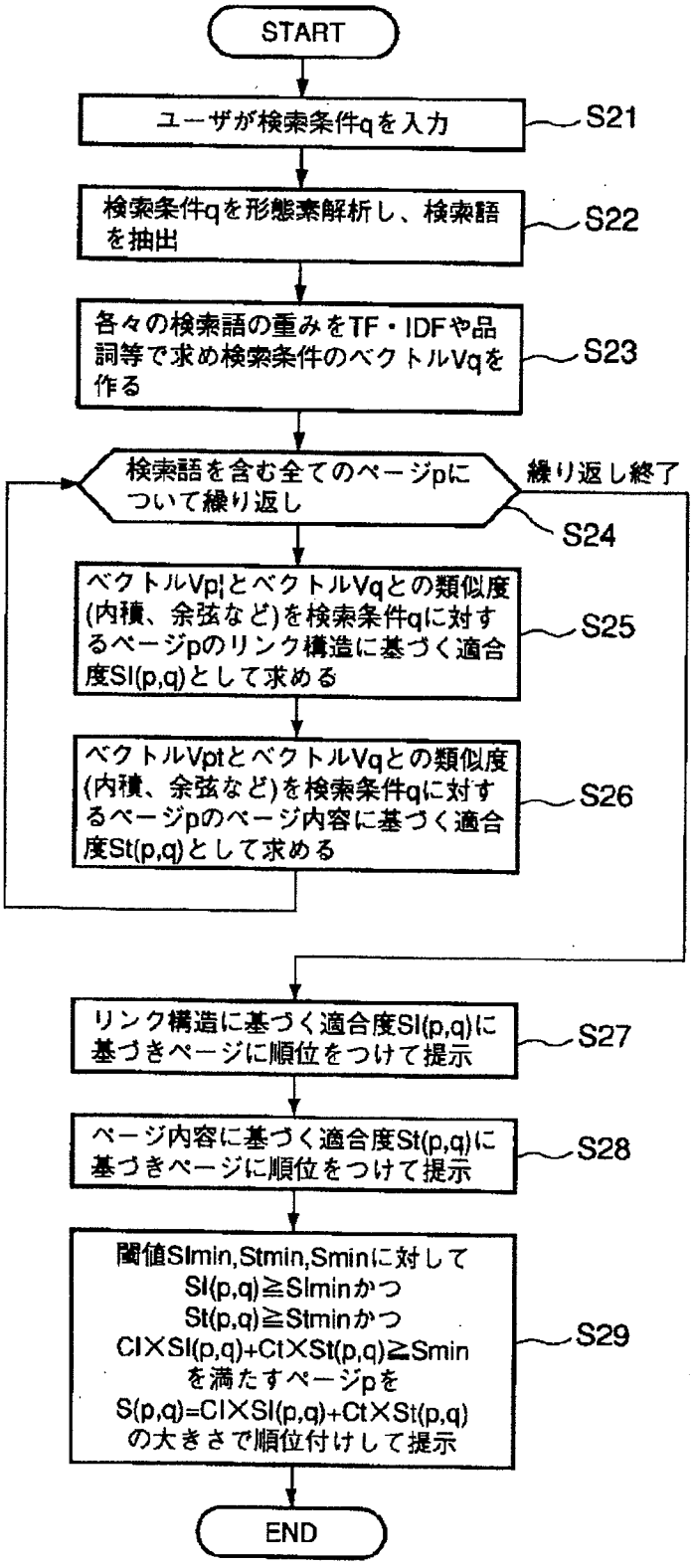
[Drawing 10]



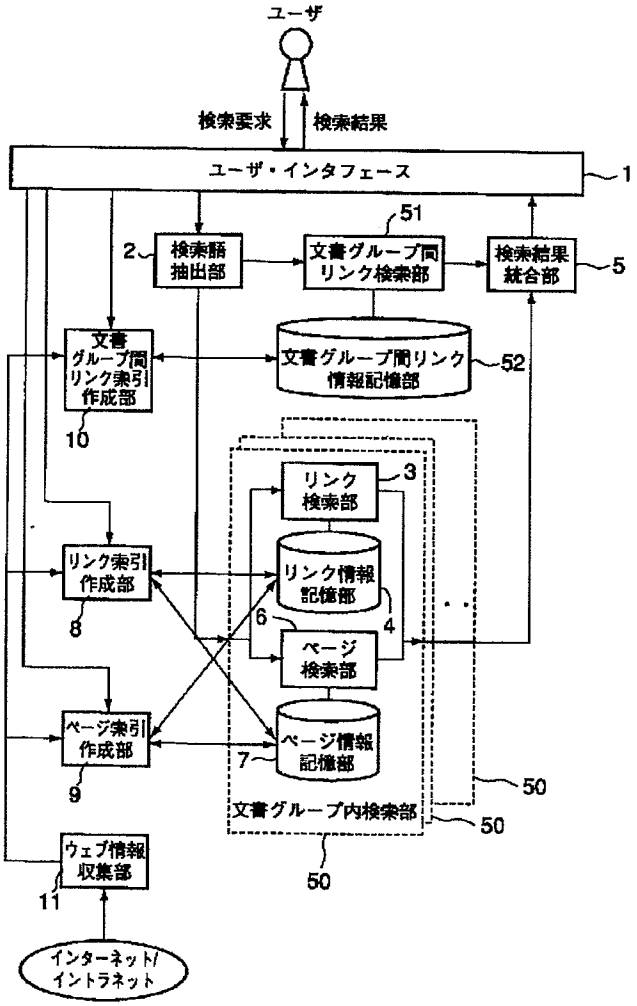
[Drawing 11]


101
検索条件：
T社のパソコン
検索方法：
リンク構造で検索
ページ内容で検索
両方の検索結果を個別に表示
両方の検索結果を総合して表示
102
103
検索
201
並び替え方法：
リンク構造に基づく適合度で並び替え
ページ内容に基づく適合度で並び替え
総合された適合度で並び替え
131
132
134
適合度の比率：リンク60%/ページ40%
並び替え
204
総合適合度
リンク構造適合度
ページ内容適合度
1位
1位
7位
T社PCウェブ
T社PCウェブはT社製パソコン製品の総合サイトです。
更新日 2001年4月7日 更新情報：パソコン新製品・・
2位
10位
1位
T社パソコンカタログ
T社パソコン2001年春モデルの一覧です。○ﾃﾞｽｸﾄｯﾌﾟﾊﾟｿｺﾝ
・・○ノートパソコン・・○周辺機器・・○ソウトウェア・・
205
3位
2位
18位
・・・
135
136
137

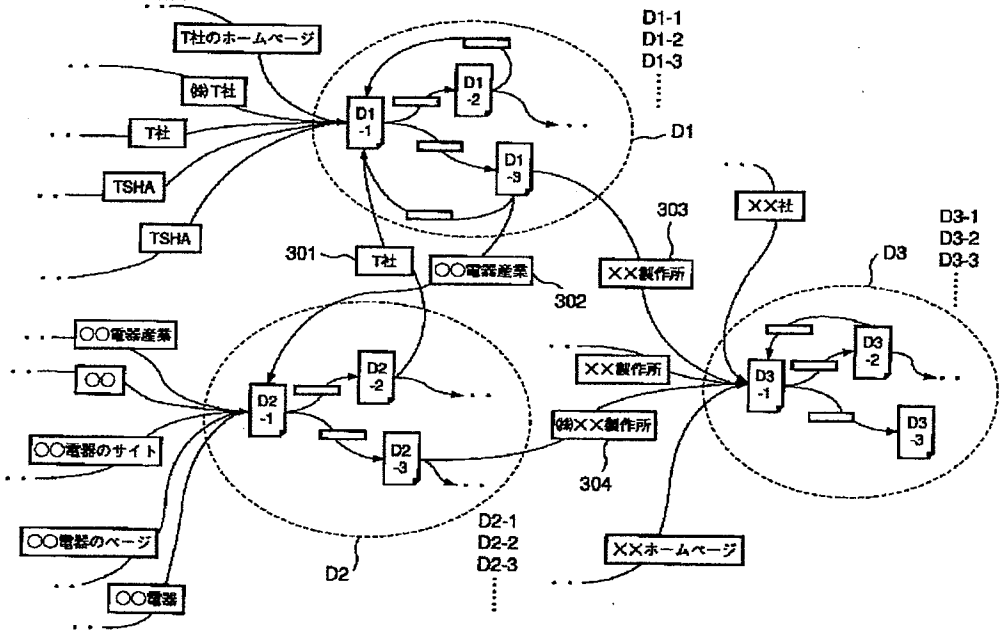
[Drawing 8]


START
ユーザが検索条件qを入力
S21
検索条件qを形態素解析し、検索語を抽出
S22
各々の検索語の重みをTF・IDFや品詞等で求め検索条件のベクトルVqを作る
S23
検索語を含む全てのページpについて繰り返し
繰り返し終了
S24
ベクトルVpiとベクトルVqとの類似度(内積、余弦など)を検索条件qに対するページpのリンク構造に基づく適合度SI(p,q)として求める
S25
ベクトルVptとベクトルVqとの類似度(内積、余弦など)を検索条件qに対するページpのページ内容に基づく適合度St(p,q)として求める
S26
リンク構造に基づく適合度SI(p,q)に基づきページに順位をつけて提示
S27
ページ内容に基づく適合度St(p,q)に基づきページに順位をつけて提示
S28
閾値SImin,Stmin,Sminに対して
SI(p,q)≧SIminかつ
St(p,q)≧Stminかつ
CI×SI(p,q)+Ct×St(p,q)≧Smin
を満たすページpを
S(p,q)=CI×SI(p,q)+Ct×St(p,q)
の大きさで順位付けして提示
S29
END

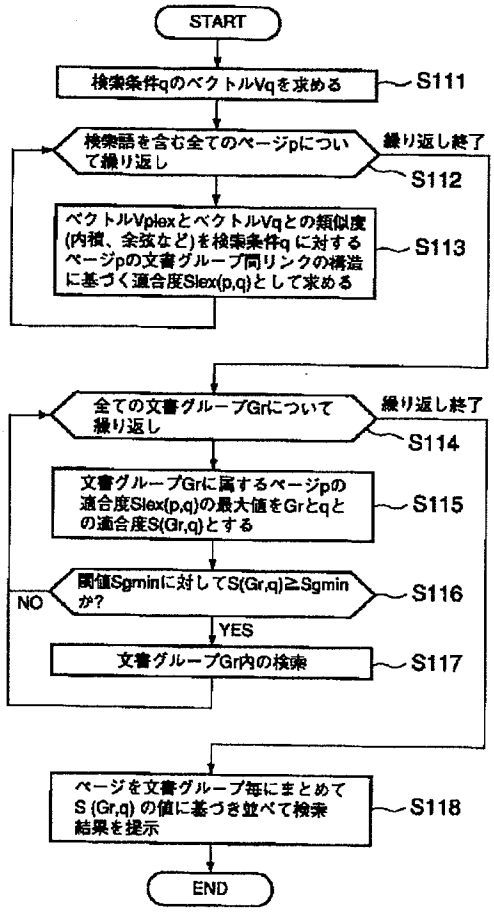
[Drawing 12]



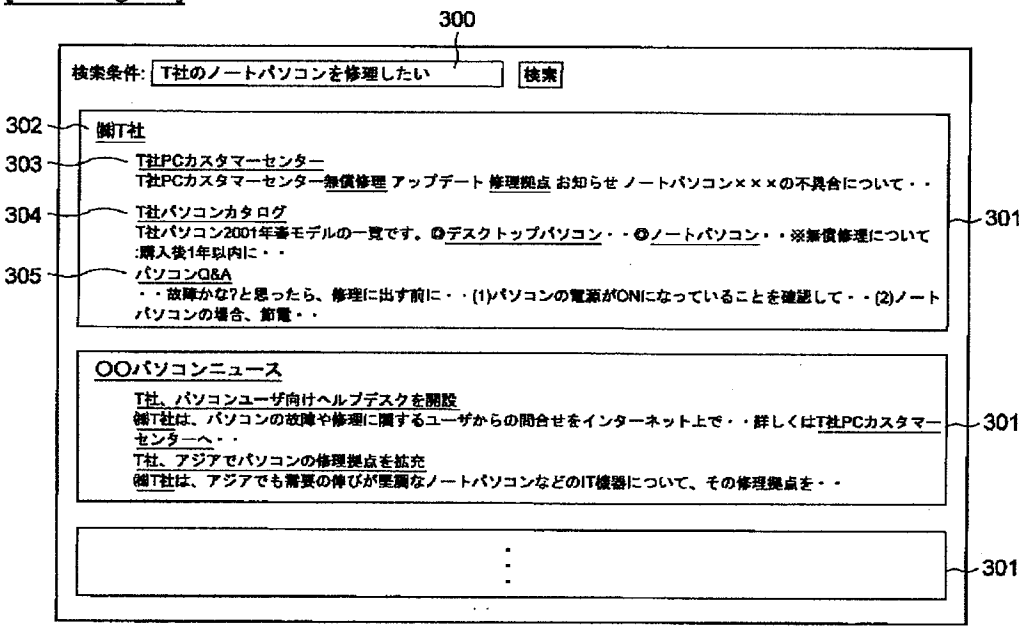
[Drawing 13]



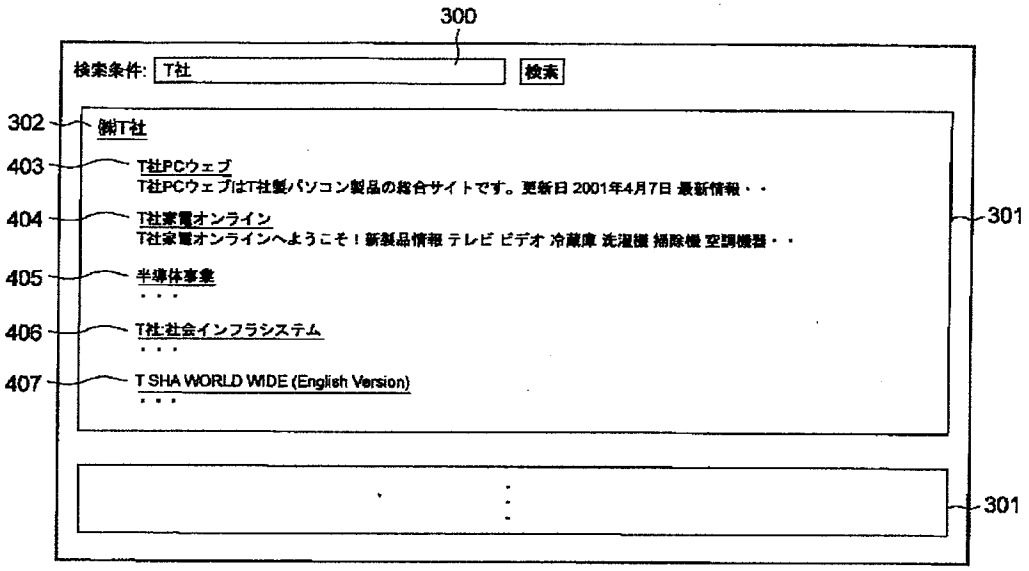
[Drawing 14]



[Drawing 15]



[Drawing 16]



[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－67419
（P2003－67419A）
(43)公開日　平成15年３月７日(2003.3.7)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/30	３８０	Ｇ０６Ｆ　17/30	３８０Ｅ	５Ｂ０７５
	３５０		３５０Ｃ	５Ｂ０８２
	４１９		４１９Ｂ	
12/00	５４６	12/00	５４６Ｔ	

審査請求　未請求　請求項の数10　ＯＬ　（全 23 頁）

(21)出願番号　特願2001－254772(P2001－254772)

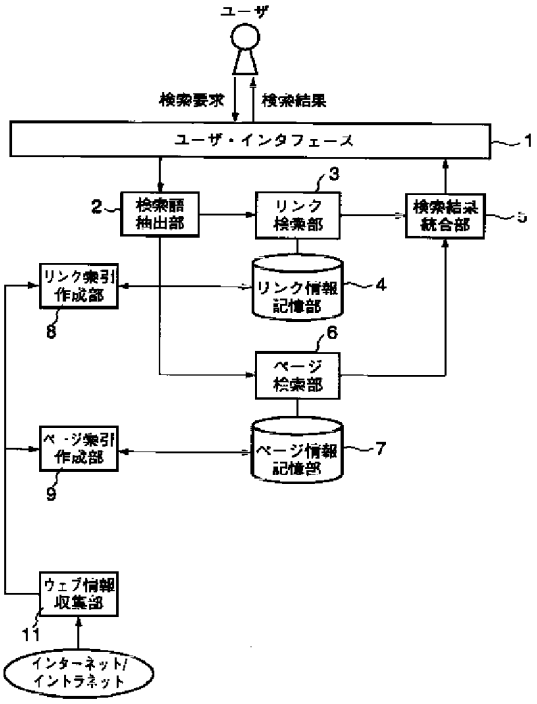
(22)出願日　平成13年８月24日(2001.8.24)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(72)発明者　後藤　和之
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(74)代理人　100058479
弁理士　鈴江　武彦　（外６名）

Ｆターム(参考)　5B075 ND03 ND36 NK02 PQ02 PQ32
PQ74 PR06 UU40
5B082 GA08 HA05

(54)【発明の名称】　情報検索方法および情報検索システム

(57)【要約】
【課題】ハイパーリンクで結ばれた複数の文書の中から、自然文または複数の語からなる複雑な検索条件に適合する文書の検索が容易に行えるとともに、高い精度の検索結果が得られる情報検索方法およびそれを用いた情報検索システムを提供する。
【解決手段】１段のハイパーリンクにより参照関係にある文書群のみならず、複数段のハイパーリンクを経て参照関係にある文書群からも、その参照関係を表すハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出して、それを検索対象の複数の文書のそれぞれについての検索時のインデックス（索引語）とし、検索条件に含まれる複数の語と上記インデックスとを比較して、複数の文書のそれぞれについて検索条件との適合度を算出し、この適合度に基づき、検索結果として表示する文書の順位を決定し、この順位に従って検索結果としての文書を表示する。



【特許請求の範囲】
【請求項１】　複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索するための情報検索方法であって、前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、（ａ）前記複数の文書のそれぞれについて、その文書と前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある文書群から、前記参照関係を表す前記ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、（ｂ）前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との適合度を算出し、（ｃ）前記適合度に基づき、検索結果として表示する文書の順位を決定し、この順位に従って前記検索結果としての文書を表示することを特徴とする情報検索方法。
【請求項２】　前記検索結果として表示する文書に、該文書を前記１段のハイパーリンクで参照する関係にあるリンク元文書があるとき、該リンク元文書の前記適合度が所定値以上であれば、該リンク元文書の存在を前記検索結果として表示する文書に関連付けて表示し、その際、所定の操作により該リンク元文書の内容表示を可能にする形態で表示することを特徴とする請求項１記載の情報検索方法。
【請求項３】　前記検索結果として表示する文書に、該文書が前記１段のハイパーリンクで参照する関係にあるリンク先文書があるとき、該リンク先文書の前記適合度が所定値以上であれば、該リンク先文書の存在を前記検索結果として表示する文書に関連付けて表示し、その際、所定の操作により該リンク先文書の内容表示を可能にする形態で表示することを特徴とする請求項１記載の情報検索方法。
【請求項４】　複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索するための情報検索方法であって、前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、（ａ）前記複数の文書のそれぞれについて、その文書と前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある文書群から、前記参照関係を表す前記ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、（ｂ）前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて前記ラベルから抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第１の適合度を算出し、この第１の適合度に基づき、検索結果として表示する文書の順位を決定し、（ｃ）前記複数の文書のそれぞれから、その文書の内容を表す語を抽出し、（ｄ）前記検索条件に含まれる複数の語と前記複数の文書のそれぞれについて、その内容から抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第２の適合度を算出し、この第２の適合度に基づき、検索結果として表示する文書の順位を決定し、（ｅ）前記第１および第２の適合度を統合した第３の適合度を算出し、この第３の適合度に基づき、検索結果として表示する文書の順位を決定し、（ｆ）前記第１～第３の適合度のそれぞれに対応して決定された順位のうちの少なくとも１つを用いて、検索結果としての文書を表示することを特徴とする情報検索方法。
【請求項５】　複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索するための情報検索方法であって、前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、前記複数の文書のそれぞれは、予め定められた複数の文書グループのうちのうちの１つに属し、（ａ）前記複数の文書のそれぞれについて、その文書の属する文書グループ内から、前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある第１の文書群を抽出して、この各第１の文書群から、前記参照関係を表す前記ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、（ｂ）前記複数の文書のそれぞれについて、２つの前記文書グループ間にまたがって２つの文書をリンクする１段の文書グループ間ハイパーリンクおよび複数段の文書グループ間ハイパーリンクを経て文書グループ間の参照関係にある第２の文書群を前記複数の文書から抽出し、この各第２の文書群から、前記文書グループ間の参照関係を表す文書グループ間ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、（ｃ）前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて前記文書グループ間ハイパーリンクのラベルから抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第１の適合度を算出し、（ｄ）前記第１の適合度に基づき前記複数の文書グループから少なくとも１つの文書グループを検索対象として選択し、（ｅ）前記検索対象として選択された文書グループのそれぞれについて、前記検索条件に含まれる複数の語と、前記文書グループ内の文書のそれぞれについて前記ハイパーリンクのラベルから抽出された語とを比較して、前記文書グループ内の文書のそれぞれについて、前記検索条件との第２の適合度を算出し、（ｆ）前記第１の適合度に基づき検索結果として表示する文書グループの順位を決定するとともに、前記文書グループ毎に前記第２の適合度に基づき検索結果として表示する文書の順位を決定し、これら順位に従って前記検索結果としての文書グループと文書を表示することを特徴とする情報検索方法。
【請求項６】　複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文

書を検索する情報検索システムであって、
前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、前記複数の文書のそれぞれについて、その文書と前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある文書群から、前記参照関係を表す前記ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出する抽出手段と、
前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて前記抽出手段で抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との適合度を算出する算出手段と、
前記適合度に基づき、検索結果として表示する文書の順位を決定し、この順位に従って前記検索結果としての文書を表示する手段と、
を具備したことを特徴とする情報検索システム。
【請求項７】　前記検索結果として表示する文書に、該文書を前記１段のハイパーリンクで参照する関係にあるリンク元文書があるとき、該リンク元文書の前記適合度が所定値以上であれば、該リンク元文書の存在を前記検索結果として表示する文書に関連付けて表示し、その際、所定の操作により該リンク元文書の内容表示を可能にする形態で表示することを特徴とする請求項６記載の情報検索システム。
【請求項８】　前記検索結果として表示する文書に、該文書が前記１段のハイパーリンクで参照する関係にあるリンク先文書があるとき、該リンク先文書の前記適合度が所定値以上であれば、該リンク先文書の存在を前記検索結果として表示する文書に関連付けて表示し、その際、所定の操作により該リンク先文書の内容表示を可能にする形態で表示することを特徴とする請求項６記載の情報検索システム。
【請求項９】　複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索する情報検索システムであって、
前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、前記複数の文書のそれぞれについて、その文書と前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある文書群から、前記参照関係を表すハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出する第１の抽出手段と、
前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて前記第１の抽出手段で抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第１の適合度を算出し、この第１の適合度に基づき、検索結果として表示する文書の順位を決定する第１の検索手段と、
前記複数の文書のそれぞれから、その文書の内容を表す語を抽出する第２の抽出手段と、
前記検索条件に含まれる複数の語と前記複数の文書のそれぞれについて前記第２の抽出手段で抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第２の適合度を算出し、この第２の適合度に基づき、検索結果として表示する文書の順位を決定する第２の検索手段と、
前記第１および第２の適合度を統合した第３の適合度を算出し、この第３の適合度に基づき、検索結果として表示する文書の順位を決定する第３の検索手段と、
前記第１～第３の適合度のそれぞれに対応して決定された順位のうちの少なくとも１つを用いて、検索結果としての文書を表示する手段と、
を具備したことを特徴とする情報検索システム。
【請求項１０】　複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索する情報検索システムであって、
前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、前記複数の文書のそれぞれは、予め定められた複数の文書グループのうちのうちの１つに属し、
前記複数の文書のそれぞれについて、その文書の属する文書グループ内から、前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある第１の文書群を抽出し、この各第１の文書群から、前記１段のハイパーリンクおよび前記複数段のハイパーリンクのそれぞれに付されたラベルに含まれる語を抽出する第１の抽出手段と、
前記複数の文書のそれぞれについて、その文書と、２つの前記文書グループ間にまたがって２つの文書をリンクする１段の文書グループ間ハイパーリンクおよび複数段の文書グループ間ハイパーリンクを経て文書グループ間の参照関係にある第２の文書群を前記複数の文書から抽出し、この各第２の文書群から、前記文書グループ間の参照関係を表す文書グループ間ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出する第２の抽出手段と、
前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて前記第２の抽出手段で抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第１の適合度を算出し、この第１の適合度に基づき前記複数の文書グループから少なくとも１つの文書グループを検索対象として選択する手段と、
前記検索対象として選択された文書グループのそれぞれについて、前記検索条件に含まれる複数の語と、前記文書グループ内の文書のそれぞれについて前記第１の抽出手段で抽出された語とを比較して、前記文書グループ内の文書のそれぞれについて、前記検索条件との第２の適合度を算出する算出手段と、
前記第１の適合度に基づき、検索結果として表示する文書グループの順位を決定するとともに、前記第２の適合

度に基づき前記文書グループ毎に、検索結果として表示する文書の順位を決定し、これら順位に従って前記検索結果としての文書グループと文書を表示する手段と、を具備したことを特徴とする情報検索システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、インターネットやイントラネットに代表される大規模なネットワーク上で、ユーザが所望する情報、たとえばウェブページを検索するための情報検索システムに関する。

【０００２】

【従来の技術】インターネットの普及により、誰もが世界に発信したい情報をウェブページという形で自由に公開できるようになった。一方、情報検索技術の進歩と計算機の性能向上により、膨大な数のページから、自分が所望する情報を検索することが可能になった。

【０００３】しかしながら、ユーザが真に欲しい情報だけを効率よく検索できるまでには至っていない。例えば、従来型の全文検索システムでは、ユーザは、所望する情報を表現する検索条件として、検索語（キーワードやフレーズ）の論理式を入力し、検索システムは、検索条件に適合するページ、すなわち、論理式を満たすように検索語を含むページを検索結果として出力する。検索結果のランキングは、主に、検索語がページ中に出現する頻度や位置によって行なわれる。しかし、このような素朴な検索方式では、数十億ともいわれるウェブページの中から、ユーザにとって価値の高い情報を見つけ出すことは不可能である。

【０００４】このような反省から、まず、価値のあるウェブサイトを人間の手作業で探して、これをユーザに提供するサービスが行なわれるようになった。その１つに、人手で集めたウェブサイトをディレクトリ構造に整理して利用しやすい形で提供するサービスがある。また、例えば、企業などの団体の名称と、その団体が運営しているウェブサイトの場所（ＵＲＬ）との組をデータベース化して、ユーザが入力した企業名に相当するウェブサイトを提示するサービスなどが運用されている。しかしながら、膨大、かつ、日々更新される情報を人手によって網羅的に整理する作業は不可能であり、労力も非常にかかかる。

【０００５】これに対し、ユーザにとって価値の高いページを自動的に求めて、これを検索結果の中で優先的に提示する方法がいくつか考えられている。例えば、米国グーグルの検索システムＧｏｏｇｌｅ（ｈｔｔｐ：／／ｗｗｗ．ｇｏｏｇｌｅ．ｃｏｍ／）では、多くのページにリンクされているページは重要なページであり、さらに、重要ページにリンクされているページは重要である、という仮説に基づいて求めた、ＰａｇｅＲａｎｋと呼ばれる重要度によって、検索結果をランキングする方法がとられている（文献１：Ｓｅｒｇｅｙ　Ｂｒｉｎ ａｎｄ Ｌａｗｒｅｎｃｅ Ｐａｇｅ．The Anatomy of a Large－Scale Hypertextual Web Search Engine．The 7th International WorldWide WebConference，１９９８）。

【０００６】この方法によれば、たとえば「Ｔ社」という検索要求に対しては、「Ｔ社」という語を含むウェブページのうち、被リンク数の多いページが、より上位にランクされる。しかしながら、この方法では、ページの重要度は、ユーザの検索要求、すなわち、「Ｔ社」という語に無関係に求められたものであるため、検索要求に関係のない内容のページが検索結果の上位にランクされる恐れがある。

【０００７】これに対して、被リンク数だけでなく、リンクのラベルの表現（ＨＴＭＬ文書の場合にはアンカーテキストと呼ばれる）を考慮して検索結果をランキングする方法が試みられている（文献２：風間一洋，原田昌紀，佐藤進也．「ハイパーリンクとアンカーテキストを利用した情報検索とランキングの一手法」情報処理学会研究報告，ＳＩＧＤＤ，Ｖｏｌ．２４，２０００）。この方法は、リンクのラベルは、リンク先のページの内容をよく示すという仮説に基づいている。この方法によれば、「Ｔ社」というラベルが付されたリンクによって他ページから多く参照されたページが、ユーザにとって重要で、かつ、検索要求ともよく適合するページであると求められる。

【０００８】同様の方法は「ハイパーテキスト検索装置」（文献３：特許公報第３１０８０１５号）でも考案されている。この特許でも、ページの内容自体の適合度に加え、ページの参照元のアンカーテキストの適合度も考慮して、ページの適合度を求める方法をとっている。

【０００９】また、参照関係にあるページ同士を統合して検索要求との適合度を求めることにより、たとえば、「青森」の語を含むページと「りんご」の語を含むページがあり、前者から後者がリンクで参照されている場合、これら２つのページを１つの文書とみなして、これを「青森りんご」という検索要求に適合する文書であるとする。また、「青森」を含むページと「りんご」を含むページを併せた形で検索結果を提示することにより、ユーザがこれらのページの参照関係を容易に把握することができるとしている。

【００１０】文献３と同様の発明としては、他に「文章検索装置および方法ならびに文書検索プログラムを記録したコンピュータ読取り可能な記憶媒体」（文献４：特開２０００－２５９６４８号公報）がある。

【００１１】上記文献１および文献２の方法によれば、いわゆる公式ウェブサイト、すなわち、企業などの団体が公式に運営しているウェブサイトのトップページを、その団体名を検索要求として、ほぼ正しく検索すること

ができると報告されている。実際、「Ｔ社」という検索要求に対しては、「（株）Ｔ社」という企業が公式に運用しているウェブサイトのトップページが、検索結果の上位にランクされる。これは、「（株）Ｔ社」の公式サイトのトップページは、「Ｔ社」という表現を含んだ他のページよりも、被リンク数が多く、かつ、「Ｔ社」という表現を含んだラベルを付されたリンクによって参照されることが多いからである。

【００１２】しかしながら、ユーザの要求は、公式サイトを見つけることだけではない。例えば、「Ｔ社のノートパソコンを修理したい」といった、より詳細で複雑な情報を要求する場合が多い。そして、このような検索要求に対しては、文献１および２の方法では、よく適合するページを検索することはできない。「Ｔ社」「ノート」「パソコン」「修理」という検索語を全て含むページは、インターネット上に少なくとも３千ページ以上ある。このうち、ユーザが所望する情報、すなわち、製造元である「（株）Ｔ社」がそのウェブサイト上で公式に公開している、ノートパソコンの修理の方法や手続きに関するページが、文献１の方法で上位にランクされるとは限らない。ウェブサイトのトップページ以外のページが、他のページから極端に多くリンクされることは稀なので、ページの重要度（文献１のＰａｇｅＲａｎｋ）に有意な差が出にくいからである。

【００１３】同様に、文献２の方法でも、正しく検索できない恐れある。ユーザが所望するページが、「Ｔ社」「ノート」「パソコン」「修理」という４つの語を含むラベルで数多くリンクされているとは限らないからである。例えば、求めるページが「Ｔ社」サイトのトップページから、「パソコン」、「ＰＣカスタマーセンター」というラベルをリンクを順に辿ったところに公開されていて、このページ自体には「ＰＣカスタマーセンター」というラベルでしかリンクされていないかもしれない。逆に、「ノート」「パソコン」「修理」という３語を含むラベルで数多くリンクされた、「Ｔ社」と関係のないページが存在するかもしれない。

【００１４】一方、文献３や文献４で考案されている方法では、複数のハイパーテキストを、リンクによる参照関係に基づいて統合したページとして検索することができるとしている。しかしながら、インターネットには膨大な数のページがあり、ページ間のリンクの個数も、平均するとページの個数の数倍～十数倍もある。「Ｔ社」「ノート」「パソコン」「修理」という４つの語のいずれか一つでも含むウェブページはインターネット上に２００万ページ以上あり、これらの膨大なページについて、その各々を逐一リンクで統合して適合度を求めることは、現実的には不可能である。さらに、複数の文書を統合したものを検索対象とすれば、一般に、検索の再現率は向上するものの、適合率は低下することが知られている。もっと効率的で、かつ、よく制御された方法によって、ユーザの複雑な検索要求に適合するページを検索する必要がある。

【００１５】また、現実のリンクの構造は複雑であり、階層的な構造に整理されているとは限らない。ウェブページのリンクは、ページの内容上の関連を表すものだけでなく、ユーザのブラウジングのしやすさを助けることを目的としたリンクや、どちらからどちらを引用しているかわからないような相互的なリンク、ループになったリンクなどが多い。

【００１６】文献４ではループになったリンクを排除できるとしているものの、やはり、内容的な関連がほとんどないページを統合してしまう恐れが多分にある。また、複雑で個数の多いリンク構造を、文献３や文献４の方法によって、検索結果上で、ユーザが見やすい形に整理して提示することは困難である。

【００１７】

【発明が解決しようとする課題】以上説明したように、従来は、ウェブページ間のリンクによる参照関係に基づき検索する際には、あるウェブページが他のウェブページを直接参照する１段のリンクのみを用いていたため、複数の語からなる複雑な検索条件に適合するウェブページの検索が容易に行えないという問題点があった。

【００１８】そこで、本発明では、上記問題点に鑑みてなされたもので、大量のハイパーテキスト形式の文書の中から、複数の語からなる複雑な検索条件に適合する文書の検索が容易に行えるとともに、高い精度の検索結果が得られる情報検索方法およびそれを用いた情報検索装置を提供することを目的とする。

【００１９】本発明は、ユーザにより入力された主に自然文による、複雑で詳細な検索条件に対して、よく適合するハイパーテキスト形式の文書を検索するための、スケーラビリティのある方法、すなわち、大量の文書に対しても高速に検索できる方法を実現することを、第一の目的とし、さらに、検索結果をユーザが理解しやすい形で提示することを第二の目的とする。

【００２０】

【課題を解決するための手段】本発明は、複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索するためのものであって、前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、（ａ）前記複数の文書のそれぞれについて、その文書と前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある文書群から、前記参照関係を表す前記ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、（ｂ）前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との適合度を算出し、（ｃ）前記適合度に基づき、検索結果として表示する文書の順位を決定し、この順位に

従って前記検索結果としての文書を表示することを特徴とする。

【００２１】本発明によれば、１段のハイパーリンクにより参照関係にある文書群のみらなず、複数段のハイパーリンクを経て参照関係にある文書群からも、その参照関係を表すハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出して、検索対象の複数の文書のそれぞれについての検索時のインデックス（索引語）とすることにより、複数の語からなる複雑な検索条件に適合する文書の検索が容易に行えるとともに、高い精度の検索結果が得られる。

【００２２】また、本発明は、複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索するためのものであって、前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、（ａ）前記複数の文書のそれぞれについて、その文書と前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある文書群から、前記参照関係を表す前記ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、（ｂ）前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて前記ラベルから抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第１の適合度を算出し、この第１の適合度に基づき、検索結果として表示する文書の順位を決定し、（ｃ）前記複数の文書のそれぞれから、その文書の内容を表す語を抽出し、（ｄ）前記検索条件に含まれる複数の語と前記複数の文書のそれぞれについて、その内容から抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第２の適合度を算出し、この第２の適合度に基づき、検索結果として表示する文書の順位を決定し、（ｅ）前記第１および第２の適合度を統合した第３の適合度を算出し、この第３の適合度に基づき、検索結果として表示する文書の順位を決定し、（ｆ）前記第１～第３の適合度のそれぞれに対応して決定された順位のうちの少なくとも１つを用いて、検索結果としての文書を表示することを特徴とする。

【００２３】本発明によれば、複数段のハイパーリンクを経て参照関係にある文書群から、その参照関係を表すハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、それと各文書から抽出したその文書内容を表した語とを、検索対象の複数の文書のそれぞれについての検索時のインデックスとすることで、複数の語からなる複雑な検索条件に適合する文書の検索が容易に行えるとともに、より高い精度の検索結果が得られる。

【００２４】本発明は、複数の文書から、ユーザにより入力された自然文または複数の語からなる検索条件に適合する文書を検索するためのものであって、前記複数の文書のうちの任意の２つの文書間は１段のハイパーリンクでリンクされ、前記複数の文書のそれぞれは、予め定められた複数の文書グループのうちのうちの１つに属し、（ａ）前記複数の文書のそれぞれについて、その文書の属する文書グループ内から、前記１段のハイパーリンクおよび複数段のハイパーリンクを経て参照関係にある第１の文書群を抽出して、この各第１の文書群から、前記参照関係を表す前記ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出し、（ｂ）前記複数の文書のそれぞれについて、２つの前記文書グループ間にまたがって２つの文書をリンクする１段の文書グループ間ハイパーリンクおよび複数段の文書グループ間ハイパーリンクを経て文書グループ間の参照関係にある第２の文書群を前記複数の文書から抽出し、この各第２の文書群から、前記文書グループ間の参照関係を表す文書グループ間ハイパーリンクについて、それぞれに付されたラベルに含まれる語を抽出すし、（ｃ）前記検索条件に含まれる複数の語と、前記複数の文書のそれぞれについて前記文書グループ間ハイパーリンクのラベルから抽出された語とを比較して、前記複数の文書のそれぞれについて、前記検索条件との第１の適合度を算出すし、（ｄ）前記第１の適合度に基づき前記複数の文書グループから少なくとも１つの文書グループを検索対象として選択し、（ｅ）前記検索対象として選択された文書グループのそれぞれについて、前記検索条件に含まれる複数の語と、前記文書グループ内の文書のそれぞれについて前記ハイパーリンクのラベルから抽出された語とを比較して、前記文書グループ内の文書のそれぞれについて、前記検索条件との第２の適合度を算出し、（ｆ）前記第１の適合度に基づき検索結果として表示する文書グループの順位を決定するとともに、前記文書グループ毎に前記第２の適合度に基づき検索結果として表示する文書の順位を決定し、これら順位に従って前記検索結果としての文書グループと文書を表示することを特徴とする。

【００２５】本発明によれば、検索対象の複数の文書を複数の文書グループに分けて、文書グループ間ハイパーリンクを利用した検索と、各文書グループ内の検索とを組み合わせることにより、複数の語からなる複雑な検索条件に適合する文書の検索が容易に行えるとともに、より高い精度の検索結果が得られる。

【００２６】

【発明の実施の形態】以下、本発明の実施形態について図面を参照して説明する。

【００２７】本発明の情報検索システムが検索対象とする文書は、インターネットのウェブページなどに用いられるハイパーリンクで他の文書を結びつけることのできる文書（ハイパーテキスト文書）である。インターネットのウェブページが、本発明の効果をもっとも発揮できる対象であるので、ここでは、検索対象のハイパーテキスト文書の一例として、ウェブページを用いて説明を行

う。従って、以下の説明では、検索対象とするハイパーテキスト文書をページと呼び、文書間のハイパーリンクをリンクと呼び、文書のユニークな位置をＵＲＬと呼ぶ。なお、ここで、リンクとは、全て、２つのページ間をリンクするハイパーリンクのことを指す。

【００２８】（第１の実施の形態）図１は、第１の実施形態に係る情報検索システムの構成を表すブロック図である。図１において、ユーザインタフェース１は、ユーザがシステムに対して検索要求を入力し、また、システムがユーザに検索結果を提示するためのものである。

【００２９】検索語抽出部２は、特にユーザが検索条件を自然文で入力した場合、その自然文から検索に用いる語（ここでは検索語と呼ぶ）を抽出する処理を行うものである。

【００３０】リンク情報記憶部４には、検索対象である複数のページが、そのリンク構造に基づき検索可能なように、これらページから予め抽出されたリンク情報が記憶されている。

【００３１】リンク検索部３は、検索語抽出部２で抽出された検索語とリンク情報記憶部４に記憶されているリンク情報とを比較して、適合する文書を検索する。

【００３２】ページ情報記憶部７には、検索対象である複数のページを、その各々の内容自体から検索可能なように、ページ情報が記憶されている。

【００３３】ページ検索部６では、検索語抽出部２で抽出された検索語とページ情報記憶部７に記憶されたページ情報とを比較して、適合する文書を検索する。

【００３４】検索結果統合部５では、リンク検索部３およびページ検索部６での検索結果をユーザが所望する形に統合して、表示用データを生成する処理を行うものである。

【００３５】ウェブ情報収集部１１は、インターネット、かつ、または、イントラネットから所定のウェブ頁を収集する手段である。これは、一般にロボット、クローラ、あるいはスパイダーなどと呼ばれるプログラムであり、ウェブページのハイパーリンクを再帰的に辿って、それぞれのページの内容や情報を収集する。この手段は従来技術に属するものである。

【００３６】リンク索引作成部８は、ウェブ情報収集部１１によって得た個々のウェブページに記述されたハイパーリンクについて、そのリンク先ＵＲＬとアンカーテキストを抽出し、リンク情報記憶部に記憶せしめる。また、リンクのアンカーテキストから索引語の単語ベクトル、すなわち、リンクのベクトルを作成する処理を行う。

【００３７】ページ索引作成部９は、ウェブ情報収集部１１によって得た個々のウェブページから、その内容、すなわち、タイトルや本文などの文章部分を抽出し、ページ情報記憶部に記憶せしめる。また、ページ内容から索引語の単語ベクトル、すなわち、ページの内容のベクトルを作成する処理を行う。

【００３８】ここで、ページ検索部６およびページ情報記憶部７は、本発明の情報検索システムに必須の構成要素ではなく、これらページ検索部６およびページ情報記憶部７を含めずにシステムを構成することも可能である。この場合、検索結果統合部５では、検索結果を統合する処理を行う必要はなく、リンク検索部３での検索結果から所定の表示用データを生成する処理を行う。

【００３９】図２、図３、図４は、上記リンク情報記憶部４およびページ情報記憶部７に記憶されているデータの記憶例を示したものである。

【００４０】図２は、本システムの検索対象である各ページのＵＲＬと、当該ページに与えられた本システム内でユニークな識別子（ＩＤ）との対応関係が記述されたデータである。このデータは、上記リンク情報記憶部４とページ情報記憶部７のいずれか一方に記憶されていればよく、また、これらとは別個の他の記憶部を設けて記憶されていてもよい。なお、ＵＲＬは通常、圧縮してデータ量を減じたり、トライ構造のような効率よくアクセスできる形式にして記憶する。

【００４１】図３は、リンク情報記憶部４に記憶されているリンク情報の記憶例を示したものである。図３に示すリンク情報では、１つのリンクを、リンク元であるページのＩＤと、リンク先であるページのＩＤと、当該リンクのラベル（リンクのラベルは、例えば、ＨＴＭＬ文書の場合、アンカーテキストと呼ばれているものである）との、三者によって表現されるている。リンク元ＩＤとリンク先ＩＤは、図２で説明したＩＤと同じである。また、ラベルについては、図３ではラベルとして記述されている文字列をそのまま図示しているが、形態素やＮグラムなどの部分文字列に分割し、転置ファイルの形式で索引が付されて記憶するようになっていてもよい。また、語の頻度や出現位置なども併せて記憶するようにしてもよい。文字列を索引を付して記憶する技術は従来技術に属するので説明は省略するが、本発明の要旨に関わる検索モデル、すなわち、ユーザの検索要求と文書との適合度を求める方法については、後に詳述する。

【００４２】図４は、ページ情報記憶部７に記憶するデータの例を表す図である。図のように、ページは、そのＩＤと、タイトルと、本文との三者で表現される。ＩＤは、前述の図２および図３の説明同様、ページをシステム内部でユニークに表現するためのＩＤである。

【００４３】タイトルと本文は、例えばＨＴＭＬ文書の場合は、タイトルタグおよびボディタグで指定された文字列である。なお、ＨＴＭＬ文書のボディタグの内部には、図表などを指定するタグが埋め込まれているが、ここではこれらのタグをパージングし、不要な部分を除いて記憶する。これら、タイトル、本文は、前述の図３のラベルと同様、転置ファイルなどの形式で索引を付して記憶するようにしてもよい。

【００４４】また、図２、図３のデータに加えて、他の属性、例えば、ページの更新日時などの情報を、必要に応じて記憶するようにしてもよい。

【００４５】図５は、リンクによる複数のページ間の参照関係を示したものである。図５中、「Ｐ１」、「Ｐ２」…などはページのＩＤを表し、矢印はページ間のリンクを表し、リンクに付された文字列はリンクのラベルを表す。

【００４６】例えば、ページ「Ｐ１」は、「お問合せ一覧」というラベルを付されたリンクによってページ「Ｐ３」を参照しており、逆に、ページ「Ｐ３」は、「ホーム」というラベルを付されたリンクによってページ「Ｐ１」を参照している。

【００４７】ページ「Ｐ１」は、ハイパーリンクにて直接ページ「Ｐ３」を参照している。この場合のハイパーリンクを１段のハイパーリンク、あるいは簡単に、１段のリンクと呼ぶ。

【００４８】ページ「Ｐ１」はページ「Ｐ７」から「Ｔ社トップ」というラベルを付されたリンクによって参照されている。この場合、ここでは、ページ「Ｐ７」はページ「Ｐ１」を介してページ「Ｐ３」と参照関係にあるということである。また、ページ「Ｐ７」とページ「Ｐ３」は、２段のリンクで参照関係にあるということである。

【００４９】インターネットのウェブページは、数十億ページとも言われる膨大なページが、このようなリンクによって複雑に接続されている。これをグラフ構造と呼ぶ。

【００５０】次に、図６～図８に示すフローチャートを参照して、図１の情報検索システムの処理動作について説明する。

【００５１】図８は、ユーザから入力された検索要求に適合するページを検索するための処理動作を説明するためのフローチャートである。

【００５２】図６，図７は、検索のための前処理を説明するためのフローチャートであり、ユーザから入力された検索要求とページとの適合度を求めるために用いるデータ、すなわち、ベクトルを作成する処理動作について説明するためのものである。これらの処理は、リンク情報、ページ情報を、リンク情報記憶部４、ページ情報記憶部７にそれぞれ記憶する際に実行する処理であり、これらの処理結果のデータ（すなわち、図６のステップＳ６で算出された各ページのリンクのベクトルＶｐ１（図１４の文書グループ間のリンクのベクトルＶｐ１ｅｘを含む）、図７のステップＳ１３で算出されたページ内容のベクトルＶｐｔなど）は、リンク情報記憶部４、ページ情報記憶部６に記憶するものである。

【００５３】図６で示した処理動作は、リンク構造に着目した検索モデルによってページをモデル化する処理である。ここではベクトル空間モデルを用いて実施しているが、他の検索モデルであってよい。

【００５４】図６に示した処理は、主に、リンク索引作成部８で実行される処理である。まず、ステップＳ１～ステップＳ５で、ページ情報記憶部７に記憶されている全てのページｐについて（ステップＳ１）、各ページｐを参照している（すなわち、ページｐをリンク先とする）リンク情報記憶部４に記憶されている全てのリンク１のラベルを調べ（ステップＳ２）、各リンク１の（ラベルの）ベクトルを作成し（ステップＳ４）、さらに、各ページｐについて、そのページｐを参照するリンクのベクトルｖ１の総和Ｖｐ１（１）を作成する（ステップＳ５）。

【００５５】すなわち、リンク索引作成部８は、ある１つのページｐを選択し、そのページｐを参照するリンクをリンク情報記憶部４を検索する。そして、１または複数個見つかれば、その１つ１つのリンク１に対し、当該リンクに付されたラベルを形態素解析して、当該ラベルから索引語を抽出する（ステップＳ３）。例えば、ラベルを形態素解析した結果得られる自立語の中から、さらに接続詞、感動詞や、その他、検索には不要であると予め定められた語など（以下、これらを不要語と呼ぶ）を取り除いて、索引語を抽出する。

【００５６】たとえば、「安くて便利なパソコン販売のページ」というラベルに対し形態素解析を行うと、「安」「く」「て」「便利」「な」「パソコン」「販売」「の」「ページ」といった形態素が解析結果として得られる。このうち、活用語尾や助詞などは自立語ではないので除かれる。また、「～のページ」という表現はウェブページのリンクに特有な表現であり、検索要求とは無関係の場合が多いので、不要語として除く。結果として、「安くて便利なパソコン販売のページ」というラベルから、「安」「便利」「パソコン」「販売」という４つの語が索引語として抽出される。

【００５７】次に、ステップＳ４では、ステップＳ３で抽出された索引語の重みを決定する。これには、一般にＴＦ・ＩＤＦと呼ばれる方法を用いる。すなわち、あるテキスト（この場合はラベル）におけるある語の重みは、そのテキストに含まれるその語の個数（ＴＦ）と、全テキストのうちその語を含むテキストの個数（ＤＦ）とによって求めることができる。ＴＦが大きいほど重みは大きくなり、ＤＦが大きいほど重みは小さくなる。一方、形容詞「安（く）」、形容動詞「便利（な）」、名詞「パソコン」などの品詞の違いを重みに反映させることも行う。一般に、名詞や固有名詞の重みを、動詞や形容詞、形容動詞などと比較して大きくすると検索精度が向上することが知られている。なお、ここでは、自立語を索引語としたが、自立語に加え、格フレームなどを索引とし、これをベクトルで表現することも可能である。

【００５８】以上の処理によって、リンク１のベクトルｖ１（ラベルに含まれる語とその重みをベクトルで表現

したデータ）が作成できる。

【００５９】例えば、図５において、ページ「Ｐ３」を参照するページ「Ｐ６」をリンク元とするリンクに付された「その他の製品に関するお問い合わせ」というラベルから、索引語として、「その他」「製品」「問い合わせ」という３つの語が索引語として抽出され、そのそれぞれに対し、重みが、「ｇ１」「ｇ２」「ｇ３」と、求められたとする。この場合、当該ラベルをもつリンク１のベクトルｖ１は、（その他、製品、問い合わせ）＝（ｇ１、ｇ２、ｇ３）となる。

【００６０】選択されたページｐについて、そのページを参照するリンク１の全てに対し、ベクトルｖ１を作成したら、次に、ステップＳ５へ進み、リンク検索部３は、当該選択されたページｐについて、そのページを参照するリンクのベクトルの総和を作成する。

【００６１】ページｐを参照するリンクのベクトルの総和Ｖｐ１（１）とは、ページｐを直接（１段で）参照する（リンク情報記憶素４に記憶されている全ての）リンクのベクトルｖ１から作成したベクトルである。

【００６２】例えば、例えば、図５において、ページ「Ｐ３」には２つのリンクにより参照され、その２つのリンクには、それぞれ「その他の製品に関するお問い合わせ」というラベルと、「お問合わせ先一覧」というラベルが付されている。この場合、各リンクのベクトルは、（その他、製品、問い合わせ）＝（ｇ１、ｇ２、ｇ３）と、（問合わせ、先、一覧）＝（ｇ４，ｇ５、ｇ６）であるとする。このとき、当該ページ「Ｐ３」を参照するリンクのベクトル総和Ｖｐ１（１）は、（その他、製品、問い合わせ、先、一覧）＝（ｇ１、ｇ２、ｇ３＋ｇ４、ｇ５，ｇ６）となる。

【００６３】選択された１つのページｐに対し、上記ステップＳ３～ステップＳ５の処理を行って、当該選択されたページｐについて、そのページを１段で参照するリンクのベクトル総和Ｖｐ１（１）を作成したら、ステップＳ１へ戻り、他のページを選択して、上記同様にして、当該ページを１段で参照するリンクのベクトル総和を作成する処理を、ページ情報記憶部７に記憶されている全てのページに対し行う（ステップＳ１）。

【００６４】次に、ステップＳ６へ進み、リンク検索部３は、ページ同士のリンクの参照関係に基づいて、各ページｐを複数段のリンクで参照する他のページがあるときは、その全てのリンクのベクトルの総和、すなわち、ベクトルＶｐ１（ｎ）を作成する。

【００６５】ここでは、ある１つのページ（第１のページ）が他のページ（第２のページ）からリンク（第１のリンク）にて直接参照されている場合、第１のページは、第２のページから１段のリンクで参照されていると表現し、第２のページがさらに他のページ（第３のページ）からリンク（第２のリンク）にて直接参照されている場合、第１のページは、第３のページから２段のリンクで参照されていると表現する。

【００６６】ステップＳ６で作成しようとしているページｐのベクトルＶｐ１（ｎ）は、ページｐがｎ段のリンクで他のページから参照されている場合に、その全てのリンクのベクトルの総和を求めることにより作成することができる。ベクトルＶｐ１（ｎ）を、以下簡単に、ページｐのｎ段のリンクのベクトルと呼ぶ。これに対応して、ベクトルＶｐ１（１）をページｐの１段のリンクのベクトルと呼ぶ。

【００６７】例えば、ここでは、段数に応じた異なる重みαをつけて和をとったベクトルＶｐ１（ｎ）を例えば次式（１）に従って作成する。なお、式（１）では、ページｐに至るまでの複数段のリンクの各ベクトルに含まれている索引語のそれぞれに対応する重みに、段数に応じて異なる重みαを乗じて総和を求めている。ページｐのｎ段のリンクのベクトルは、各リンクのベクトルに含まれている索引語の並びと、その各索引語に対する式（１）で求めた重みの値で表されるものである。

【００６８】

【数１】

$$Vpl(n) = Vpl(n-1) + (\alpha) \times \Sigma(r \rightarrow p)\, V(r)l(n-1) + \beta \times \Sigma(p \rightarrow r)\, Vrl(n-1) \quad \cdots(1)$$

式(1)において、ｒ → ｐとは、ページｒがページｐを一段のリンクで参照するということを意味し、Σ(ｒ → ｐ)は、そのようなページｒについて、ｎ － １段のリンクのベクトルの和を(1)式を用いて求めることを意味する。

【００６９】式（１）において、αとβは係数であるが、どちらも例えば「０」以上「１」未満とする。αは、ページｐに向かうリンクに関する係数であり、βは、ページｐから出ていくリンクに関する係数であるため、βよりもαの値を大きくする。βは「０」であってもよい。

【００７０】なお、式（１）において、係数βを「０」とした場合、ページｐのｎ段のリンクのベクトルは、ページｐに向かう方向のリンクのみから作成されるものである。以下、説明の簡単のため、係数βが「０」の場合について説明する。

【００７１】式（１）を用いて、ページｐがｎ段のリン

クにて参照されているときは、式（１）を用いた計算をｎ回繰り返すことにより、ページｐに対する、ｎ段のリンク構造を反映したベクトルを求めることができる。

【００７２】例えば、図５のページ「Ｐ７」の場合、上記ステップＳ１～ステップＳ６を経た結果、ページ「Ｐ７」の３段のリンクのベクトルに含まれる索引語は、例えば、「ノート」「ＰＣ」「超」「薄型」「家庭」「向け」「パソコン」「周辺」「機器」であり、その各索引語の重みは式（１）を用いて計算した結果、それぞれ「ｇ１１」「ｇ１２」「ｇ１３」「ｇ１４」「ｇ１５」「ｇ１６」「ｇ１７」「ｇ１８」「ｇ１９」であるとすると、ページ「Ｐ７」のリンクのベクトルは、（ノート、ＰＣ、超、薄型、家庭、向け、パソコン、周辺、機器、Ｔ社、修理、拠点、…）＝（ｇ１１、ｇ１２、ｇ１３、ｇ１４、ｇ１５、ｇ１６、ｇ１７、ｇ１８、ｇ１９、ｇ２０、ｇ２１、ｇ２２、…）となる。

【００７３】ページ「Ｐ７」の場合、「ノート」「ＰＣ」などの索引語は、ページ「Ｐ７」に近い段数のリンクのラベルに含まれ、かつ、頻度が多いので、重みが大きくなる。

【００７４】なお、式（１）の係数αの値は、例えば、対象としているページｐ（例えば、ページ「Ｐ７」）に近い段数のリンクほど大きい値となるように定めてもよい。すなわち、ページ「Ｐ７」を直接参照している１段目のリンクを加算するときには、αを最も大きくする。

【００７５】ページｐのｎ段のリンクのベクトルＶｐ１（ｎ）を計算する際の段数ｎは、検索システムの目的や要求される検索精度に応じて設定すればよい。ｎを大きくするほど、ベクトルの語の数が増えることになるが、多くてもｎ＝５程度でよく、ｎ＝２か３でも実用的な検索が可能であることが分かっている。以下、ページｐのｎ段のリンクのベクトルＶｐ１（ｎ）の表記を、単にＶｐ１と記述する。

【００７６】図７に示すフローチャートは、ページの内容に着目してページ内容のベクトルを作成するための処理動作を示したもので、ページ索引作成部９での処理動作を示したものである。

【００７７】ページ情報記憶部７に記憶されている全てのページｐについて（ステップＳ１１）、その内容（すなわち、図４に示したタイトルと本文）を、形態素解析し、図６のステップＳ３において、ラベルから索引語を抽出するのと同様にして、ページｐの内容から索引語を抽出する（ステップＳ１２）。そして、図６のステップＳ４の説明と同様にして、各索引語の重みを求め、ページ情報記憶部７に記憶されている各ページについて、ページ内容のベクトルＶｐｔを作成する（ステップＳ１３）。

【００７８】なお、図７に示した処理自体は従来技術に属するものである。

【００７９】次に、図８に示すフローチャートを参照して、ユーザが検索条件ｑを入力したときに、その入力された検索条件に適合するページを検索するための処理動作について説明する。

【００８０】ここで、ユーザにより入力される検索条件ｑとは、ページを検索するためのキーワード（語）が複数含まれるものであって、自然文、または複数の語を羅列したもの、複数の語を論理式で結合したものなどである。

【００８１】ユーザにより検索条件ｑが入力される（ステップＳ２１）。検索語抽出部２では、まず、これを形態素解析して、例えば、図６のステップＳ３で索引語を抽出するのと同様にして、検索語を抽出する（ステップＳ２１）。すなわち、検索条件ｑを形態素解析した結果得られる自立語の中から、さらに不要語を取り除いて、検索語を抽出する。そして、各検索語の重みを図６のステップＳ４の説明と同様にして求め、さらに、図６のステップＳ５の説明と同様にして、検索条件ｑのベクトルＶｑを作成する（ステップＳ２３）。

【００８２】以上のようにして作成された検索条件ｑのベクトルＶｑを用いて、ページ情報記憶部７に記憶されている全てのページに対し、以下のステップＳ２５、ステップＳ２６の処理を実行する。

【００８３】なお、ページ情報記憶部７に記憶されている全てのページの中から、検索条件ｑのベクトルＶｑに含まれる検索語を少なくとも１つ含むページを予め検索し、検索結果として得られたページを処理対象として、ステップＳ２５、ステップＳ２６の処理を実行するようにしてもよいし、処理の高速化のため、他と比べて少ない個数の検索語しか含まないページについては、検索条件との適合度が他と比べて小さくなると見込まれるので、それらについては処理を省略してもよい。

【００８４】また、ステップＳ２５とＳ２６は、後述するように、ユーザの要求や使い方に応じて一方を省略してもよい。

【００８５】以下、ここでは、検索条件ｑのベクトルＶｑに含まれる検索語を少なくとも１つ含むページをステップＳ２５，ステップＳ２６の処理対象とした場合を例にとり説明を行う。

【００８６】ステップＳ２５では、図６のステップＳ６で求めたページｐのｎ段のリンクのベクトルＶｐ１（すなわち、ページのリンク構造に着目して作成したベクトル）と、検索条件のベクトルＶｑとを比較し、その類似度を求める。類似度の算出方法としては、一般に、ベクトルの内積や余弦をとる方法がよいとされている。こうして求めたＶｐ１とＶｑとの類似度を、検索条件ｑに対するページｐのリンク構造に基づく適合度Ｓ１（ｐ，ｑ）とする。

【００８７】同様にして、ステップＳ２６では、図７の処理で求めた各ページのページ内容のベクトルＶｐｔ（すなわち、ページの内容に着目して作成したベクト

ル）と、検索条件のベクトルＶｑとを比較して、検索条件ｑに対するページｐのページ内容に基づく適合度Ｓｔ（ｐ，ｑ）を求める。

【００８８】処理対象の各ページから、ページのリンクのベクトルと検索条件ｑのベクトルとの類似度（リンク構造に基づく適合度Ｓ１（ｐ，ｑ））と、ページ内容のベクトルと検索条件ｑのベクトルとの類似度（ページ内容に基づく適合度Ｓｔ（ｐ，ｑ））が算出されたら、次に、ステップＳ２７～ステップＳ２９の検索結果を表示するための処理を行う。なお、ステップＳ２７～ステップＳ２８の処理は、後述するように、ユーザが検索条件などの入力の段階で、どの検索方法を選択したかにより省略される場合もある。

【００８９】ステップＳ２７では、リンク構造に基づく適合度Ｓ１（ｐ，ｑ）に基づき、ページの順位を付けて、それを検索結果として表示する。

【００９０】ステップＳ２８では、ページ内容に基づく適合度Ｓｔ（ｐ，ｑ）に基づきページに順位を付けて、それを検索結果として表示する。

【００９１】ステップＳ２９では、適合度Ｓ１（ｐ，ｑ）、Ｓｔ（ｐ，ｑ）を統合した適合度Ｓ（ｐ，ｑ）を各ページについて算出する。そして、この適合度Ｓ（ｐ，ｑ）に基づいて、各ページを順位付けしたものを検索結果としてユーザに提示する。

【００９２】以下、ステップＳ２９の処理について説明する。

【００９３】たとえば、適合度Ｓ１（ｐ，ｑ）、Ｓｔ（ｐ，ｑ）から、これらを統合した適合度Ｓ（ｐ，ｑ）を算出するには、次式（２）を用いればよい。

【００９４】Ｓ（ｐ，ｑ）＝Ｃ１×Ｓ１（ｐ，ｑ）＋Ｃｔ×Ｓｔ（ｐ，ｑ）　　…（２）なお、式（２）において、Ｃ１、Ｃｔは、予め定められた定数で、適合度Ｓ（ｐ，ｑ）に占める適合度Ｓ１（ｐ，ｑ）、Ｓｔ（ｐ，ｑ）のそれぞれの比率、すなわち、重要度を定めるものである。

【００９５】また、ここで、ページｐを検索結果に含めてよいかどうかの判定は、予め定められた閾値との比較によって行う。すなわち、リンクに基づく適合度Ｓ１（ｐ，ｑ）については、これが閾値Ｓ１ｍｉｎ以上であれば、ページｐを検索結果に含めてよいとする。ページ内容に基づく適合度Ｓｔ（ｐ，ｑ）についても同様に、閾値Ｓｔｍｉｎ以上であれば、ページｐを検索結果に含めてよいとする。

【００９６】統合した適合度Ｓ（ｐ，ｑ）についても同様に、閾値Ｓｍｉｎより大きければページｐを検索結果に含めてよいする。

【００９７】Ｃ１とＣｔは、各々定数である。閾値Ｓ１ｍｉｎ、Ｓｔｍｉｎ、Ｓｍｉｎのいずれかを「０」に設定すれば、その閾値での判定は行わないことになる。また、Ｃ１とＣｔのいずれかを「０」に設定すれば、Ｓ１（ｐ，ｑ）あるいはＳｔ（ｐ，ｑ）は、統合された適合度Ｓ（ｐ，ｑ）の値には反映されないことになる。

【００９８】図９～図１１は、図１の情報検索システムのユーザインタフェース１の画面表示例を示す図である。

【００９９】図９に示した画面は、検索要求を入力する領域２０１からなる入力画面である。領域２０１には、ユーザが検索条件を入力する領域１０１と、検索方法を指定する領域１０２からなる。

【０１００】ユーザは、図１の情報検索システムに検索要求を行う場合には、領域１０１に、例えば「Ｔ社のパソコン」といった自然文で記述した検索条件ｑを入力する。領域１０２は、ユーザが検索方法（図９では、「リンク構造で検索」「ページ内容で検索」「両方の検索結果を個別に表示」「両方の検索結果を総合して表示」の４つがある）を指定するための領域である。

【０１０１】図９に示した上記４つの検索方法とは、それぞれ、（１）リンク構造で検索する方法、（２）ページ内容で検索する方法、（３）リンク構造での検索結果とページ内容での検索結果をそれぞれ個別に表示する方法、（４）リンク構造での検索結果とページ内容での検索結果を統合して表示する方法である。

【０１０２】ユーザは、領域１０１に検索条件を入力し、上記４つの検索方法から所望の方法を選択した後、「検索」ボタン１０３をマウス等で選択する（押す）ことにより、図８に示した検索処理が実行される。すると、ユーザインタフェースには、図１０～図１１に示すような検索結果が表示される。

【０１０３】図１０に示した画面は、図９に示した入力画面からユーザが、検索方法として、「リンク構造で検索」「ページ内容で検索」「両方の検索結果を個別に表示」のいずれかを選択したときの検索結果の表示方法を説明するための図である。なお、図１０に示した画面表示例そのものは、検索方法として「両方の検索結果を個別に表示」が選択されたときの検索結果の表示例を示したものである。

【０１０４】図１０に示した画面は、大きく分けて３つの領域に分かれている。１つは、検索要求を入力する領域２０１であり、他の１つは、リンク構造に基づく検索結果を表示する領域２０２であり、さらに他の１つは、ページ内容に基づく検索結果を表示する領域２０３である。

【０１０５】検索方法のうち、（１）リンク構造で検索する方法とは、前述した適合度Ｓ１（ｐ，ｑ）のみに基づいて検索結果を求める方法である。この方法が選択された場合には、図８のステップＳ２８，ステップＳ２９の処理は省略してもよい。ユーザインターフェース１には、検索結果として領域２０２に示したような、リンクに基づく検索結果が表示される。なお、このとき、領域２０３のページ内容に基づく検索結果は表示されない。

【０１０６】検索方法のうち、（２）ページ内容で検索する方法とは、前述した適合度Ｓｔ（ｐ，ｑ）のみに基づいて検索結果を求める方法である。この方法が選択された場合には、図８のステップＳ２７，ステップＳ２９の処理は省略してもよい。ユーザインターフェース１には、検索結果として、領域２０３に示したような、ページ内容に基づく検索結果が表示される。なお、このとき、領域２０２のリンク構造に基づく検索結果は表示されない。

【０１０７】検索方法のうち、（３）リンク構造での検索結果とページ内容での検索結果をそれぞれ個別に表示する方法とは、リンク構造による検索（すなわち適合度Ｓ１（ｐ，ｑ）に基づく検索）と、ページ内容による検索（すなわち適合度Ｓｔ（ｐ，ｑ）に基づく検索結果）とを両方行い、それぞれの検索結果を領域２０２，２０３に表示する方法である。この方法が選択された場合には、図８のステップＳ２９を省略してもよい。検索結果は、領域２０２と領域２０３に表示される。

【０１０８】図１１に示した画面は、図９に示した入力画面からユーザが、検索方法として、「両方の検索結果を総合して表示」を選択したときの検索結果の表示例を示したものである。

【０１０９】両方の検索結果を総合して表示する検索方法が選択されたときは、図８に示したフローチャートに従って、ステップＳ２９までの処理を全て実行して、式（２）のＣ１、Ｃｔを適宜指定して（あるいは、予め定められた値をそのまま用いてよい）求めた適合度Ｓ（ｐ，ｑ）に基づいた検索結果を含めた検索結果の表示を行う方法である。なお、図１１については、後述する。

【０１１０】図１０の領域２０２には、リンク構造に基づいた検索結果が表示される。順位１０５の高い方から順に、検索されたページの見出し１０７が並べられている。なお、順位１０５は上記Ｓ１（ｐ，ｑ）の大きい順に、１位、２位、…と検索された各ページに与えたものである。また、ページの見出し１０６は、図１のページ情報記憶部７に記憶されたページのタイトル（図４のタイトル）を表示してもよいが、ページ情報記憶部７がない場合や、当該ページの情報をシステムが取得していない場合は、図１のリンク情報記憶部４に記憶した当該ページを参照するリンクのラベルのうち、代表的なもの（例えば、検索条件に最も合致するもの）を見出し１０６として用いてもよい。また、見出し１０６の文字列の中で検索条件に関連する部分、例えば、検索条件に「Ｔ社」「パソコン」などの語が含まされているとき、見出し１０６に含まれているこれらの語は強調して表示する。

【０１１１】ページの見出し１０６に、記号「＜」にて追加されている情報は、見出し１０６に対応するページをリンクで参照するリンク元ページの見出し１０７である。この見出し１０７には、この見出し１０７に対応するページへジャンプするリンクが埋め込まれていて、この見出し１０７をマウス等でクリックすれば、当該ページが表示可能になっている。リンク元ページは複数あり得るが、ここでは、そのうち、Ｓ１（ｐ，ｑ）が最も大きいページを１つ表示することとする。

【０１１２】一方、ページの見出し１０６に、記号「＞」にて追加されている情報は、見出し１０６に対応するページがリンクで参照するリンク先ページの見出し１０８である。この見出し１０８には、見出し１０８に対応するページへジャンプするリンクが埋め込まれていて、この見出し１０８をマウス等でクリックすれば、当該ページが表示可能になっている。リンク先ページは複数あり得るが、これについても、Ｓ１（ｐ，ｑ）が大きいものを複数（例えば５つまで）表示することにする。

【０１１３】このように、検索結果のページの見出し１０６に対して、これとリンクにより参照関係にあるページの見出し１０７，１０８を表示することにより、ユーザは、見出し１０６に対応するページの内容自体を見なくても、見出し１０６に対応するページと参照関係にあるページのうち所望のページを直接アクセスして、見出し６に対応するページがどのような位置付けのページなのかを理解することも容易になる。

【０１１４】一般に、リンク構造は必ずしも階層構造に整理されているわけではないが、上述した方法で、ユーザの検索要求によく適合するリンクを選択的に表示すれば、リンク構造の複雑さによるユーザの混乱は避けられる。

【０１１５】図１０の領域２０２には、ページ内容に基づいた検索結果が表示される。順位１１０の高い方から順に、検索されたページの見出し１１１が並べられている。なお、見出し１１１は、検索結果のページのタイトルであり、順位１１０は、上記Ｓｔ（ｐ，ｑ）が大きい順に、１位、２位、…と検索された各ページに与えたものである。

【０１１６】ページの見出し１１１の下には、当該ページの内容の要約１１２が表示されている。要約１１２は、ここでは、当該ページの本文から検索条件に合致する表現、すなわち、例えば、検索条件に含まれている「Ｔ社」「パソコン」等の語をよく含む部分（文など）を抜きだして表示する。

【０１１７】また、見出し１１１に対応する検索結果のページから他のページを参照するリンクがある場合は、そのうち、検索条件と関連するリンク１１３が見出し１１１に対応させて表示されている。例えば、見出し１１１に対応するページの本文に「デスクトップパソコン」「ノートパソコン」「周辺機器」「ソフトウェア」などのラベルをもつリンクがある場合にも、ユーザの検索条件に関連のあるリンク「デスクトップパソコン」「ノートパソコン」のみがリンク１１３としてに表示される。

すなわち、これらリンク１１３をマウス等でクリックすると、当該リンクにて参照している他のページが表示されるようになっている。これにより、ユーザは、見出し１１１に対応するページの内容自体を見ずとも、見出し１１１に対応するページから参照されているページのうち所望するものに直接アクセスすることができる。

【０１１８】次に、図１１について説明する。

【０１１９】図１１は、リンク構造に基づく検索と、ページ内容に基づく検索との両方の検索結果を１つに統合した検索結果を、ユーザが所望する方法で順位付けてユーザに提示する場合の画面表示例である。

【０１２０】上述の通り、検索により得たページには、リンク構造に基づく適合度Ｓｌ（ｐ，ｑ）と、ページ内容に基づく適合度Ｓｔ（ｐ，ｑ）と、Ｓｌ（ｐ，ｑ）とＳｔ（ｐ，ｑ）とを総合して求めた適合度Ｓ（ｐ，ｑ）がある。ユーザが、図９の入力画面において、検索方法として、「両方の検索結果を統合して表示」を選択したときには、これらＳｌ（ｐ，ｑ）、Ｓｔ（ｐ，ｑ）、Ｓ（ｐ，ｑ）のうちいずれの適合度によっても検索結果の並び替えが行えるように、領域２０４にて、並び替え方法を選択するようになっている。

【０１２１】並び替え方法の選択肢としては、図１１に示すように、「リンク構造に基づく適合度（Ｓｌ（ｐ，ｑ））で並び替え」、「ページ内容に基づく適合度（Ｓｔ（ｐ，ｑ））で並び替え」、「統合された適合度（Ｓ（ｐ，ｑ））で並び替え」がある。

【０１２２】例えば、統合された適合度によれば１位のページは、リンク構造に基づく適合度やページ内容に基づく適合度では、２位以下のこともある。統合された適合度で並び替えを行うと、当該ページは一番上に表示されるが、他の適合度で並び替えを行うと、２番目以下に表示される。

【０１２３】ユーザは、例えば、統合された適合度Ｓ（ｐ，ｑ）で並び替えを行いたい場合には、３番目の選択肢を選択する。そして、比率設定領域１３２において、統合された適合度を求めるための式（２）の定数Ｃ１、Ｃｔに対応する、統合された適合度中に占めるリンク構造に基づく適合度Ｓｌ（ｐ，ｑ）の比率、ページ内容に基づく適合度Ｓｔ（ｐ，ｑ）の比率をそれぞれ設定する。その後、「並び替え」ボタン１３４を押すことよって並び替えの実行を指示する。

【０１２４】例えば、比率設定領域１３２で設定されたＳｌ（ｐ，ｑ）とＳｔ（ｐ，ｑ）の比率が６０％対４０％であったときは、式（２）において、Ｃ１＝０．６、Ｃｔ＝０．４として統合された適合度Ｓ（ｐ，ｑ）を求めた結果、この値の大きい順に順位が設定され、この設定された順位の順に検索結果が領域２０５に表示される。

【０１２５】領域２０５に表示されている検索結果は、領域２０４で並び替え方法として、３番目の選択肢が選択されたときの表示例である。「統合適合度」「リンク構造適合度」「ページ内容適合度」と付された欄１３５，１３６，１３７には、それぞれ、適合度Ｓ（ｐ，ｑ）、Ｓｌ（ｐ，ｑ）、Ｓｔ（ｐ，ｑ）に基づくページの順位が表示されている。

【０１２６】例えば、「統合適合度」が１位で、「リンク構造適合度」が１位で、「ページ内容適合度」が７位である一番上に表示されているページの見出し、すなわち、ここでは、タイトルは、「Ｔ社ＰＣウェブ」であり、この見出しとともに、前述同様、当該ページの内容の要約と、当該ページから他のページを参照するリンクがある場合は、そのうち、検索条件と関連するが表示されている。

【０１２７】以上説明したように、第１の実施形態の情報検索システムは、リンク構造に基づいた検索を行うことに特徴があり、このような検索手法によって、ユーザの検索条件によく適合する文書を効率よく検索できる。複数段のリンク構造を反映した検索を行うため、自然文などで記述されたユーザの複雑な検索要求に対しても、その要求に合致する検索結果を求めることができる。

【０１２８】なお、リンク構造のみによっても検索が可能であるが、上述のように、ページ内容に基づいた従来型の検索方式と統合した方法で検索を行うことも可能である。その統合の方法は柔軟であるため、ユーザは、リンク構造に基づく検索、すなわち、他のページから検索条件によく適合する表現で多くリンクされているページを検索する方法と、ページ内容に基づく検索、すなわち、内容自体が検索条件によく適合する表現で記述されているページを検索する方法を、目的に応じて自由に使い分けることができる。さらに、上述のように、検索結果のページとともに、これと参照関係にある他のページをユーザの検索条件を反映した形で整理して提示するため、検索結果の理解や利用が容易である。

【０１２９】（第２の実施の形態）第２の実施形態に係る情報検索システムは、検索方法としては、第１の実施形態と類似した方法をとるが、ハイパーテキスト形式の文書を、複数の文書グループに分けて考えることにより、検索性能をさらに向上させることに特徴がある。

【０１３０】例えば、インターネット上のウェブページは、個々のページの一つ一つが個別の場所に存在するのではなく、サイトやドメインなどと呼ばれる管理単位でまとめて配置されている。このような文書のまとまりを、ここでは文書グループと呼ぶことにする。当然ながら、同一の文書グループに属するページは互いに内容が類似していたり、意味的な関連性が強い場合が多い。

【０１３１】また、同一の文書グループに属するページ間のリンクと、別々の文書グループに属するページ間のリンクとは、性質や意味が異なる。

【０１３２】図１３は、ページ間の参照関係を、文書グループの概念を加えて摸式的に表現した図である。図１

３において、文書グループは、例えば、文書グループＤ１～Ｄ３の３つである。また、図５と同様に、図１３中、「Ｄ１－１」「Ｄ１－２」…「Ｄ２－１」「Ｄ２－２」…「Ｄ３－１」「Ｄ３－２」…は、ページのＩＤを表し、矢印はページ間のリンクを表し、リンクに付された文字列はリンクのラベルを表す。

【０１３３】例えば、文書グループＤ１は、インターネット上で（株）Ｔ社が運用管理しているサイトに相当する。ページ「Ｄ１－１」「Ｄ１－２」「Ｄ１－３」は、同一の文書グループＤ１に属する。

【０１３４】一方、図１３において、リンク３０１、３０２、３０３、３０４は、異なる文書グループに属するページ間のリンクである。このようなリンクを、ここでは、文書グループ間リンクと呼ぶ。

【０１３５】なお、文書グループの定義としては、サイトやドメインといった大まかな単位を文書グループと見なすのが最も単純な方法であるが、さらに細かく文書グループを分割したり、文書グループを階層的に構成する方法も可能である。

【０１３６】このような文書グループを用いた情報検索システムの構成例を図１２に示す。なお、図１２において、図１と同一部分には同一符号を付し、異なる部分についてのみ説明する。

【０１３７】すなわち、図１のリンク検索部３、リンク情報記憶部４，ページ検索部６，ページ情報記憶部７からなる文書グループ内検索部５０を文書グループ毎に設け、文書グループ間リンクの情報を記憶するための文書グループ間リンク情報記憶部５２と、この情報を用いてユーザの検索要求に適合するページを検索する文書グループ間リンク検索部５１、が新たに追加されている。

【０１３８】文書グループ間リンク情報記憶部５２には、文書グループ間リンク情報が図３と同様にして記憶されている。ただし、この場合、１つのリンク（文書グループ間リンク）について、リンク元であるページと、リンク先であるページのそれぞれの属する文書グループは必ず異なっている。

【０１３９】文書グループ間リンク検索部５１は、索引語抽出部２で抽出された索引語とリンク情報とを比較して、適合する文書を検索する点でリンク検索部３と基本的には同様であるが、文書グループ間リンク検索部５１は、文書グループ間リンク情報記憶部５２に記憶されている文書グループ間リンクのみを処理対象とする点で異なる。

【０１４０】複数の文書グループにそれぞれ対応する複数の文書グループ内検索部５０のそれぞれは、同一の文書グループ内に存在するリンクとページを検索対象とするものである。なお、ここでは、説明の簡単のため、１つの文書グループ内検索部５０が１つの文書グループに１対１で対応しているものとするが、この場合に限らず、１つの文書グループ内検索部５０が複数の文書グループに対応していてもよいし、１つの文書グループ内検索部５０が全ての文書グループのそれぞれに対応していてもよい。ただし、１つの文書グループに１つの文書グループ内検索部５０を割り当て、並列分散して動作するように構成すれば、個々の検索部の負荷が減じ、大量のページに対しても高速に検索できるようになる。

【０１４１】図１２に示した情報検索システムでは、１つの文書グループに１つの文書グループ内検索部５０を割り当てているので、リンク情報記憶部４に記憶されているリンク情報は、検索対象である１つの同じ文書グループに属するページからページへのリンクに関するものだけである。また、ページ情報記憶部７に記憶されているページ情報も検索対象である１つの同じ文書グループに属するページに関するものだけである。

【０１４２】図１２の検索結果統合部５では、文書グループ間リンク検索部５１による検索結果と、複数の文書グループ内検索部５０のそれぞれによる検索結果とを、統合する処理を行う。

【０１４３】次に、図１２に示した情報検索システムの処理動作について、図１４に示すフローチャートを参照して説明する。

【０１４４】ユーザにより入力された検索条件ｑから検索語を抽出して、ベクトルＶｑを作成するステップＳ１１１の処理は、図８のステップＳ２１～ステップＳ２３と同様である。

【０１４５】また、図１４ののステップＳ１１２およびステップＳ１１３の処理は、図８のステップＳ２４およびＳ２５とほぼ同様であるが、ステップＳ１１３では文書グループ間リンクの構造のみを用いて適合度Ｓｉｍ（ｐ，ｑ）を求める点で異なる。ステップＳ１１２とＳ１１３の処理は、文書グループ間リンク検索部５１で行う。

【０１４６】なお、ステップＳ１１３で、検索条件ｑのベクトルＶｑと比較するベクトルＶｐｉｍは、図６を参照して説明した、ページｐのベクトルＶｐｉの作成するための処理と同様にして、文書グループ間リンク情報記憶部５２に記憶されている文書グループ間リンク情報のみに基づいて、文書グループ間リンク索引作成部１０で、各ページｐについて、あらかじめ作成したものである。ベクトルＶｐｉｍを、ここでは、ページｐのｎ段の文書グループ間リンクのベクトルと呼ぶ。

【０１４７】ここで、文書グループ間リンク索引作成部１０の処理動作について、図６を参照して、リンク索引作成部８での処理動作と異なる部分についてのみ説明する。すなわち、図６のステップＳ５では、ページｐと１段のリンクで参照関係にある、ページｐとは異なる文書グループに属する文書からのリンクのベクトルの総和、すなわち、Ｖｐｉ（１）を作成する。また、図６のステップＳ６では、ページｐがｎ段のリンクで参照されていて、このｎ段のリンクのそれぞれが異なる２つの文書グ

ループ間にまたがるリンク（このような２つのページ間をリンクするハイパーリンクを文書グループ間リンクと呼ぶ）であるとき、図６のステップＳ６では、この全ての文書グループ間リンクのベクトルの総和を求めることにより作成することができる。そして得られたＶｐｌ（ｎ）を、Ｖｐｌｅｘに置き換えればよい。

【０１４８】例えば、図１３において、ページ「Ｄ３－１」は、１段の文書グループ間リンク３０３でページ「Ｄ１－３」から参照されており、ページ「Ｄ１－３」は、１段の文書グループ間リンク３０２でページ「Ｄ２－１」から参照されているので、ページ「Ｄ３－１」は、２段の文書グループ間リンクでページ「Ｄ２－１」から参照されていることになる。このような複数段の文書グループ間リンクを用いて、そのそれぞれのラベルのベクトルｖｌから、上記したようにして、Ｖｐｌ（１）、Ｖｐｌｅｘを作成する。

【０１４９】文書グループ間のリンクのベクトルＶｐｌｅｘは、上述したように、文書グループ間リンク情報のみに基づいて作成されたものであり、言い換えれば、文書グループ間のリンクの参照関係に基づき作成されたものである。

【０１５０】ステップＳ１１３では、各ページ毎に求めた文書グループ間のリンクのベクトルＶｐｌｅｘと、検索条件のベクトルＶｑとを比較し、その類似度を求める。類似度の算出方法としては、一般に、ベクトルの内積や余弦をとる方法がよいとされている。こうして求めたＶｐｌｅｘとＶｑとの類似度を、検索条件ｑに対する文書グループ間リンク構造に基づく適合度Ｓｌｅｘ（ｐ，ｑ）とする。

【０１５１】次に、全ての文書グループＧｒについて（Ｓ１１４）、ステップＳ１１５からＳ１１７の処理を行う。

【０１５２】ステップＳ１１５では、文書グループＧｒに属するページについて、上記ステップＳ１１３で求めた適合度Ｓｌｅｘ（ｐ，ｑ）のうち、その最大値を、Ｇｒと検索条件ｑとの適合度Ｓ（Ｇｒ，ｑ）とする。このＳ（Ｇｒ，ｑ）は、検索条件ｑに対して、文書グループＧｒがどれくらい適合しているかを表す値と考えることができる。なお、同じ文書グループに属するページｐのＳｌｅｘ（ｐ，ｑ）の最大値をＳ（Ｇｒ，ｑ）とするのではなく、これらページｐのＳｌｅｘ（ｐ，ｑ）の総和や平均値などをＳ（Ｇｒ，ｑ）としてもよい。

【０１５３】Ｓ（Ｇｒ，ｑ）が予め定められた閾値Ｓｇｍｉｎより大きい文書グループＧｒについては（ステップＳ１１６）、その文書グループ内での検索を行う（ステップＳ１１７）。すなわち、ステップＳ１１７では、当該文書グループ対応の文書グループ内検索部５０のそれぞれに、図８のステップＳ２４～ステップＳ２９までの処理を行う。

【０１５４】Ｓ（Ｇｒ，ｑ）が予め定められた閾値Ｓｇｍｉｎより大きい文書グループＧｒ内における検索結果は、Ｓ（Ｇｒ，ｑ）の値が大きい文書グループの順に、文書グループ毎にまとめられて、ユーザに提示する（Ｓ１１８）。

【０１５５】次に、ユーザが検索条件として、例えば「Ｔ社のノートパソコンを修理したい」を所定の入力画面から入力して、検索の実行を図１２の情報検索システムに指示した場合の検索結果の表示例について説明する。

【０１５６】図１５～図１６は、図１２の情報検索システムのユーザインタフェース１の画面表示例を示す図である。

【０１５７】図１５において、領域３００には、先にユーザにより入力された検索条件が表示されている。

【０１５８】図１４に示したようにして検索した結果は、Ｓ（Ｇｒ，ｑ）の値が大きい文書グループから順に表示されるが、図１５では、１つの文書グループに１つの表示領域３０１を割り当てて表示している。

【０１５９】１番目の領域３０１には、Ｓ（Ｇｒ，ｑ）の値が最も大きい文書グループ内の検索結果が表示される。

【０１６０】例えば、図１５では、（株）Ｔ社のサイトが、検索条件に最も適合する文書グループとして求められる。例えば「Ｔ社」という語と「パソコン」という語を比較した場合、文書グループ間リンクだけに着目すれば、「Ｔ社」という固有名詞をラベルに含んだリンクは、少数のＵＲＬを集中して参照する傾向にある。これに対し、「パソコン」という一般名詞をラベルに含むリンクは、多数のＵＲＬを参照する傾向にあり、少数のＵＲＬに集中して参照することは稀である。このような性質が検索語の重みに影響するため、Ｔ社のサイトが、より適合度の大きい文書グループとして選ばれる。このことはユーザの検索条件に合致する。

【０１６１】図１５において、１番目の領域３０１の１番目の行には、当該文書グループＧｒに属するページのうち、図１４のステップＳ１１３で求めた適合度Ｓｌｅｘ（ｐ，ｑ）が最も大きいページの見出し（例えば、ここではタイトル）３０２が表示されている。このページは、文書グループ間リンクの構造に基づいた適合度が大きいページであるから、当該文書グループ１を代表的するページと見なすことができる。一方、２番目、３番目、４番目に表示されているページの見出し３０２，３０４，３０５は、図１４のステップＳ１１７の処理によって、当該文書グループの中で求められた検索結果のページである。

【０１６２】文書グループ内のリンク構造だけに着目すれば、「Ｔ社」という語は、当該文書グループ内では数多く使われる語であるため、文書グループ内でページを特定する働きが弱い。これと比較して、「ノート」「パソコン」「修理」という一連の語は、ページを特定する

働きが強い。このような性質は、リンクのラベルについても、ページの内容についても成り立つ。この性質が検索語の重みに影響するため、ステップＳ１１７では、文書グループ内からユーザの検索条件によく適合するページが検索できる。

【０１６３】なお、図１５の表示例では、１番目のページの見出し３０２以外の各ページの見出し３０３，３０４，３０５には、そのページの本文の要約を対応付けて表示している。これは、前述の第１の実施形態の図１０の領域２０３で要約を表示している場合と同様である。

【０１６４】また、検索結果として表示される全ての見出しは、マウス等でクリックすることにより、その見出しに対応するページを表示するようになっていることが望ましい。

【０１６５】さらに、各文書グループ内の検索は、第１の実施形態で説明したように、検索実行に先だって、検索方法を指定することにより、リンク構造に基づく検索、ページ内容に基づく検索、それらを総合した検索を行って、指定された検索方法に対応する検索結果の表示を行ってもよい。

【０１６６】ユーザにより入力された検索条件によっては、上記処理によってユーザが所望する文書グループを求めることができても、その文書グループ内でユーザがどのページを所望しているかということまでは求めることができないときがある。例えば単に「Ｔ社」という検索条件が入力された場合には、ユーザが所望するページを文書グループ内で検索するための条件がユーザから与えられていない。このような場合の処理としては、文書グループ内の検索を行わないという方法と、検索条件がなくてもユーザにとって有用と思われるページを文書グループ内から選んで提案する方法の、二通りが考えられる。

【０１６７】前者の方法では、例えば、図１４のステップＳ１１７の処理は省略し、検索結果として、文書グループに属するページのうち、図１４のステップＳ１１３で求めた適合度Ｓ１ｅｘ（ｐ，ｑ）が最も大きいページの見出し３０２のみをユーザに提示する。

【０１６８】後者の方法の一つの実施形態としては、ハイパーリンクによって他のページから参照されている数の多いページを有用なページであると見なし、そのような有用度の大きいページをユーザにいくつか提示する方法がある。すなわち、図４に示したようなリンク情報から各ページについての被リンク数が求まるので、この被リンク数の単調増加関数としてページの有用度を定義する。ステップＳ１１７では、文書グループＧｒ内の個々のページについて有用度を求め、有用度が大きいページを所定の個数（例えば上位５件）だけを選ぶという処理を行う。この場合の検索結果の表示例を図１６に示す。

【０１６９】なお、図１６において、図１５と同一部分には、同一符号を付し、異なる部分について説明する。すなわち、検索結果として得られた各文書グループの表示領域３０１では、１番目の行には、当該文書グループに属するページのうち、図１４のステップＳ１１３で求めた適合度Ｓ１ｅｘ（ｐ，ｑ）が最も大きいページの見出し（例えば、ここではタイトル）３０２を表示するものの、それ以下には、被リンク数に基づいた上述の方法で有用度が大きいと見なされたページの見出し（例えば、ここでは、タイトル）４０３～４０７が表示される。

【０１７０】図１６に示した形式で検索結果を表示することによって、ユーザは、簡単な検索要求を入力した場合にも、有用なページを容易に見つけることができる。なお、ページの有用度を被リンク数によって求める処理では、文書グループ内リンクと文書グループ間リンクとを区別し、文書グループ間リンクの方を重要視する方法が効果的である。なぜなら、異なる文書グループから多く参照されているページは、その価値がより客観的に支持されているページであると考えられるからである。

【０１７１】以上説明したように、文書グループを考慮した上記第２の実施形態によれば、文書グループ間リンクと、文書グループ内のリンクとの、性質の違いに着目することにより、ユーザの検索条件によく適合するページを検索できるのみならず、得られた検索結果が、文書グループ毎にまとめた形で提示されるため、ユーザは、検索結果から所望のページを簡単に見つけることができる。さらに、文書グループ内の検索処理を、複数の文書グループ内検索部５０に分散し、並列して実行することにより、大量の文書についても高速に検索できるという利点がある。

【０１７２】上記第１～第２の実施形態によれえば、大量のハイパーテキスト形式の文書の中から、ユーザが自然文または複数の語で記述した複雑な検索条件に適合する文書を、効率よく検索することができる。また、ハイパーリンクによる参照関係や文書グループに基づいて検索結果を整理して提示することにより、ユーザは、所望する情報を検索結果の中から容易に見つけ出すことができる。

【０１７３】なお、上記第１～第２の実施形態において、検索結果として、ページの見出しを表示する際には、その全ての見出しは、マウス等によりクリックされることにより、その見出しに対応するページが表示されるようになっていることが望ましい。

【０１７４】また、本発明の実施の形態に記載した本発明の手法は、コンピュータに実行させることのできるプログラムとして、磁気ディスク（フロッピー（登録商標）ディスク、ハードディスクなど）、光ディスク（ＣＤ－ＲＯＭ、ＤＶＤなど）、半導体メモリなどの記録媒体に格納して頒布することもできる。

【０１７５】さらに、本発明は、上記実施形態に限定されるものではなく、実施段階ではその要旨を逸脱しない

範囲で種々に変形することが可能である。さらに、上記実施形態には種々の段階の発明は含まれており、開示される複数の構成用件における適宜な組み合わせにより、種々の発明が抽出され得る。例えば、実施形態に示される全構成要件から幾つかの構成要件が削除されても、発明が解決しようとする課題の欄で述べた課題（の少なくとも１つ）が解決でき、発明の効果の欄で述べられている効果（のなくとも１つ）が得られる場合には、この構成要件が削除された構成が発明として抽出され得る。

【０１７６】

【発明の効果】以上説明した様に本発明によれば、大量のハイパーテキスト形式の文書の中から、複数の語からなる複雑な検索条件に適合する文書の検索が容易に行えるとともに、高い精度の検索結果が得られる。

【図面の簡単な説明】

【図１】本発明の第１の実施形態に係る情報検索システムの構成例を示した図。

【図２】検索対象の複数のページのそれぞれに与えられた識別子と、各識別子に対応する各ページのＵＲＬの記憶例を示した図。

【図３】リンク情報記憶部におけるリンク情報の記憶例を示した図。

【図４】ページ情報記憶部におけるページ情報の記憶例を示した図。

【図５】検索対象の複数のページのハイパーリンクによる参照関係の一例を示した図。

【図６】各ページについて、ｎ段のリンクのベクトルを作成するための処理を説明するためのフローチャート。

【図７】各ページについて、ページ内容のベクトルを作成するための処理を説明するためのフローチャート。

【図８】図１の情報検索システムの検索処理動作を説明するためのフローチャート。

【図９】図１の情報検索システムのユーザインタフェースの画面表示例を示す図で、検索要求を入力する入力画面の一例を示した図。

【図１０】図１の情報検索システムのユーザインタフェースの画面表示例を示す図で、検索結果の表示例を示した図。

【図１１】図１の情報検索システムのユーザインタフェースの画面表示例を示す図で、検索結果の表示例を示した図。

【図１２】本発明の第２の実施形態に係る情報検索システムの構成例を示した図。

【図１３】検索対象の複数のページと文書グループのハイパーリンクによる参照関係の一例を示した図。

【図１４】図１２の情報検索システムの検索処理動作を説明するためのフローチャート。

【図１５】図１２の情報検索システムのユーザインタフェースの画面表示例を示す図で、検索結果の表示例を示した図。

【図１６】図１２の情報検索システムのユーザインタフェースの画面表示例を示す図で、検索結果の他の表示例を示した図。

【符号の説明】

１…ユーザインタフェース
２…検索語抽出部
３…リンク検索部
４…リンク情報記憶部
５…検索結果統合部
６…ページ検索部
７…ページ情報記憶部
８…リンク索引作成部
９…ページ索引作成部
１０…文書グループ間リンク索引作成部
１１…ウェブ情報収集部
５０…文書グループ内検索部
５１…文書グループ間リンク検索部
５２…文書グループ間リンク情報記憶部

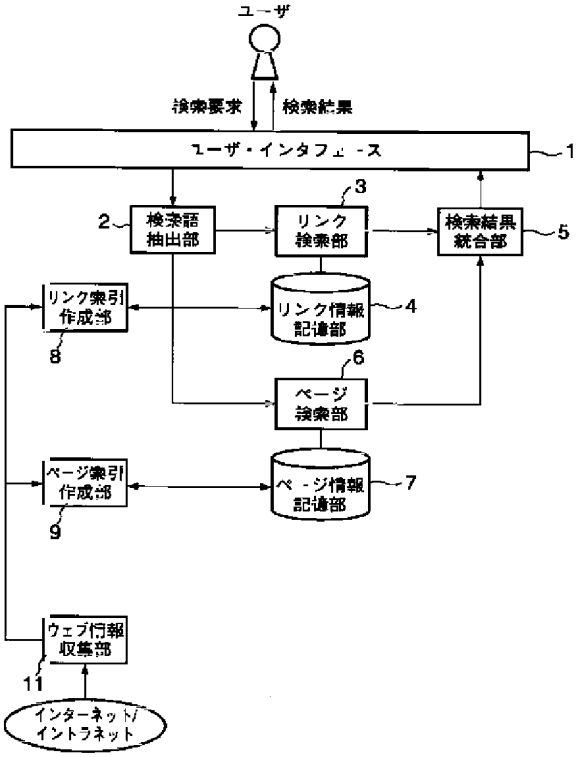
【図２】

ID	URL
1-1	http://www.foo.co.jp/
1-2	http://www.foo.co.jp/products/
1-3	http://www.foo.co.jp/products/pc.html
2-1	http://www.bar.com/
2-2	http://www.bar.com/about/

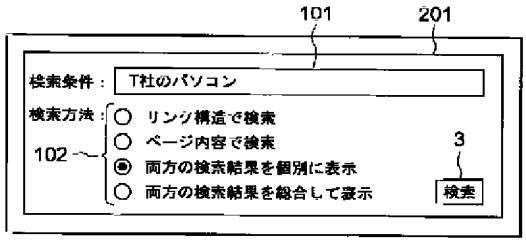
【図３】

リンク元ID	リンク先ID	ラベル
8-16	5-1	株式会社Ｔ社
27-368	5-1	㈱Ｔ社のホームページ
32-59	5-1	Ｔ社へのリンク
5 4	5-1	Ｔ社トップ

【図１】



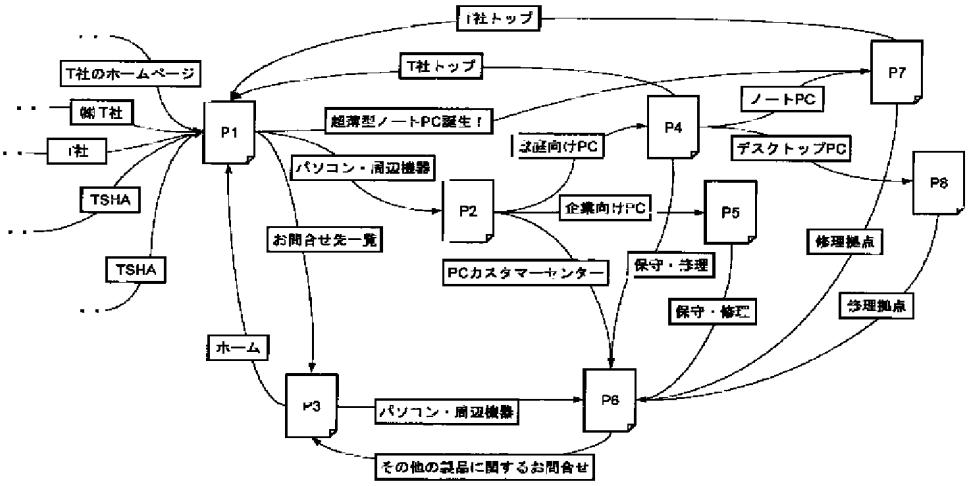
【図９】



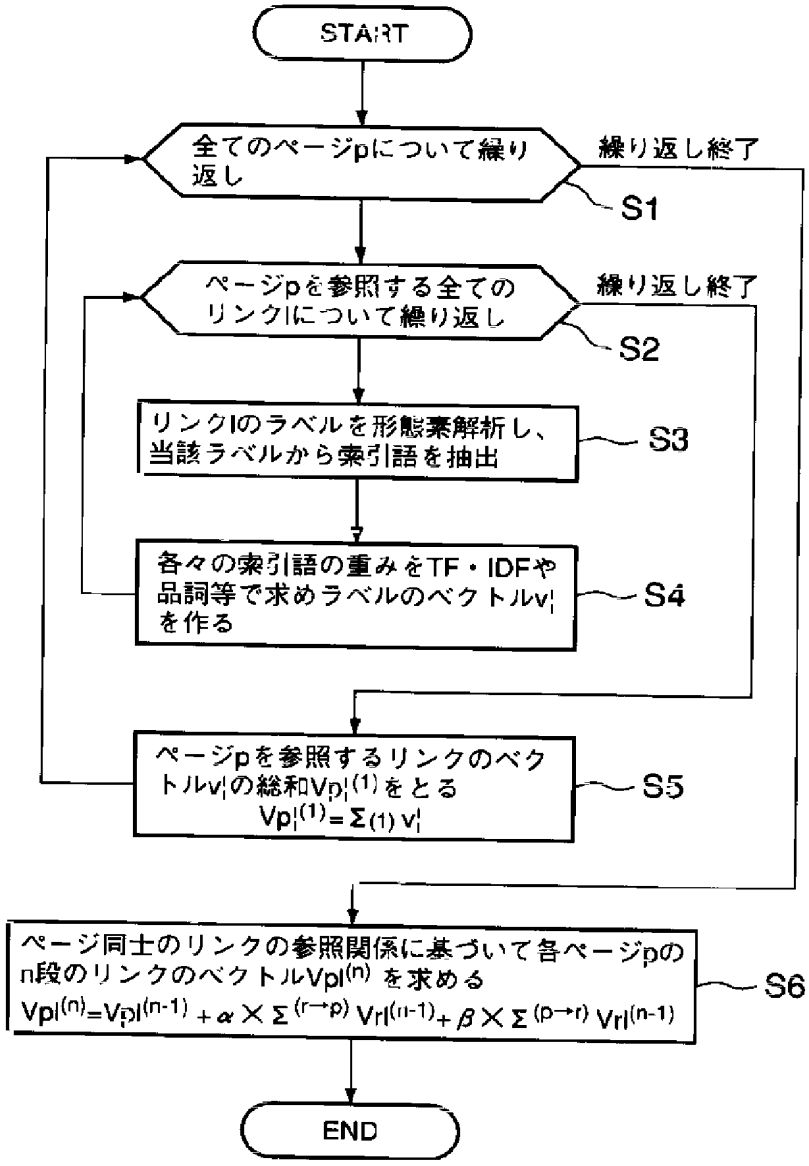
【図４】

ID	タイトル	本　文
5-1	(株)T社	T社のウェブサイトへようこそ！ 製品紹介 企業情報 IR情報 採用情報 お問合せ・・
5-2	T社PCウェブ	T社PCウェブはT社製パソコン製品の総合サイトです。更新:1 2001年4月7日最新情報・・
5-3	T社PCカスタマーセンター	T社PCカスタマーセンター無償修理 アップデート 修理拠点 お知らせ・・

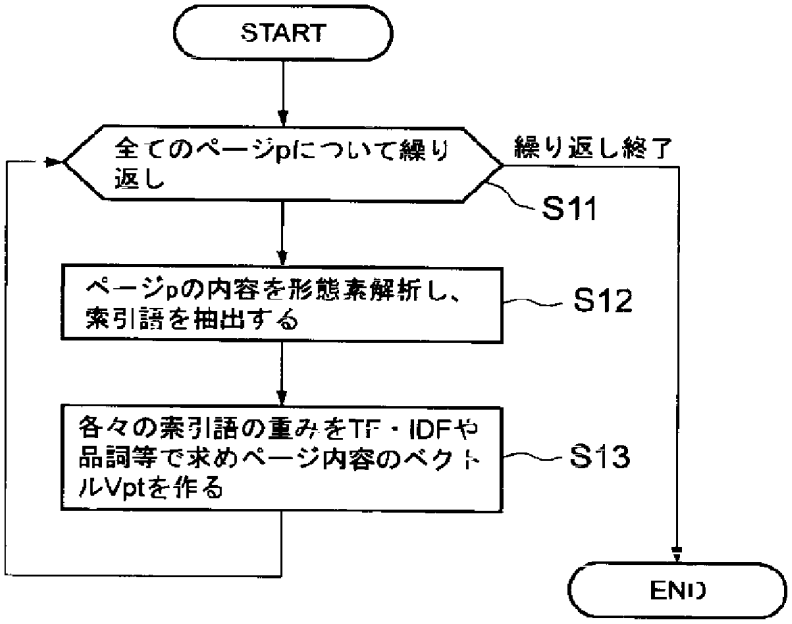
【図５】



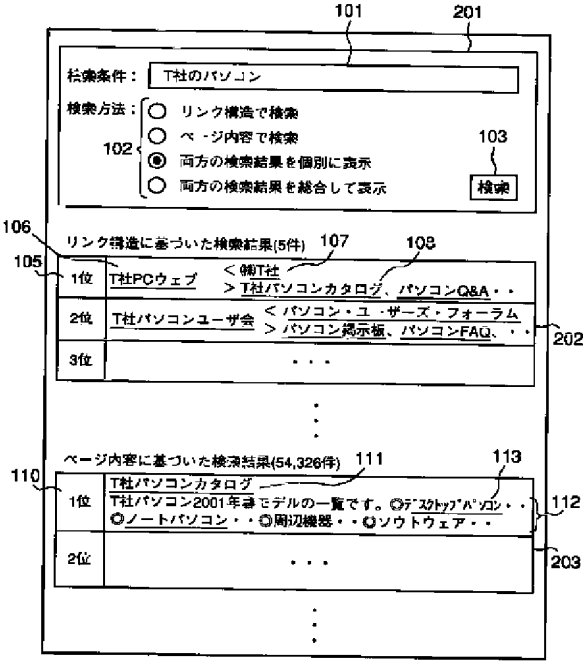
【図6】


START
全てのページpについて繰り返し
繰り返し終了
S1
ページpを参照する全てのリンクについて繰り返し
繰り返し終了
S2
リンクlのラベルを形態素解析し、当該ラベルから索引語を抽出
S3
各々の索引語の重みをTF・IDFや品詞等で求めラベルのベクトルvlを作る
S4
ページpを参照するリンクのベクトルvlの総和Vpl(1)をとる
Vpl(1)=Σ(1) vl
S5
ページ同士のリンクの参照関係に基づいて各ページpのn段のリンクのベクトルVpl(n) を求める
Vpl(n)=Vpl(n-1) + α×Σ(r→p) Vrl(n-1)+ β×Σ(p→r) Vrl(n-1)
S6
END

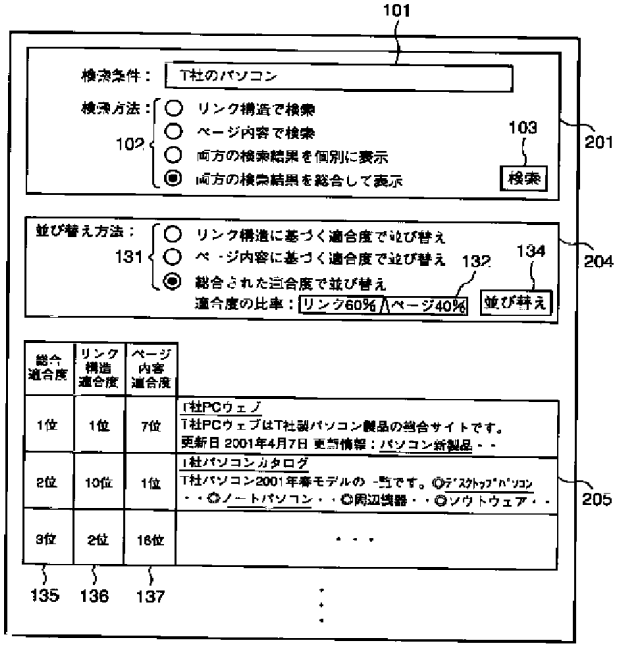
【図7】


START
全てのページpについて繰り返し
繰り返し終了
S11
ページpの内容を形態素解析し、索引語を抽出する
S12
各々の索引語の重みをTF・IDFや品詞等で求めページ内容のベクトルVptを作る
S13
END

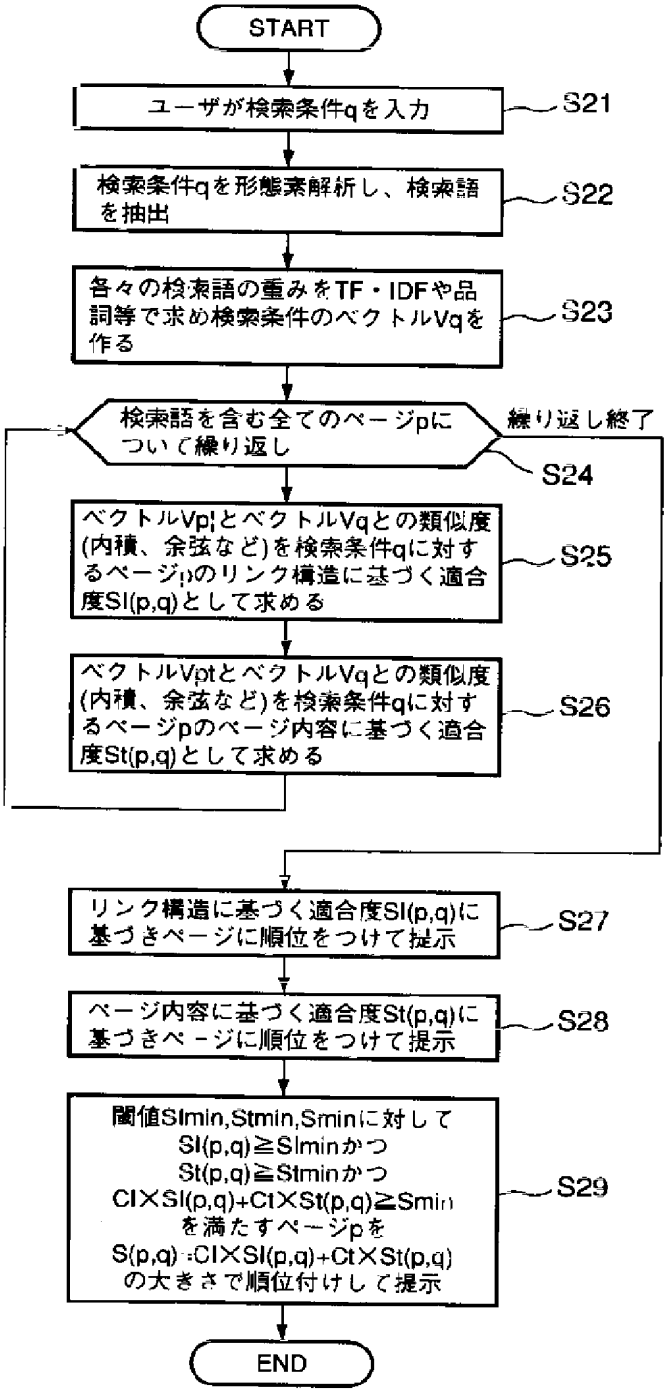
【図10】


101
201
検索条件：
T社のパソコン
検索方法：
リンク構造で検索
ページ内容で検索
両方の検索結果を個別に表示
両方の検索結果を総合して表示
102
103
検索
リンク構造に基づいた検索結果(5件)
106
105
107
108
1位
T社PCウェブ
＜例T社
＞T社パソコンカタログ、パソコンQ&A・・
2位
T社パソコンユーザ会
＜パソコン・ユーザーズ・フォーラム
＞パソコン掲示板、パソコンFAQ、・・
3位
・・・
202
ページ内容に基づいた検索結果(54,326件)
110
111
113
1位
T社パソコンカタログ
T社パソコン2001年春モデルの一覧です。◎ﾃﾞｽｸﾄｯﾌﾟﾊﾟｿｺﾝ・・◎ノートパソコン・・◎周辺機器・・◎ソウトウェア・・
112
2位
・・・
203

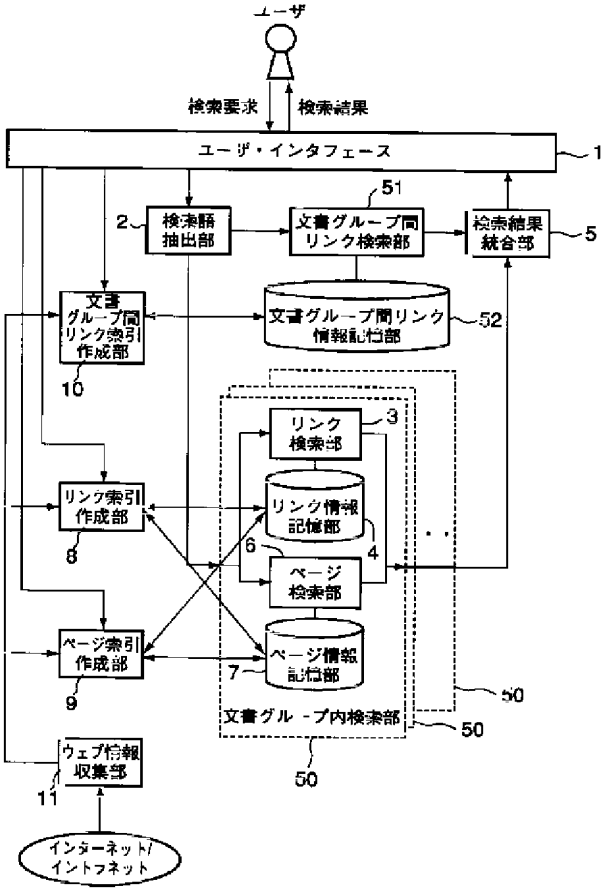
【図11】


101
検索条件：
T社のパソコン
検索方法：
リンク構造で検索
ページ内容で検索
両方の検索結果を個別に表示
両方の検索結果を総合して表示
102
103
検索
201
並び替え方法：
リンク構造に基づく適合度で並び替え
ページ内容に基づく適合度で並び替え
総合された適合度で並び替え
131
132
134
適合度の比率：リンク60%/ページ40%
並び替え
204
総合適合度
リンク構造適合度
ページ内容適合度
1位
1位
7位
T社PCウェブ
T社PCウェブはT社製パソコン製品の総合サイトです。
更新日 2001年4月7日 更新情報：パソコン新製品・・
2位
10位
1位
T社パソコンカタログ
T社パソコン2001年春モデルの一覧です。◎ﾃﾞｽｸﾄｯﾌﾟﾊﾟｿｺﾝ・・◎ノートパソコン・・◎周辺機器・・◎ソウトウェア・・
205
3位
2位
16位
・・・
135
136
137

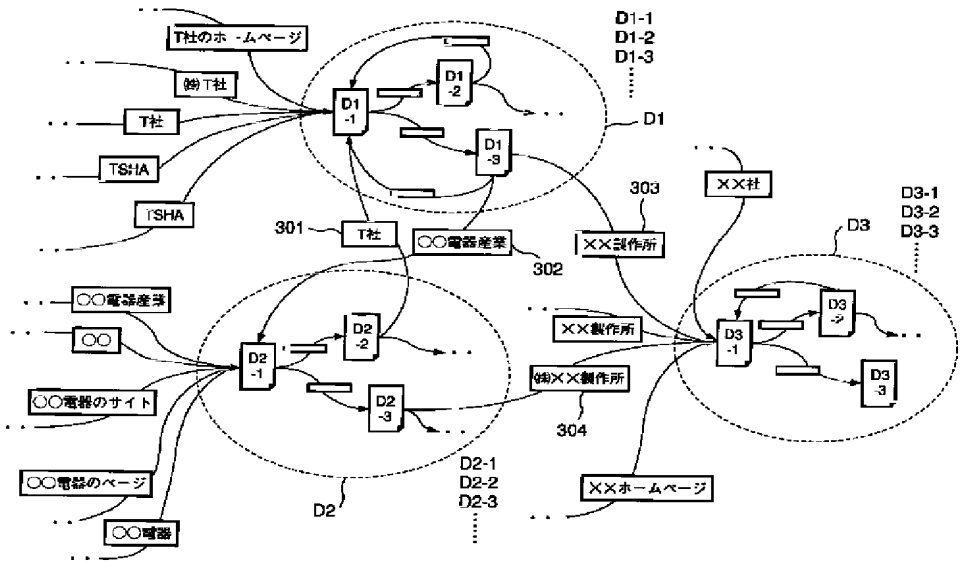
【図8】


START
ユーザが検索条件qを入力
S21
検索条件qを形態素解析し、検索語を抽出
S22
各々の検索語の重みをTF・IDFや品詞等で求め検索条件のベクトルVqを作る
S23
検索語を含む全てのページpについて繰り返し
繰り返し終了
S24
ベクトルVpとベクトルVqとの類似度(内積、余弦など)を検索条件qに対するページpのリンク構造に基づく適合度SI(p,q)として求める
S25
ベクトルVptとベクトルVqとの類似度(内積、余弦など)を検索条件qに対するページpのページ内容に基づく適合度St(p,q)として求める
S26
リンク構造に基づく適合度SI(p,q)に基づきページに順位をつけて提示
S27
ページ内容に基づく適合度St(p,q)に基づきページに順位をつけて提示
S28
閾値SImin,Stmin,Sminに対して
SI(p,q)≧SIminかつ
St(p,q)≧Stminかつ
CI×SI(p,q)+Ct×St(p,q)≧Smin
を満たすページpを
S(p,q)=CI×SI(p,q)+Ct×St(p,q)
の大きさで順位付けして提示
S29
END

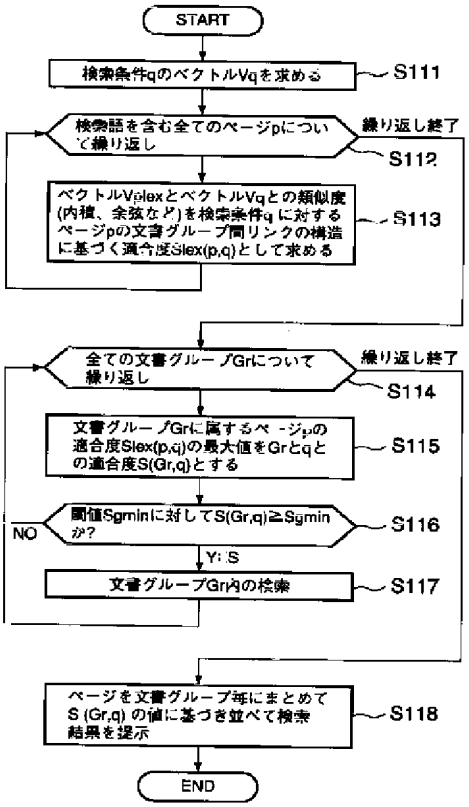
【図12】


ユーザ
検索要求
検索結果
ユーザ・インタフェース
1
検索語抽出部
2
文書グループ間リンク検索部
51
検索結果統合部
5
文書グループ間リンク索引作成部
10
文書グループ間リンク情報記憶部
52
リンク検索部
3
リンク索引作成部
8
リンク情報記憶部
4
6
ページ検索部
ページ索引作成部
9
ページ情報記憶部
7
50
文書グループ内検索部
50
50
ウェブ情報収集部
11
インターネット/イントラネット

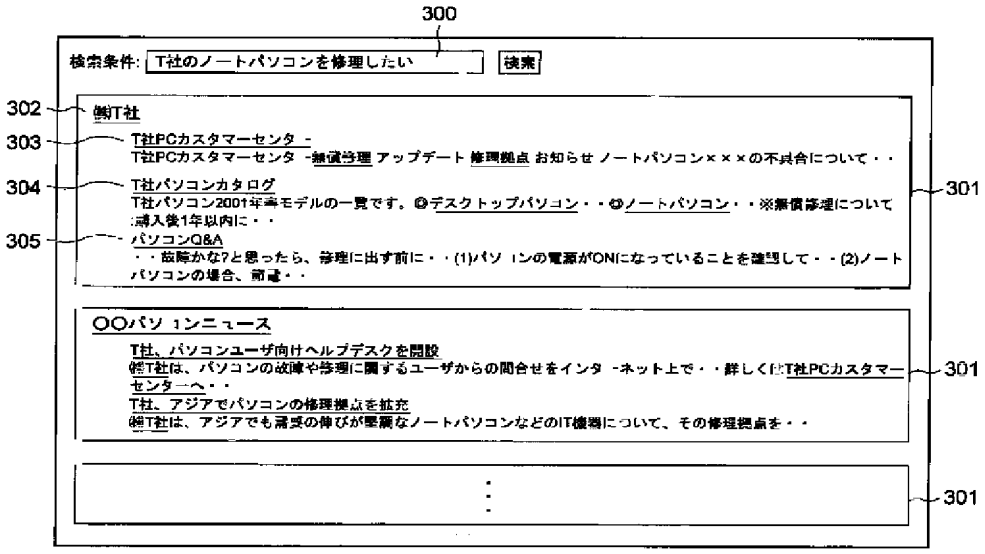
【図１３】


T社のホームページ
(株)T社
T社
TSHA
TSHA
D1-1
D1-2
D1-3
D1
D1-1
D1-2
D1-3
301
T社
○○電器産業
302
303
××製作所
××社
D3
D3-1
D3-2
D3-3
○○電器産業
○○
○○電器のサイト
○○電器のページ
○○電器
D2-1
D2-2
D2-3
D2
××製作所
(株)××製作所
304
××ホームページ

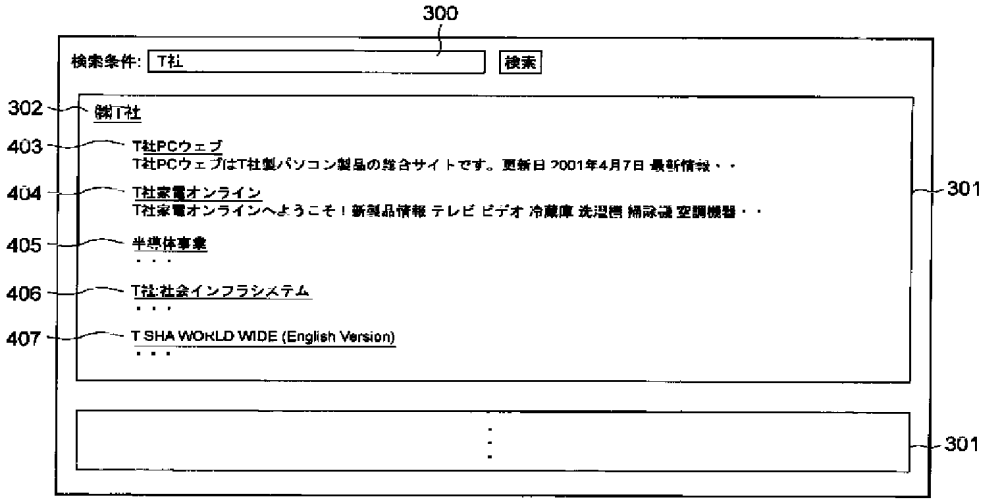
【図１４】


START
検索条件qのベクトルVqを求める
S111
検索語を含む全てのページpについて繰り返し
繰り返し終了
S112
ベクトルVplexとベクトルVqとの類似度(内積、余弦など)を検索条件qに対するページpの文書グループ間リンクの構造に基づく適合度Slex(p,q)として求める
S113
全ての文書グループGrについて繰り返し
繰り返し終了
S114
文書グループGrに属するページpの適合度Slex(p,q)の最大値をGrとqとの適合度S(Gr,q)とする
S115
閾値Sgminに対してS(Gr,q)≧Sgminか?
NO
S116
Y: S
文書グループGr内の検索
S117
ページを文書グループ毎にまとめてS(Gr,q)の値に基づき並べて検索結果を提示
S118
END

【図１５】


300
検索条件: T社のノートパソコンを修理したい
検索
302
㈱T社
303
T社PCカスタマーセンタ -
T社PCカスタマーセンタ -無償修理 アップデート 修理拠点 お知らせ ノートパソコン×××の不具合について・・
304
T社パソコンカタログ
T社パソコン2001年春モデルの一覧です。◎デスクトップパソコン・・◎ノートパソコン・・※無償修理について :購入後1年以内に・・
305
パソコンQ&A
・・故障かな?と思ったら、修理に出す前に・・(1)パソコンの電源がONになっていることを確認して・・(2)ノートパソコンの場合、前述・・
301
○○パソコンニュース
T社、パソコンユーザ向けヘルプデスクを開設
㈱T社は、パソコンの故障や修理に関するユーザからの問合せをインターネット上で・・詳しくはT社PCカスタマーセンターへ・・
T社、アジアでパソコンの修理拠点を拡充
㈱T社は、アジアでも需要の伸びが著しいノートパソコンなどのIT機器について、その修理拠点を・・
301
301

【図１６】


300
検索条件: T社
検索
302
㈱T社
403
T社PCウェブ
T社PCウェブはT社製パソコン製品の総合サイトです。更新日 2001年4月7日 最新情報・・
404
T社家電オンライン
T社家電オンラインへようこそ！新製品情報 テレビ ビデオ 冷蔵庫 洗濯機 掃除機 空調機器・・
405
半導体事業
・・・
406
T社:社会インフラシステム
・・・
407
T SHA WORLD WIDE (English Version)
・・・
301
301

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-173280

(43)Date of publication of application : 20.06.2003

(51)Int.Cl. G06F 12/00
G06F 17/30

(21)Application number : 2001-371636

(22)Date of filing : 05.12.2001

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

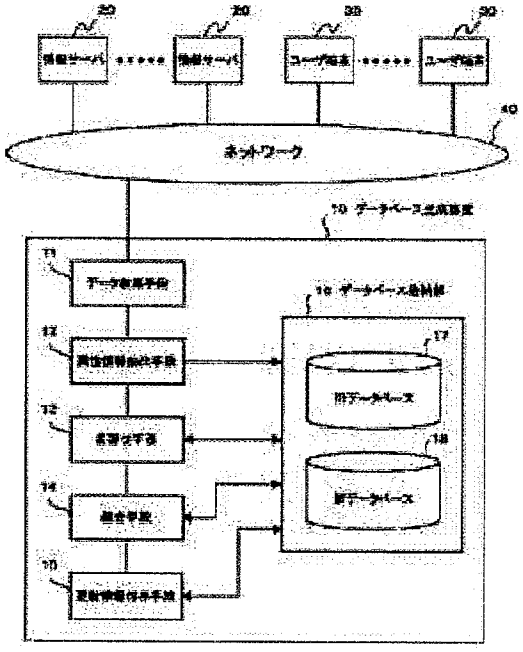
(72)Inventor : BESSHO KATSUTO
IWASE SHIGETO

(54) APPARATUS, METHOD AND PROGRAM FOR GENERATING DATABASE

(57)Abstract:

PROBLEM TO BE SOLVED: To generate a database (DB) enabling updating of data to be displayed without redundant data by collecting data from a plurality of servers which independently manage and operate store information and the like on a network.

SOLUTION: This database generating apparatus 10 comprises a means 11 for collecting data, the ID of the data, and information about updating date and the like from each server 20 on the network 40, a means 12 for extracting from the collected data an attribute value which characterizes the data to create a new DB 18 made up of the attribute value, ID, updating date and the like, a means 13 for collating the names of the data whose attribute values in the new DB can be considered the same, a means 14 for associating with one another the data whose attribute values in the new DB and the old DB 17 previously generated can be considered the same, and a means 15 for comparing the IDs and updating dates of the data in the new and old DBs and imparting updating information to the data in the new DB.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention is distributed on networks, such as the Internet, from two or more servers which are managing and managing the notice information of a store etc., etc. independently, collects data and relates to the device which generates the database for searching and showing around, a method, and its program.
[0002]
[Description of the Prior Art]The data of the notice information of a store, etc. is independently created in some organizations, and is updated if needed in many cases. Since the data set which one organization owns does not cover all the notice information of stores, if these data sets created and updated independently are unified, the more substantial information search service can be performed. The data set which each organization holds is kept in the computer connected to networks, such as the Internet, and an inspection is presented with it. Henceforth, such a computer will be called an information server.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL FIELD

[Field of the Invention]This invention is distributed on networks, such as the Internet, from two or more servers which are managing and managing the notice information of a store etc., etc. independently, collects data and relates to the device which generates the database for searching and showing around, a method, and its program.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art]The data of the notice information of a store, etc. is independently created in some organizations, and is updated if needed in many cases. Since the data set which one organization owns does not cover all the notice information of stores, if these data sets created and updated independently are unified, the more substantial information search service can be performed. The data set which each organization holds is kept in the computer connected to networks, such as the Internet, and an inspection is presented with it. Henceforth, such a computer will be called an information server.
[0003]Data sets were collected from two or more information servers, and what merged simply the data sets collected from two or more information servers was used as the database in the Prior art which generates a database. The link information to the source data in the information server with which this data exists is given to each data in the generated database.
When a user searches data from a database using a terminal, the source data of this data can be accessed by the link information which accompanies the data displayed on the terminal. Drawing 11 shows an example of the conventional search-results display screen when a type of industry searches from a database the store whose address is "Kagurazaka, Shinjuku-ku" "Chinese", for example. When a user clicks link information on a screen, the detailed screen of the store of a link destination is displayed.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

EFFECT OF THE INVENTION

[Effect of the Invention]As explained above, according to this invention, when the data which agrees in a user's demand from the generated database is searched, it becomes possible to display search results in the form where there is no duplicate data and the update information of data was added.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]In the data set independently created in some ****, it is registered with the form that a name differs from an address also at the same store, and expression, in many cases. Therefore, in the Prior art which merges simply the data sets collected from two or more information servers, and generates a database, the overlapping same store cannot be summarized to one, but into the store group of search results, two or more same stores are intermingled and may be displayed. In such a case, search results increase redundantly, and a user examines to the contents of unnecessary data and has forced the complicated work of judging whether it being the same as the data which it already looked at. For example, in the search-results display screen of drawing 11, the 1st store and the 4th store are the same, and the 3rd store and the 6th store are the same.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

MEANS

[Means for Solving the Problem]A memory measure which memorizes a database (the old database) with which a database generating device of this invention was generated in the past, A data collection means which collects from a point two or more information, including discernment ID of data and this data, an update date, etc., containing a value of the attribute of a name, an address, etc., An attribute information extracting means which generates a database (new database) of composition of that extract a value of an attribute from said each collected data, and each data consists of a value of said extracted attribute, discernment ID, an update date, etc., A data set which can be regarded as a value of an attribute in said generated new database being the same between an identification-of-multiple-accounts-under-the-same-name-as-a-single-entity means to classify into the same group, and a new database and said old database, A coupling means which judges the data which can be regarded as a value of an attribute being the same to be the same, and matches each data between both databases, By comparing information, including discernment ID of data in said new database, an update date, etc., with information, including discernment ID of data in said old database matched with said data, an update date, etc., It has an update information grant means to give update information to corresponding data in said new database.
[0008]In an identification-of-multiple-accounts-under-the-same-name-as-a-single-entity means, overlapping data is set to one in the generated new database. For this reason, when data corresponding to a user's demand is searched and displayed from this database, two or more data of the same store is not displayed, and grasp of search results can carry out more easily. Between the generated new database and the old database generated last time, specify a coupling means and data of the same store etc. in an update information grant means. Since update information of data is derived by comparing those discernment ID (for example, name), update dates, etc., the database generated eventually can display data, after update information of data is given.

[0009]Next, a data collection process in which a database generation method of this invention collects information, including discernment ID of data and this data, an update date, etc., containing a value of the attribute of a name, an address, etc. from two or more points, An attribution information extraction process which generates a database (new database) of composition of that extract a value of an attribute from said each collected data, and each data consists of a value of said extracted attribute, discernment ID, an update date, etc., A data set which can be regarded as a value of an attribute in said generated new database being the same between an identification-of-multiple-accounts-under-the-same-name-as-a-single-entity process classified into the same group, and the old database currently generated and held in a new database and the past, A joint process in which judge the data which can be regarded as a value of an attribute being the same to be the same, and each data between both databases is matched, By comparing information, including discernment ID of data in said new database, an update date, etc., with information, including discernment ID of data in said old database matched with said data, an update date, etc., It has an update information grant process in which update information is given to corresponding data in a new database.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a figure showing the composition of the database generating device of one embodiment of this invention.
[Drawing 2]It is a flow chart figure of the database generation method of one embodiment of this invention.
[Drawing 3]It is a figure showing an example of the data downloaded from the information server.
[Drawing 4]It is a figure showing an example of the new database generated by the attribute information extracting means.
[Drawing 5]It is a figure showing an example of the new database updated by the identification-of-multiple-accounts-under-the-same-name-as-a-single-entity means.
[Drawing 6]It is a figure showing an example of the collation rule applied by an identification-of-multiple-accounts-under-the-same-name-as-a-single-entity means.
[Drawing 7]It is a figure showing an example of the collation rule applied by a coupling means.
[Drawing 8]It is a figure showing an example of the old database generated last time.
[Drawing 9]It is a figure showing an example of the new database with which update information was given by an update information grant means.
[Drawing 10]It is a figure showing an example of the display screen of the search results from the database generated by this invention.
[Drawing 11]It is a figure showing the example of the display screen of the search results from the database generated by the conventional database generation art.
[Description of Notations]
10 Database generating device
11 Data collection means
12 Attribute information extracting means

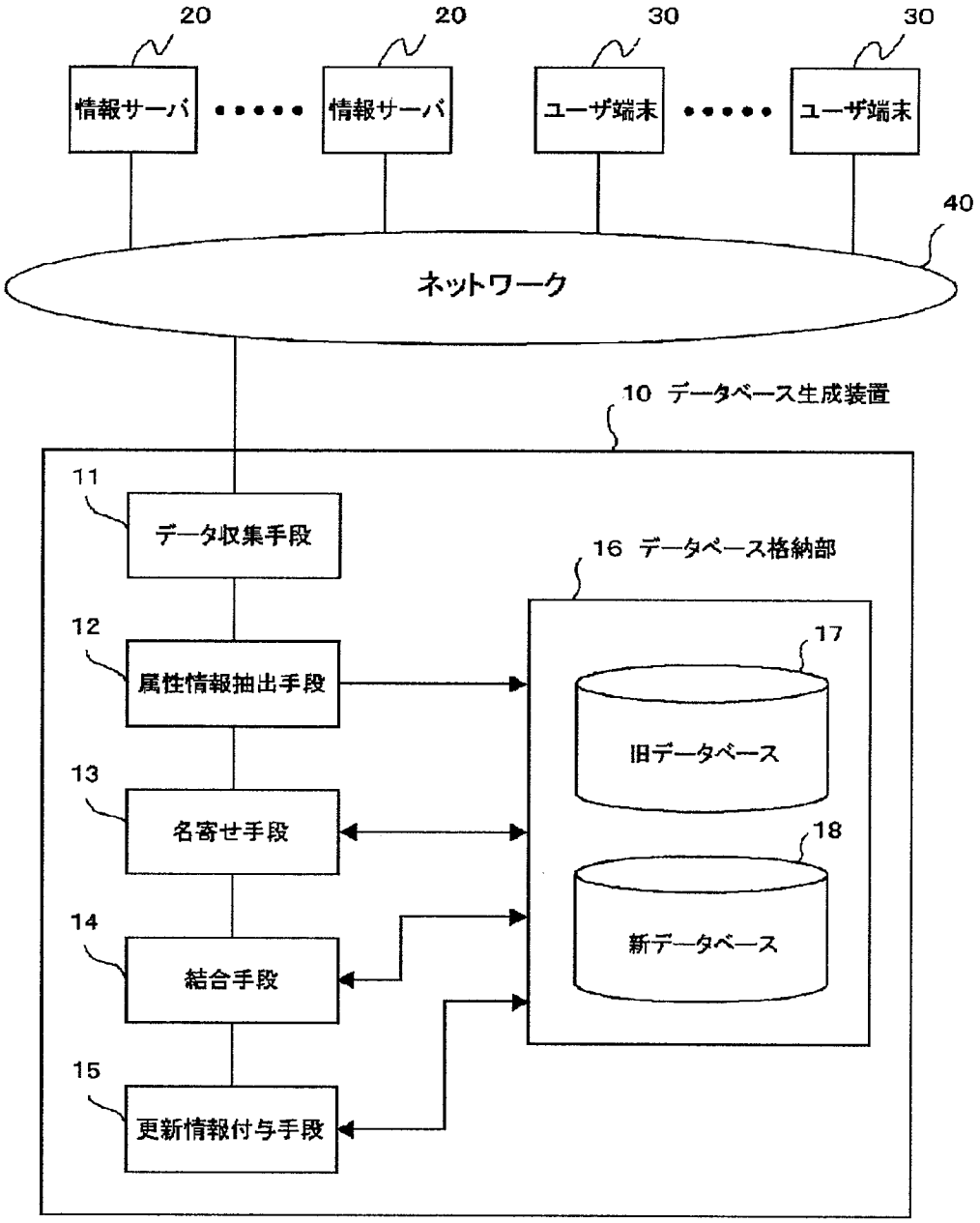
13 Identification-of-multiple-accounts-under-the-same-name-as-a-single-entity means
14 Coupling means
15 Update information grant means
16 Database storage
17 The old database
18 New database
20 Information server
30 User terminal
40 Network

* NOTICES *

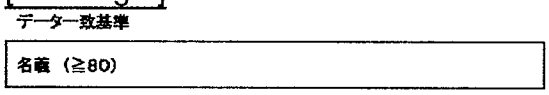
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

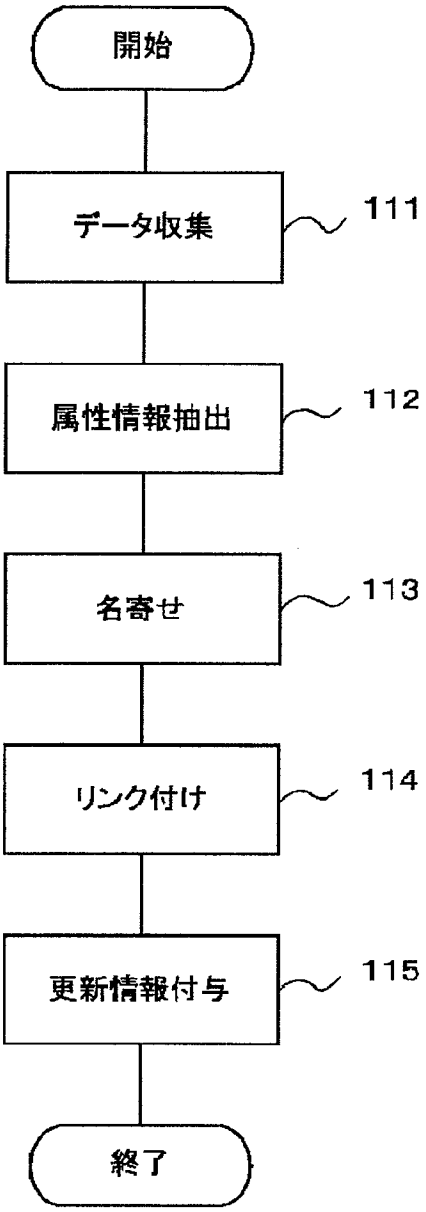
[Drawing 1]



[Drawing 7]



[Drawing 2]



[Drawing 3]

情報サーバAの一つのデータ

紅蘭亭 住所　:新宿区神楽坂1−2−3 電話番号:XXX−YYY 毎月20日は感謝デー ……	
リンク情報	更新日時
http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/10 15:30

情報サーバBの一つのデータ

ラーメンショップ紅蘭亭 住所　:新宿区神楽坂1丁目2番地3号 最寄駅　:新宿 ランチ 500円 日曜定休 ……	
リンク情報	更新日時
http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45

[Drawing 5]

業種	名義	住所	情報サーバ	リンク情報	更新日時	更新情報
中華	紅蘭亭	新宿区神楽坂 1-2-3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/11 15:30	
			B	http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45	
中華	大森飯店	新宿区神楽坂 5-2-4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14	
中華	大竹亭	新宿区神楽坂 3-8-6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19	
			B	http://www.bserv.co.jp/data/data5568.html	2001/08/11 14:10	
中華	龍王飯店	新宿区神楽坂 2-7-4	B	http://www.bserv.co.jp/data/data5238.html	2001/08/11 10:27	

[Drawing 4]

	業種	名義	住所	情報サーバ	リンク情報	更新日時
(a)	ラーメン	紅蘭亭	新宿区神楽坂 1-2-3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/11 15:30
(a)	台湾料理	大森飯店	新宿区神楽坂 5-2-4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14
(a)	中華料理	大竹亭	新宿区神楽坂 3-8-6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19
(b)	中華全般	ラーメンショップ紅蘭亭	新宿区神楽坂1丁目2番地3号	B	http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45
(b)	四川料理	龍王飯店	新宿区神楽坂2丁目7番地4号	B	http://www.bserv.co.jp/data/data5238.html	2001/08/11 10:27
(b)	広東料理	大竹亭本店	新宿区神楽坂3丁目8番地6号	B	http://www.bserv.co.jp/data/data5568.html	2001/08/11 14:10

[Drawing 6]

(a) データー致基準

名義（≧90）＋ 住所（≧80）
名義（≧80）＋ 住所（≧90）

(b) 名義の照合方法

照合方法	評価値
完全一致	100
文字単位一致	一致した文字の数の割合 × 100
単語単位一致	一致した単語の数の割合 × 100

(c) 住所の照合方法

照合方法	評価値
完全一致	100
単語単位一致	一致した文字の数の割合 × 100

[Drawing 8]

業種	名義	住所	情報サーバ	リンク情報	更新日時	更新情報
中華	紅蘭亭	新宿区神楽坂 1-2-3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/10 16:23	なし
			B	http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45	なし
中華	大森飯店	新宿区神楽坂 5-2-4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14	なし
中華	大竹亭	新宿区神楽坂 3-8-6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19	なし

[Drawing 9]

業種	名義	住所	情報サーバ	リンク情報	更新日時	更新情報
中華	紅蘭亭	新宿区神楽坂 1-2-3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/11 15:30	更新
			B	http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45	なし
中華	大森飯店	新宿区神楽坂 5-2-4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14	更新
中華	大竹亭	新宿区神楽坂 3-8-6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19	なし
			B	http://www.bserv.co.jp/data/data5568.html	2001/08/11 14:10	新規
中華	龍王飯店	新宿区神楽坂 2-7-4	B	http://www.bserv.co.jp/data/data5238.html	2001/08/11 10:27	新規

[Drawing 10]

業種	名義	住所	リンク情報
中華	紅蘭亭	新宿区神楽坂 1-2-3	情報サーバA 更新 / 情報サーバB
中華	大森飯店	新宿区神楽坂 5-2-4	情報サーバA 更新
中華	大竹亭	新宿区神楽坂 3-8-6	情報サーバA / 情報サーバB 新規
中華	龍王飯店	新宿区神楽坂 2-7-4	情報サーバB 新規

[Drawing 11]

業種	名義	住所	リンク情報
中華	紅蘭亭	新宿区神楽坂 1-2-3	情報サーバA
中華	大森飯店	新宿区神楽坂 5-2-4	情報サーバA
中華	大竹亭	新宿区神楽坂 3-8-6	情報サーバA
中華	ラーメンショップ紅蘭亭	新宿区神楽坂1丁目2番地3号	情報サーバB
中華	龍王飯店	新宿区神楽坂2丁目7番地4号	情報サーバB
中華	大竹亭本店	新宿区神楽坂3丁目8番地6号	情報サーバB

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－173280
（P2003－173280A）
(43)公開日　平成15年６月20日(2003.6.20)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　12/00	５１２	Ｇ０６Ｆ　12/00	５１２	５Ｂ０７５
17/30	１１０	17/30	１１０Ｆ	５Ｂ０８２
	２４０		２４０Ａ	

審査請求　未請求　請求項の数５　ＯＬ　（全 11 頁）

(21)出願番号　特願2001－371636(P2001－371636)

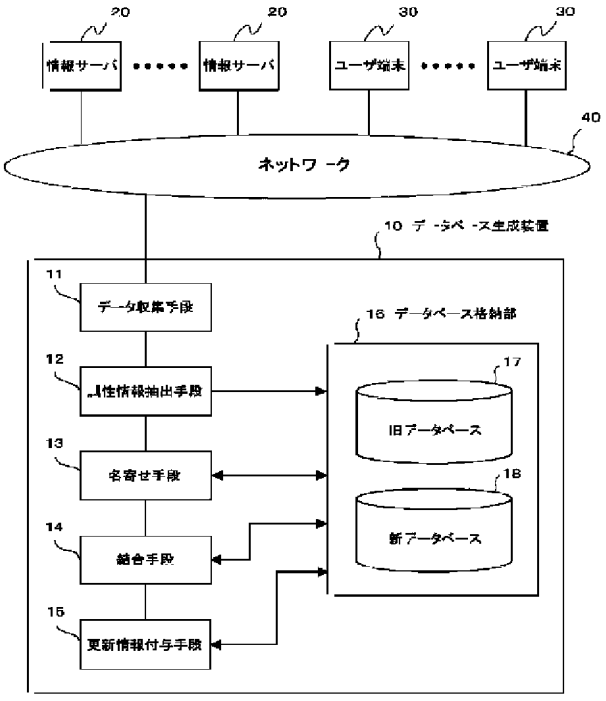
(22)出願日　平成13年12月５日(2001.12.5)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　別所　克人
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　岩瀬　成人
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(74)代理人　100073760
弁理士　鈴木　誠

最終頁に続く

(54)【発明の名称】　データベース生成装置、データベース生成方法及びデータベース生成プログラム

(57)【要約】
【課題】　ネットワーク上で店舗情報などを独立に管理・運営している複数サーバからデータを収集し、重複データがなく、かつ、データの更新状況の表示が可能なデータベース（ＤＢ）を生成する。
【解決手段】　本データベース生成装置１０は、ネットワーク４０上の各サーバ２０からデータと、該データの識別ＩＤ、更新日時などの情報を収集する手段１１、収集した各データから、該データを特徴付ける属性値を抽出し、該属性値、識別ＩＤ、更新日時などからなる新ＤＢ１８を生成する手段１２、新ＤＢ内の属性値が同一とみなせるデータを名寄せする手段１３、該新ＤＢと前回生成した旧ＤＢ１７間で属性値が同一とみなせるデータ同士を対応づける手段１４、新ＤＢのデータと対応する旧ＤＢのデータの識別ＩＤや更新日時を比較し、新ＤＢの該当データに更新情報を付与する手段１５を有する。



【特許請求の範囲】
【請求項１】　複数地点からデータを収集してデータベースを生成する装置であって、
過去に生成されたデータベース（以下、旧データベース）を記憶する記憶手段と、
各前記地点から、名義や住所などの属性の値を含むデータ、該データの識別ＩＤ、更新日時情報を収集するデータ収集手段と、
前記収集された各データから属性の値を抽出し、各データが、少なくとも前記抽出した属性の値、識別ＩＤ、更新日時からなる構成のデータベース（以下、新データベース）を生成する属性情報抽出手段と、
前記生成された新データベース内の属性の値が同一とみなせるデータ集合を同一グループに分類する名寄せ手段と、
前記新データベースと前記旧データベースとの間で、属性の値が同一とみなせるデータ同士を同一と判断して両データベース間の各データを対応付けする結合手段と、
前記新データベース中のデータの識別ＩＤや更新日時などの情報と、前記データと対応付けされた前記旧データベース中のデータの識別ＩＤや更新日時などの情報とを比較することによって、前記新データベース内の該当データに更新情報を付与する更新情報付与手段と、を有することを特徴とするデータベース生成装置。
【請求項２】　請求項１記載のデータベース生成装置において、更新情報付与手段は、新データベースのデータの識別ＩＤや更新日時の情報が、旧データベースの対応するデータの識別ＩＤや更新日時の情報と不一致の場合は更新あり、一致の場合は更新なしを表わす更新情報を新データベースの該当データに付与し、旧データベースに新データベースのデータと対応付けられたデータが存在しない場合には、新データベースの該当データに新規を表わす更新情報を付与することを特徴とするデータベース生成装置。
【請求項３】　複数地点からデータを収集してデータベースを自動生成する方法であって、
各前記地点から、名義や住所などの属性の値を含むデータ、該データの識別ＩＤ、更新日時情報を収集するデータ収集過程と、
前記収集された各データから属性の値を抽出し、各データが、少なくとも前記抽出した属性の値、識別ＩＤ、更新日時からなる構成のデータベース（以下、新データベース）を生成する属性情報抽出過程と、
前記生成された新データベース内の属性の値が同一とみなせるデータ集合を同一グループに分類する名寄せ過程と、
前記新データベースと過去に生成して保持されている旧データベースとの間で、属性の値が同一とみなせるデータ同士を同一と判断して両データベース間の各データを対応付けする結合過程と、
前記新データベース中のデータの識別ＩＤや更新日時などの情報と、前記データと対応付けされた前記旧データベース中のデータの識別ＩＤや更新日時などの情報とを比較することによって、新データベース内の該当データに更新情報を付与する更新情報付与過程と、を有することを特徴とするデータベース生成方法。
【請求項４】　請求項３記載のデータベース生成方法において、
前記名寄せ過程あるいは前記結合過程の少なくとも一方を省略し、
更新情報付与過程では、少なくとも前記結合過程が省略された場合には、データ中の不変情報にもとづいて新データベースのデータと旧データベースのデータとの対応を認識することを特徴とするデータベース生成方法。
【請求項５】　複数地点からデータを収集してデータベースを生成するための、コンピュータで実行可能なプログラムであって、
各前記地点から、名義や住所などの属性の値を含むデータ、該データの識別ＩＤ、更新日時情報を収集するデータ収集プロセスと、
前記収集された各データから属性の値を抽出し、各データが、少なくとも前記抽出した属性の値、識別ＩＤ、更新日時からなる構成のデータベース（以下、新データベース）を生成する属性情報抽出プロセスと、
前記生成された新データベース内の属性の値が同一とみなせるデータ集合を同一グループに分類する名寄せプロセスと、
前記新データベースと過去に生成した旧データベースとの間で、属性の値が同一とみなせるデータ同士を同一と判断して両データベース間の各データを対応付けする結合プロセスと、
前記新データベース中のデータの識別ＩＤや更新日時などの情報と、前記データと対応付けされた前記旧データベース中のデータの識別ＩＤや更新日時などの情報とを比較することによって、前記新データベース内の該当データに更新情報を付与する更新情報付与プロセスと、を有することを特徴とするデータベース生成プログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、インターネット等のネットワーク上に分散配置され、店などの案内情報等を独立に管理・運営している複数のサーバ等からデータを収集し、検索・案内するためのデータベースを生成する装置及び方法、並びにそのプログラムに関する。
【０００２】
【従来の技術】店の案内情報などのデータは、いくつかの組織において独立に作成され、必要に応じて更新される場合が多い。一つの組織が所有しているデータ集合が全ての店の案内情報をカバーしているわけではないので、独立に作成・更新されているこれらのデータ集合を

統合すれば、より充実した情報検索サービスを行うことができる。各組織が保有するデータ集合は、インターネット等のネットワークに接続されたコンピュータ内に保管され、閲覧に供される。以後、このようなコンピュータを情報サーバと呼ぶことにする。

【０００３】複数の情報サーバからデータ集合を収集し、データベースを生成する従来の技術においては、複数の情報サーバから収集したデータ集合を単純にマージしたものをデータベースとしていた。生成されたデータベース中の各データには、該データが存在する情報サーバ内の元データへのリンク情報が付与されており、ユーザが端末を用いてデータベースからデータを検索した際は、端末に表示されたデータに付随するリンク情報により、該データの元データにアクセスを行うことができる。図１１は、データベースから、例えば業種が「中華」で住所が「新宿区神楽坂」である店を検索したときの、従来の検索結果表示画面の一例を示したものである。ユーザがリンク情報を画面上でクリックすることにより、リンク先の店の詳細画面が表示される。

【０００４】

【発明が解決しようとする課題】いくつかの組識において独立に作成されたデータ集合では、同一店舗でも名義や住所が異なる形式、表現で登録されることが多い。従って、複数の情報サーバから収集したデータ集合を単純にマージしてデータベースを生成する従来の技術では、重複する同一店舗を一つにまとめることができず、検索結果の店舗群の中に、同一店舗が複数混在して表示されることがある。このような場合、検索結果が冗長に多くなり、ユーザは不必要なデータの中身まで吟味し、それが既に見たデータと同じかどうか判断するといった煩雑な作業を強いられることになる。例えば、図１１の検索結果表示画面では、１番目の店舗と４番目の店舗が同一であり、３番目の店舗と６番目の店舗が同一である。

【０００５】また、店などの情報を検索するユーザにとって特に興味のあるのは、店の新しい情報や、新規に出来たお店などの情報である。このため、データ集合の収集とデータベースの生成を定期的に実行する場合、生成されたデータベースからデータを検索するユーザにとっては、表示されたデータの内、どのデータが更新されたものであるか、または新規のものであるかの情報がついていると、新しい情報を迅速に取得することが出来る。しかしながら、従来技術においては、各データにこのような更新情報は付与されない。

【０００６】本発明の目的は、ネットワーク上に分散して存在している複数の情報サーバ等からデータ集合を収集して、冗長性がないようにデータを統合し、かつデータの更新情報が付加されたデータベースを生成することを可能とするデータベース生成装置及び方法、並びにそのためのデータベース生成プログラムを提供することにある。

【０００７】

【課題を解決するための手段】本発明のデータベース生成装置は、過去に生成されたデータベース（旧データベース）を記憶する記憶手段と、複数地点から、名義や住所などの属性の値を含むデータ、該データの識別ＩＤ、更新日時などの情報を収集するデータ収集手段と、前記収集された各データから属性の値を抽出し、各データが、前記抽出した属性の値、識別ＩＤ、更新日時などからなる構成のデータベース（新データベース）を生成する属性情報抽出手段と、前記生成された新データベース内の属性の値が同一とみなせるデータ集合を同一グループに分類する名寄せ手段と、新データベースと前記旧データベースとの間で、属性の値が同一とみなせるデータ同士を同一と判断して両データベース間の各データを対応付けする結合手段と、前記新データベース中のデータの識別ＩＤや更新日時などの情報と、前記データと対応付けされた前記旧データベース中のデータの識別ＩＤや更新日時などの情報とを比較することによって、前記新データベース内の該当データに更新情報を付与する更新情報付与手段とを有することを特徴とする。

【０００８】名寄せ手段では、生成された新データベースにおいて、重複するデータが一つにされている。このため、このデータベースからユーザの要求に合致するデータを検索し表示したとき、同一店舗のデータが複数個表示されることはなく、検索結果の把握がより容易に行える。また、結合手段は、生成した新データベースと、前回生成した旧データベースとの間で、同一店舗等のデータを特定し、更新情報付与手段では、それらの識別ＩＤ（例えば名称）や更新日時などを比較することにより、データの更新情報を導出するので、最終的に生成されたデータベースは、データの更新情報が付与された上で、データを表示することが可能である。

【０００９】次に、本発明のデータベース生成方法は、複数地点から、名義や住所などの属性の値を含むデータ、該データの識別ＩＤ、更新日時などの情報を収集するデータ収集過程と、前記収集された各データから属性の値を抽出し、各データが、前記抽出した属性の値、識別ＩＤ、更新日時などからなる構成のデータベース（新データベース）を生成する属性情報抽出過程と、前記生成された新データベース内の属性の値が同一とみなせるデータ集合を同一グループに分類する名寄せ過程と、新データベースと過去に生成して保持されている旧データベースとの間で、属性の値が同一とみなせるデータ同士を同一と判断して両データベース間の各データを対応付けする結合過程と、前記新データベース中のデータの識別ＩＤや更新日時などの情報と、前記データと対応付けされた前記旧データベース中のデータの識別ＩＤや更新日時などの情報とを比較することによって、新データベース内の該当データに更新情報を付与する更新情報付与過程とを有することを特徴とする。

【００１０】次に、本発明のコンピュータで実行可能なプログラムは、複数地点から、名義や住所などの属性の値を含むデータ、該データの識別ＩＤ、更新日時などの情報を収集するデータ収集プロセスと、前記収集された各データから属性の値を抽出し、各データが、前記抽出した属性の値、識別ＩＤ、更新日時などからなる構成のデータベース（新データベース）を生成する属性情報抽出プロセスと、前記属性情報抽出プロセスで生成された新データベース内の属性の値が同一とみなせるデータ集合を同一グループに分類する名寄せプロセスと、前記新データベースと過去に生成した旧データベースとの間で、属性の値が同一とみなせるデータ同士を同一と判断して両データベース間の各データを対応付けする結合プロセスと、前記新データベース中のデータの識別ＩＤや更新日時などの情報と、前記データと対応付けされた前記旧データベース中のデータの識別ＩＤや更新日時などの情報とを比較することによって、前記新データベース内の該当データに更新情報を付与する更新情報付与プロセスとを有することを特徴とする。

【００１１】

【発明の実施の形態】以下に、本発明の一実施例について、図面を参照して説明する。図１は、本発明の一実施の形態のデータベース生成装置の構成例を示す図である。データベース生成装置１０は、インターネット等のネットワーク４０に接続されるものであり、該ネットワーク４０を介して、店の案内情報などのデータ集合を管理・運営している複数の情報サーバ２０と、ユーザが使用するユーザ端末３０とに接続している。ネットワーク４０に接続された個々の情報サーバ２０はそのＵＲＬ（uniform resource locator）によって識別される。各情報サーバ２０は、その内部にデータ集合をもち、当該データ集合を、他の情報サーバとは独立して管理・運営している。したがって、同一店舗の案内情報などが、複数の情報サーバ２０内に存在することが多々ある。データベース生成装置１０自体も、その内部にデータ集合をもって、管理・運営するという形態をとっていてもよい。ユーザ端末３０としては、典型的には、WWWソフトウェア（WWWブラウザ）が組み込まれたパーソナルコンピュータ（パソコン）や携帯端末が使用される。各ユーザは、該ユーザ端末３０を用いて情報検索などを行うほか、必要ならデータベース生成装置１０に対して要望等を通知する。

【００１２】データベース生成装置１０は、データ収集手段１１、属性情報抽出手段１２、名寄せ手段１３、結合手段１４、更新情報付与手段１５の各処理手段、及び、データベース格納部１６を具備する。データベース格納部１６には、過去（ここでは前回とする）に生成したデータベース（旧データベース）１７と新規に生成したデータベース（新データベース）１８が格納される。データベース生成装置１０は、所謂コンピュータで実現されるものであり、各処理手段１１～１５はＣＰＵやその内蔵メモリ（ＲＡＭ、ＲＯＭ等）が受け持ち、データベース格納部１６はハードディスク、その他の外部記憶装置などが受け持つ。

【００１３】なお、データベース生成装置１０自体、ユーザ端末３０から検索要求を受けて情報検索サービスを実施してもよい。この場合、図１では省略したが、データベース生成装置１０は情報検索手段も具備することになる。また、情報検索装置は該データベース生成装置１０とは別構成として、データベース生成装置１０で生成したデータベースを別の情報検索装置で利用することでもよい。

【００１４】図２は、本発明の一実施形態のデータベース生成方法のフローチャートを示す図である。以下、図２のフローチャートに従って、本データベース生成装置１０の動作を詳しく説明する。

【００１５】データベース生成装置１０では、データ収集手段１１において、一定期間や特定日時ごと（例えば、１日、１週間、毎月曜日など）に、各情報サーバ２０にアクセスし、各情報サーバ２０内のデータ（データ集合）を収集する（ステップ１１１）。ここで、各データは一つのファイルであり、全てのファイルがあるディレクトリ配下にあるものとする。このディレクトリの所在は、データベース生成装置１０の管理者と各情報サーバ２０の管理者との間であらかじめ取り決めがなされており、データ収集手段１１は、各情報サーバ２０の該ディレクトリ配下のファイル群をダウンロードし、例えばＲＡＭやハードディスク等に一時的に格納する。ここで、ファイルとともに、データの名称となるファイル名（これが当該データのリンク情報となる）とファイルの更新日時の情報も取得する。

【００１６】図３は、情報サーバＡ及びＢからダウンロードしたデータの一例を示したものである。この例では、同一店舗「紅蘭亭」のデータが情報サーバＡにもＢにも登録されているものとし、そのデータを示したものである。図３に示すように、情報サーバＡとＢでは、同一店舗「紅蘭亭」でも、名義や住所等が異なる形式、表現で登録されている。

【００１７】次に、属性情報抽出手段１２において、上記データ収集手段１１で収集した各データから名義や住所などの該データを特徴付ける属性の値を抽出する（ステップ１１２）。各データファイルは典型的にはＨＴＭＬ文書やＸＭＬ文書であり、ユーザはユーザ端末３０を用いてWWWソフトウェア（WWWブラウザ）から該当ファイルのＵＲＬにアクセスすることにより、その内容を閲覧することができるものである。各データファイルの内容が、どういった属性からなり、各属性がどのようなフォーマットで記述されているかといったフォーマット情報は、各情報サーバ２０ごとに決められている。ここでは、各情報サーバに対応したデータファイルフォー

マット解析ルーチンを属性情報抽出手段１２が保持しているとする。属性情報抽出手段１２は、各情報サーバに対応したデータファイルフォーマット解析ルーチンにより、データファイルから名義や住所などの属性値を抽出する。次に、属性情報抽出手段１２では、抽出した名義や住所などの属性値と該データが存在する情報サーバ名及び該データのリンク情報及び更新日時の情報等からなるデータ（レコード）を作成し、このようなデータが集積したデータベースを生成してデータベース格納部１６に格納する。この新たに生成されたデータベースを新データベース１８とする。また、前回（１日前、１週間前など）、同様に各情報サーバ２０からデータを収集して生成し、後述の名寄せ、結合、更新情報付与等の処理を施したデータベースを旧データベース１７とする。

【００１８】図４は、新たに生成されたデータベース（新データベース）１８の一例を示したもので、（ａ）は情報サーバAのデータ、（ｂ）は情報サーバBのデータである。ここでは、抽出する属性として業種、名義、住所をとっている。業種体系は情報サーバ２０ごとに一般に異なっている。また、同一店舗のデータでも、情報サーバが異なれば、名義や住所の表記には揺れがある。

【００１９】なお、情報サーバ２０が、店のデータファイルの他に、各店の名義や住所、電話番号、リンク情報などの基本情報のみが記載されているデータのリストからなるファイルをもっている場合、データ収集手段１１において、データファイル群の代わりに、そのようなリストファイルをダウンロードしてもよい。この場合、属性情報抽出手段１２においては、リストファイルの各データから名義、住所、リンク情報などを抽出し、抽出したリンク情報をもとに、再び情報サーバ２０にアクセスし、データファイルの更新日時情報を取得する。そして、同様に図４のような新データベース１８を生成する。

【００２０】次に、名寄せ手段１３において、新データベース１８内の名義や住所などの属性の値が同一とみなせるデータ（レコード）を同一グループに分類する（ステップ１１３）。即ち、同一店舗として名寄せする。

【００２１】例えば、図４に示した新データベース１８の任意の２データ間において、名義及び住所の属性の値同士を照合し、マッチしたレコード同士を同一グループに分類する。名義文字列や住所文字列の照合方法には例えば次のようなものが考えられる。一つには完全一致したときマッチするとみなす方法（完全一致と呼ぶ）があり、また、両方に共通して含まれる文字の数の割合がある閾値以上のときマッチするとみなす方法（文字単位一致と呼ぶ）がある。他には、文字列を単語分割して両方に共通して含まれる単語の数の割合がある閾値以上のときマッチするとみなす方法（単語単位一致と呼ぶ）がある。いずれの方法も、漢数字を算用数字に変換したり、英字を大文字に統一化するといった表記の揺れを解消する処理を事前に行うことにより、より照合の精度を高めることが可能である。　照合の結果、図４の例では、１番目と４番目のデータ（レコード）がマッチし、３番目と６番目のデータがマッチする。このマッチしたレコード同士を同一グループに分類する。ここで、各グループを通常のデータと区別して、名寄せデータと呼ぶことにする。

【００２２】名寄せ手段１３では、各名寄せデータの名義や住所の属性値として、例えば当該名寄せデータに含まれるデータの名義や住所の属性値から一つだけ選んで、その値そのものを用いるか、あるいは正規化した値に変換する。また、各データの業種名は、データベース生成装置１０独自の業種体系における対応する業種名に変換する。

【００２３】図４について、こうして更新された新データベース１８の一例を図５に示す。例えば、データベース生成装置１０独自の業種体系では、業種として「和食」、「中華」などがあり、図４におけるデータの業種名はいずれも「中華」に変換される。図５において、同一グループに分類された１番目と４番目のデータの業種名はともに「中華」に変換されるので、名寄せデータとしての業種名も「中華」となる。３番目と６番目のデータに関しても同様である。また、名寄せデータの名義や住所の属性値としては、１番目と４番目のデータでは、名義は「紅蘭亭」を選択し、住所は「新宿区神楽坂１－２－３」を選択している。同様に、３番目と６番目のデータでは、名義は「大竹亭」を選択し、住所は「新宿区神楽坂３－８－６」を選択している。なお、図５中の新データベース１８の「更新情報」の欄は後述の更新情報付与手段１５で書き替えられるもので、ここでは全て空（ＮＵＬＬ）としておく。

【００２４】ここで、どの属性値同士をどの照合方法で照合させるかといった照合ルールは、名寄せ手段１３を実現するプログラム内に記述してもよいし、データベース生成装置１０内の、プログラムが参照する外付けテーブルに記述して、データベース生成装置１０の管理者が、この外付けテーブルを自由に変更できるようにしておいてもよい。

【００２５】図６は、このような外付けテーブルの内容の一例である。図６（ａ）では、データが一致する基準を記述する。この例では、照合項目として名義と住所を指定している。名義の照合結果の評価値が９０点以上かつ住所の照合結果の評価値が８０点以上の場合、あるいは名義の照合結果の評価値が８０点以上かつ住所の照合結果の評価値が９０点以上の場合、２データが一致すると判定する。図６（ｂ）では、名義の照合方法を記述する。ここでは、照合方法として完全一致、文字単位一致、単語単位一致を指定しており、各方法による照合を行う。完全一致の照合処理で一致したならば評価値１００とし、一致しなければ評価値０とする。文字単位一致

の照合結果の評価値は一致した文字の数の割合に１００を乗じたものとする。単語単位一致の照合結果の評価値も一致した単語の数の割合に１００を乗じたものとする。一番高い評価値を返した照合方法の評価値を名義の評価値とする。図６（ｃ）では、住所の照合方法を同様に記述する。ここでは、照合方法として完全一致、単語単位一致を指定している。一番高い評価値を返した照合方法の評価値を住所の評価値とする。

【００２６】次に、結合手段１４において、データベース格納部１６にある、名寄せ後の新データベース１８と、前回各情報サーバ２０からデータを収集して、生成した旧データベース１７との間で、名義や住所などの属性の値が同一とみなせる名寄せデータ同士を同一と判断してリンク付けし、対応付けする（ステップ１１４）。例えば、新旧データベース１７、１８内の同一と判断された両データに、同一なデータであることを示す情報を付与するなどしてリンク付けし、対応付けする。

【００２７】ここでは、情報サーバ２０において、同一データのリンク情報が時の経過とともに変わり得るという前提であるものとする。各データの更新情報を導出するにあたっては、新データと旧データの更新日時などを比較する必要があり、そのためには、新旧データベースにおいて、どのデータが同一かを判断しなければならない。リンク情報が不変であれば、リンク情報が同一かで判断できるが、リンク情報が変わり得るという前提のもとでは、データがもつ名義や住所の属性値が同一かで判断する必要があるわけである。ここで、同一データであっても時の経過とともに、名義などが微妙に変更される場合もありうるので、照合は、表記の揺れを考慮して行う。具体的には、例えば完全一致以外に文字端単位一致や単語単位一致といった照合方法で行う。基本的には名寄せの場合と同様である。また、照合の対象となる項目を、例えば名義のみにすると、同一店の住所が変更しても、新旧のデータはマッチすることになる。このように、どのような条件で新旧のデータを同一視するかは、照合ルールを変更することにより調節可能である。図７に、外付けテーブルに記述する照合ルールにおけるデータ一致基準の一例を示す。ここでは照合項目として名義のみを指定した例を示している。名義の照合方法の記述は、例えば図６と同様にすればよい。

【００２８】図８は、旧データベース１７の一例である。便宜上、図８では、各データは前々回から更新がなかったとしている。結合手段１４では、図５に示した新データベース１８の各名寄せデータと同一な旧データベース１７の名寄せデータを、名義のみあるいは名義及び住所の属性値同士を照合することによって特定する。その結果、図５の新データベース１８の１番目、２番目、３番目の名寄せデータがそれぞれ、図８の旧データベース１７の１番目、２番目、３番目の名寄せデータにリンク付けされる。図５の新データベース１８の４番目の名寄せデータにリンク付けされる名寄せデータは、図８の旧データベース１７には存在しない。なお、リンク付けされた名寄せデータ内の同一の対応情報サーバをもつデータ同士も、同一のデータとしてリンク付けされる。以後、図５、図８の各データを上から何番目かで表現する。

【００２９】次に、更新情報付与手段１５において、新データベース１８のデータのリンク情報や更新日時の情報と、結合手段１４により該データと同一と判断された旧データベース１７中のデータのリンク情報や更新日時の情報とを比較することにより、新データベース１８中の該当データに更新情報を設定･付与する（ステップ１１５）。即ち、新データベース１８中のデータとリンク付けされた旧データベース１７のデータがあり、かつリンク情報または更新日時が変更されているとき、該データは更新されたものと判断し、いずれも変更されていないとき、該データは更新なしと判断し、新データベース１８中の該当データの更新情報を「更新」あるいは「更新なし」とする。また、新データベース１８中のデータとリンク付けされた旧データベース１７のデータがない場合、該データは新規に作成されたものと判断し、新データベース１８中の該当データの更新情報を「新規」とする。

【００３０】例えば、図５の新データベース１８の１番目のデータは、リンク付けされた図８の旧データベース１７の１番目のデータと、リンク情報が同じで、更新日時が変わっているので、当該データは更新されたものと判断する。

【００３１】図５の新データベース１８の２番目のデータは、リンク付けされた図８の旧データベース１７の２番目のデータと比べ、リンク情報も更新日時も不変なので、当該データは更新されていないものと判断する。図５の新データベース１８の４番目のデータについても同様である。

【００３２】図５の新データベース１８の３番目のデータは、リンク付けされた図８の旧データベース１７の３番目のデータと比べ、更新日時は変わらないが、リンク情報が変わっているので、当該データは更新されたものと判断する。

【００３３】図５の新データベース１８の５番目のデータは、名寄せデータとしては、図８の旧データベース１７の３番目の名寄せデータとリンクしているが、データとしてリンク付けされたデータは図８の旧データベース１７にないので、新規に作成されたものと判断する。

【００３４】図５の新データベース１８の６番目のデータにリンク付けされたデータは、図８の旧データベース１７にないので、当該データは新規に作成されたものと判断する。

【００３５】このようにして、図５の新データベース１８と図８の旧データベース１７の場合、図９に示すよう

な更新情報の付与された新データベース１８が最終的に生成される。更新情報付与手段１５では、この最終的に生成された新データベース１８でもって旧データベース１７を上書きする。

【００３６】以上によりデータベースの生成が終了する。最終的に生成されたデータベースにユーザ端末３０からアクセスし、ユーザの要求に合致するデータを検索し表示したときには、名寄せデータの業種、名義、住所の情報と、該データが存在する情報サーバ２０内のファイルへのリンク情報及び更新情報が表示される。図１０は、図９の生成データベースにより、業種が「中華」で住所が「新宿区神楽坂」である店を検索したときの検索結果の表示例である。ユーザはこのリンク情報を画面上でクリックすることにより、リンク先のファイルの内容である店の詳細情報にアクセスすることができる。

【００３７】以上、本発明の典型的な一実施例について述べたが、名寄せ前の旧データベースを保持しておき、結合手段１４のリンク付けを、名寄せ後の新旧データベース（図５及び図８）間で実行するのではなく、名寄せ前の新旧データベース（図４及び図４相当の古いデータベース）間で実行してもよい。例えば、この場合、対応情報サーバが同一なデータ同士を照合させる。

【００３８】情報サーバ２０において、同一データのリンク情報が時の経過とともに変わりえても、各データにとって恒久的に不変なＩＤ情報がデータ中に含まれている場合は次のように処理を行うこともできる。属性情報抽出手段１２において、このＩＤ情報を抽出し、結合手段１４におけるリンク付けを、新データベース中の各データに対し、当該データと同一のＩＤ情報をもつ旧データベース中のデータをリンク付けることによって行う。

【００３９】また、情報サーバ２０において、同一データのリンク情報が常に不変であれば、結合手段１４のリンク付けは必要でない。更新情報付与手段１５において、生成したデータベース（名寄せ前のものでも名寄せ後のものでもよい）中のデータの更新日時が、前にデータ集合を収集した時点以降ならば、該データは更新されたデータか新規データであることが分かる。さらに、該データの対応情報サーバとリンク情報がともに同一であるデータが、前に生成したデータベース中にあれば、該データは更新されたデータであり、なければ新規データであることが判明する。

【００４０】上記に挙げた以外にも、本発明は特許請求の範囲の記載内で、様々な変更や拡張が可能である。例えば、名寄せ手段や名寄せ過程をなくして、各データの更新情報のみを付与する構成も考えられる。

【００４１】なお、図１で示した装置における各部の一部もしくは全部での処理機能をコンピュータのプログラムで構成し、そのプログラムをコンピュータを用いて実行して本発明を実現することができること、あるいは、図２で示した処理手順をコンピュータのプログラムで構成し、そのプログラムをコンピュータに実行させることができることは言うまでもない。また、コンピュータでその処理機能を実現するためのプログラム、あるいは、コンピュータにその処理手順を実行させるためのプログラムを、そのコンピュータが読み取り可能な記録媒体、例えば、ＦＤや、ＭＯ、ＲＯＭ、メモリカード、ＣＤ、ＤＶＤ、リムーバブルディスクなどに記録して、保存したり、提供したりすることができるとともに、インターネット等のネットワークを通してそのプログラムを配布したりすることが可能である。

【００４２】

【発明の効果】以上説明したように、本発明によれば、生成されたデータベースからユーザの要求に合致するデータを検索したとき、重複データがなく、かつデータの更新情報が付加された形で検索結果を表示することが可能となる。

【図面の簡単な説明】

【図１】本発明の一実施形態のデータベース生成装置の構成を示す図である。

【図２】本発明の一実施形態のデータベース生成方法のフローチャート図である。

【図３】情報サーバからダウンロードしたデータの一例を示す図である。

【図４】属性情報抽出手段で生成された新データベースの一例を示す図である。

【図５】名寄せ手段で更新された新データベースの一例を示す図である。

【図６】名寄せ手段で適用される照合ルールの一例を示す図である。

【図７】結合手段で適用される照合ルールの一例を示す図である。

【図８】前回生成した旧データベースの一例を示す図である。

【図９】更新情報付与手段で更新情報が付与された新データベースの一例を示す図である。

【図１０】本発明により生成されたデータベースからの検索結果の表示画面の一例を示す図である。

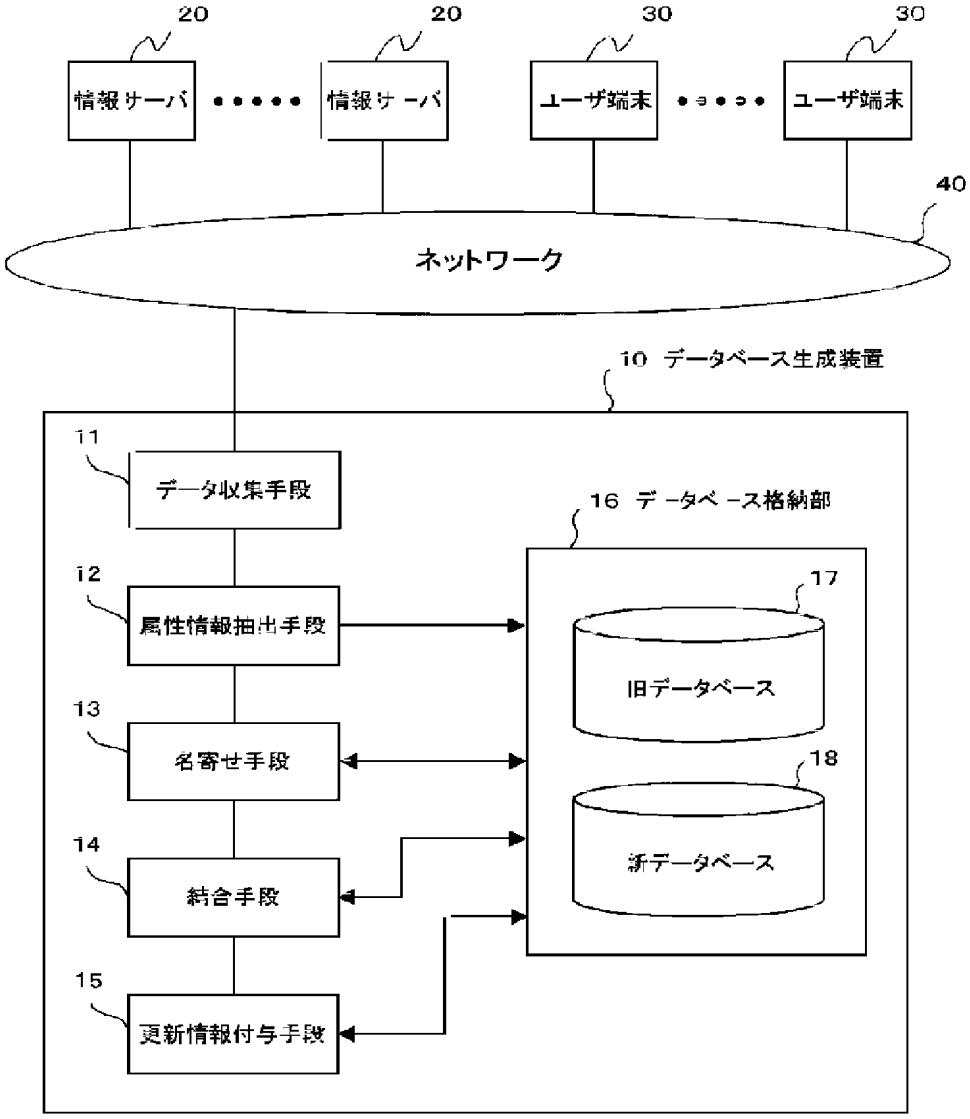
【図１１】従来のデータベース生成技術により生成したデータベースからの検索結果の表示画面の例を示す図である。

【符号の説明】

１０　データベース生成装置
１１　データ収集手段
１２　属性情報抽出手段
１３　名寄せ手段
１４　結合手段
１５　更新情報付与手段
１６　データベース格納部
１７　旧データベース
１８　新データベース

２０　情報サーバ　　４０　ネットワーク
３０　ユーザ端末

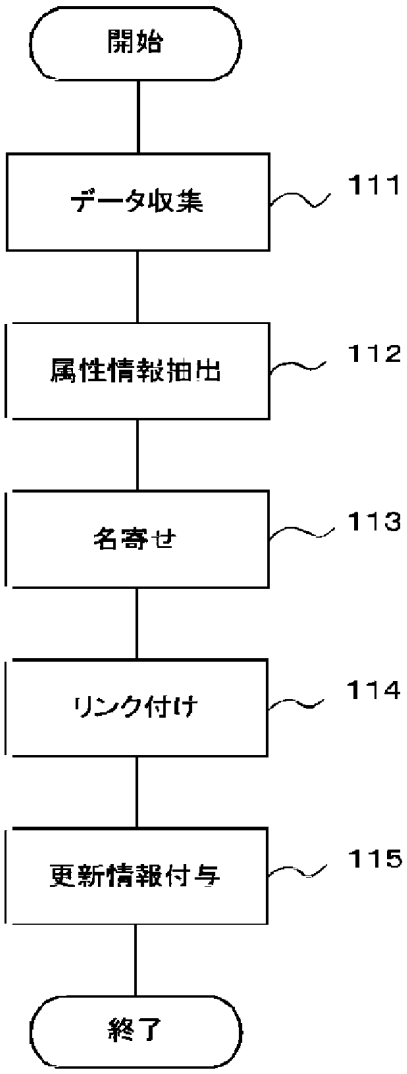
【図１】



【図７】

データー致基準
名義（≧80）

【図2】



【図3】

情報サーバAの一つのデータ

紅蘭亭
住所　：新宿区神楽坂1−2−3
電話番号：XXX −YYY
毎月20日は感謝デー
……

リンク情報	更新日時
http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/10　15:30

情報サーバBの一つのデータ

ラーメンショップ紅蘭亭
住所　：新宿区神楽坂1丁目2番地3号
最寄駅　：新宿
ランチ 500円
日曜定休
……

リンク情報	更新日時
http://www.bserv.co.jp/data/data5138.html	2001/06/23　13:45

【図5】

業種	名義	住所	情報サーバ	リンク情報	更新日時	更新情報
中華	紅蘭亭	新宿区神楽坂1−2−3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/11 15:30	
			B	http://www.bserv.co.jp/data/data5138.html	2001/06/23 13:45	
中華	大森飯店	新宿区神楽坂5−2−4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14	
中華	大竹亭	新宿区神楽坂3−8−6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19	
			B	http://www.bserv.co.jp/data/data5568.html	2001/08/11 14:10	
中華	龍王飯店	新宿区神楽坂2−7−4	B	http://www.bserv.co.jp/data/data5238.html	2001/08/11 10:27	

【図４】

	業種	名義	住所	情報サーバ	リンク情報	更新日時
(a)	ラーメン	紅蘭亭	新宿区神楽坂 1-2-3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/11 15:30
(a)	台湾料理	大森飯店	新宿区神楽坂 5-2-4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14
(a)	中華料理	大竹亭	新宿区神楽坂 3-8-6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19
(b)	中華全般	ラーメンショップ紅蘭亭	新宿区神楽坂1丁目2番地3号	B	http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45
(b)	四川料理	龍王飯店	新宿区神楽坂2丁目7番地4号	B	http://www.bserv.co.jp/data/data5238.html	2001/08/11 10:27
(b)	広東料理	大竹亭本店	新宿区神楽坂3丁目8番地6号	B	http://www.bserv.co.jp/data/data5568.html	2001/08/11 14:10

【図６】

(a) データー致基準

名義（≧90）＋住所（≧80）
名義（≧80）＋住所（≧90）

(b) 名前の照合方法

照合方法	評価値
完全一致	100
文字単位 -致	一致した文字の数の割合 × 100
単語単位 -致	一致した単語の数の割合 × 100

(c) 住所の照合方法

照合方法	評価値
完全一致	100
単語単位 -致	一致した文字の数の割合 × 100

【図８】

業種	名義	住所	情報サーバ	リンク情報	更新日時	更新情報
中華	紅蘭亭	新宿区神楽坂 1-2-3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/10 16:23	なし
			B	http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45	なし
中華	大森飯店	新宿区神楽坂 5-2-4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14	なし
中華	大竹亭	新宿区神楽坂 3-8-6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19	なし

【図９】

業種	名義	住所	情報サーバ	リンク情報	更新日時	更新情報
中華	紅蘭亭	新宿区神楽坂１-2-3	A	http://www.aserv.co.jp/shop/shopfile0018.html	2001/08/11 15:30	更新
			B	http://www.bserv.co.jp/data/data5136.html	2001/06/23 13:45	なし
中華	大森飯店	新宿区神楽坂５-2-4	A	http://www.aserv.co.jp/shop/shopfile3870.html	2001/04/13 11:14	更新
中華	大竹亭	新宿区神楽坂3-8-6	A	http://www.aserv.co.jp/shop/shopfile2657.html	2001/08/01 13:19	なし
			B	http://www.bserv.co.jp/data/data5568.html	2001/08/11 14:10	新規
中華	龍王飯店	新宿区神楽坂2-7-4	B	http://www.bserv.co.jp/data/data5238.html	2001/08/11 10:27	新規

【図１０】

業種	名義	住所	リンク情報	
中華	紅蘭亭	新宿区神楽坂１-2-3	情報サーバA 更新	情報サーバB
中華	大森飯店	新宿区神楽坂５-2-4	情報サーバA 更新	
中華	大竹亭	新宿区神楽坂3-8-6	情報サーバA	情報サーバB 新規
中華	龍王飯店	新宿区神楽坂2-7-4	情報サーバB 新規	

【図１１】

業種	名義	住所	リンク情報
中華	紅蘭亭	新宿区神楽坂１-2-3	情報サーバA
中華	大森飯店	新宿区神楽坂５-2-4	情報サーバA
中華	大竹亭	新宿区神楽坂3-8-6	情報サーバA
中華	ラーメンショップ紅蘭亭	新宿区神楽坂1丁目2番地3号	情報サーバB
中華	龍王飯店	新宿区神楽坂2丁目7番地4号	情報サーバB
中華	大竹亭本店	新宿区神楽坂3丁目8番地6号	情報サーバB

フロントページの続き

Ｆターム(参考)　5B075 KK03 KK07 ND20 NK02 NK46
NR03 NR14 NR20 UU40
5B082 EA08 EA10

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2004-227165

(43)Date of publication of application : 12.08.2004

(51)Int.Cl. G06F 17/30

(21)Application number : 2003-012517

(22)Date of filing : 21.01.2003

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

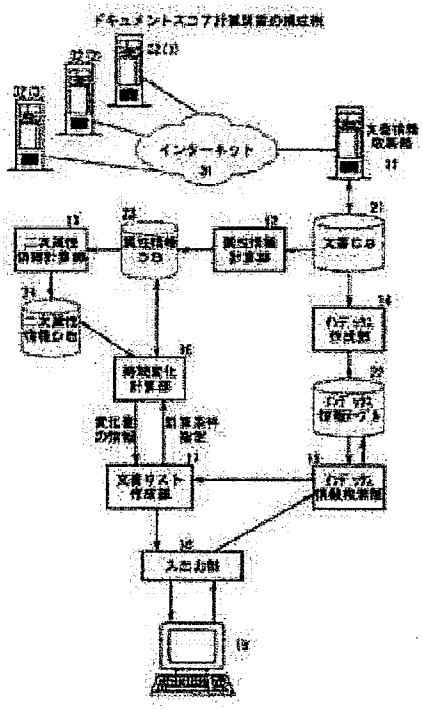
(72)Inventor : OMORI NOBUYUKI
INOUE TAKASHI
TAKENO HIROSHI
IBARAKI HISASHI

(54) DOCUMENT SCORE CALCULATION METHOD AND DEVICE, AND PROGRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a document score calculating method by which a user can acquire results more closer to a necessary document in retrieving large amounts of documents, and to provide a device and a program therefor.

SOLUTION: In this document score calculation method for collecting the information of a plurality of documents to be updated, and for assigning scores to be used for extracting a specific document from a collected document group to each document, attribute information showing characteristics is extracted from each document, and the attribute information extracted at two or more points of time is simultaneously stored for each document, and the secular change of the attribute information is calculated based on a plurality of attribute information stored for each document, and the calculated secular change of the attribute information is reflected on the calculation of the scores.



* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]
In a document score calculation method for assigning a score used in order to collect information on two or more documents which may be updated and to extract a specific document out of a collected document group for every document,
Attribution information showing the characteristic is extracted from each document,
Attribution information extracted for every document at the two or more times is held simultaneously,

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]
This invention relates to the document score calculation method, device, and program which are used when extracting automatically the candidate of the document which a user should refer to out of the document of a large number which exist, for example on the Internet etc.
[0002]
[Description of the Prior Art]
It is indicated by the nonpatent literature 1 about the art of the conventional common text browsing system. It is indicated by the nonpatent literature 2 about the conventional system to which it refers by collecting information from Web.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL FIELD

[Field of the Invention]
This invention relates to the document score calculation method, device, and program which are used when extracting automatically the candidate of the document which a user should refer to out of the document of a large number which exist, for example on the Internet etc.
[0002]

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

PRIOR ART

[Description of the Prior Art]
It is indicated by the nonpatent literature 1 about the art of the conventional common text browsing system. It is indicated by the nonpatent literature 2 about the conventional system to which it refers by collecting information from Web.
For example, various computers are connected on a network like the Internet or LAN, and huge document information exists in the accessible state. A possibility that the target document exists in a vast quantity of these documents is high. However, the number of documents is huge, and since the places where each document exists also differ, respectively, it is difficult [it] to discover the specific document which a user needs out of such huge document information.
[0003]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

EFFECT OF THE INVENTION

[Effect of the Invention]
According to this invention, since change of attribution information can be investigated at two or more:00 using the attribution information of a point and it can be reflected in search results, an improvement of retrieval precision has an effect.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]
In the system which provides the above search services, about the document information updated, for example by collecting document information periodically at the fixed interval, it updates to the newest document information, and old document information is discarded. And a search is performed only based on the collected newest text data.
[0005]
However, when searching the target document using the conventional search service, what the document which a user does not need is extracted from as search results in many cases is the actual condition.
Namely, since various documents, such as a reliable document, an unreliable document, a high document of utility value, and a low document of utility value, are intermingled on the Internet, Only by identifying the conformity of a keyword, it is unavoidable that an unreliable document and the low document of utility value are extracted as search results.
[0006]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

MEANS

[Means for Solving the Problem]
In a document score calculation method for assigning a score used in order for Claim 1 to collect information on two or more documents which may be updated and to extract a specific document out of a collected document group for every document, Based on two or more attribution information which holds simultaneously attribution information which extracted attribution information showing the characteristic from each document, and was extracted for every document at the two or more times, and is held for every document, aging of attribution information is calculated and called-for aging of attribution information is reflected in calculation of a score.
[0008]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a block diagram showing the example of composition of a document score computing device.
[Drawing 2]It is a flow chart which shows the procedure of temporal change computation.
[Drawing 3]It is a flow chart which shows the procedure of document list creation processing.
[Drawing 4]It is a mimetic diagram showing the composition of the attribution information DB.
[Drawing 5]It is a mimetic diagram showing the composition of the secondary attribute information DB.
[Drawing 6]It is a mimetic diagram showing the example of composition of a document size table.
[Drawing 7]It is a mimetic diagram showing the example of composition of the number table of tags.
[Drawing 8]It is a mimetic diagram showing the example of composition of the number table of links.
[Drawing 9]It is a mimetic diagram showing the example of composition of an update date table.
[Drawing 10]It is a mimetic diagram showing the example of composition of a linking number table.
[Drawing 11]It is a mimetic diagram showing the example of composition of the degree-of-association table between documents (1).
[Drawing 12]It is a front view showing the example of composition of a user interface.
[Drawing 13]It is a front view showing the display example (1) of a user interface.
[Drawing 14]It is a front view showing the display example (2) of a user interface.
[Description of Notations]
11 Document information collecting part

12 Attribution information calculation part
13 Secondary attribute information calculation part
14 Index preparing part
15 Index information retrieval section
16 Temporal change calculation part
17 Document-list preparing part
18 Input output section
19 User terminal
21 Document DB
22 Index information table
23 Attribution information DB
24 Secondary attribute information DB
31 Internet
32 Document server

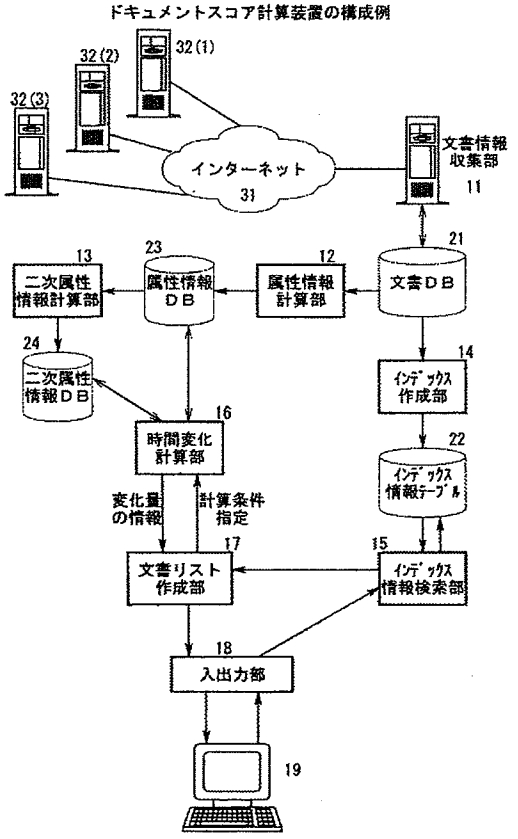
[Translation done.]

* NOTICES *

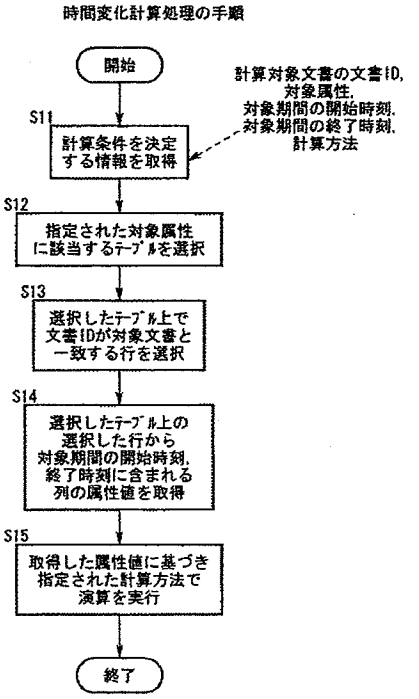
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

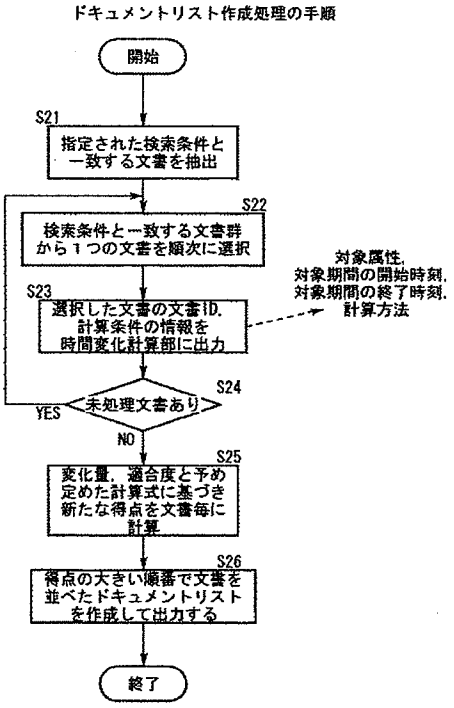
[Drawing 1]



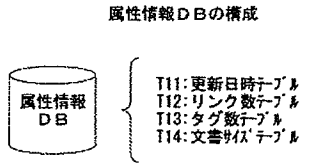
[Drawing 2]


時間変化計算処理の手順
開始
計算対象文書の文書ID,
対象属性,
対象期間の開始時刻,
対象期間の終了時刻,
計算方法
S11
計算条件を決定
する情報を取得
S12
指定された対象属性
に該当するテーブルを選択
S13
選択したテーブル上で
文書IDが対象文書と
一致する行を選択
S14
選択したテーブル上の
選択した行から
対象期間の開始時刻,
終了時刻に含まれる
列の属性値を取得
S15
取得した属性値に基づき
指定された計算方法で
演算を実行
終了

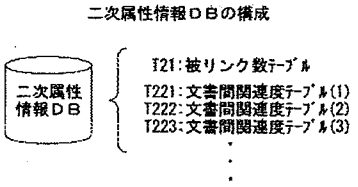
[Drawing 3]


ドキュメントリスト作成処理の手順
開始
S21
指定された検索条件と
一致する文書を抽出
S22
検索条件と一致する文書群
から1つの文書を順次に選択
対象属性,
対象期間の開始時刻,
対象期間の終了時刻,
計算方法
S23
選択した文書の文書ID,
計算条件の情報を
時間変化計算部に出力
S24
未処理文書あり
YES
NO
S25
変化量，適合度と予め
定めた計算式に基づき
新たな得点を文書毎に
計算
S26
得点の大きい順番で文書を
並べたドキュメントリスト
を作成して出力する
終了

[Drawing 4]



[Drawing 5]



[Drawing 6]

文書ｻｲｽﾞﾃｰﾌﾞﾙ

文書ID	記録年月日			
	2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	871	773	136	34
url2.co.jp	874	732	629	719
url3.ne.jp	333	633	890	777
url4.go.jp	978	249	92	174
url6.gr.jp	96	814	259	421
url7.jp	190	966	115	718
url8.com	216	51	801	185

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
17	411	55	263	853
300	138	215	403	874
95	298	276	536	247
751	942	395	7	309
118	839	738	491	884
134	106	20	62	852
382	887	875	503	578

[Drawing 7]

タグ数テーブル

文書ID	記録年月日			
	2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	32	88	3	31
url2.co.jp	53	18	9	28
url3.ne.jp	25	23	28	16
url4.go.jp	.	.	.	.
url6.gr.jp	.	.	.	.
url7.jp	.	.	.	.
url8.com				

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
56	68	82	93	97
72	42	54	71	67
26	27	29	29	37
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

[Drawing 8]

リンク数テーブル

文書ID	記録年月日			
	2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	0	1	1	1
url2.co.jp	3	8	2	2
url3.ne.jp	5	2	2	1
url4.go.jp	.	.	.	.
url6.gr.jp	.	.	.	.
url7.jp	.	.	.	.
url8.com				

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
2	2	1	0	3
7	4	4	1	3
2	2	2	2	3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

[Drawing 9]

更新日時テーブル

文書ID	記録年月日			
	2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	568	568	568	263040
url2.co.jp	.	.	.	.
url3.ne.jp	.	.	.	.
url4.go.jp	.	.	.	.
url6.gr.jp	.	.	.	.
url7.jp	.	.	.	.
url8.com				

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
356521	356521	356521	356521	356521
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

[Drawing 10]

被リンク数テーブル

文書ID	記録年月日 2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	3	1	4	1
url2.co.jp	3	9	2	0
url3.ne.jp	5	3	8	8
url4.go.jp	.	.	.	.
url6.gr.jp	.	.	.	.
url7.jp	.	.	.	.
url8.com				

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
9	6	8	3	0
5	7	6	3	1
6	7	9	9	7
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

[Drawing 11]

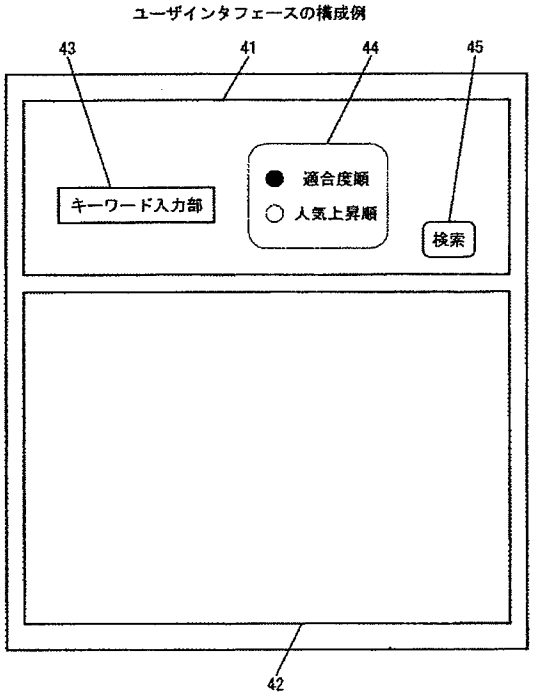
文書間関連度テーブル(1)

関連度計算対象：url1.co.jp

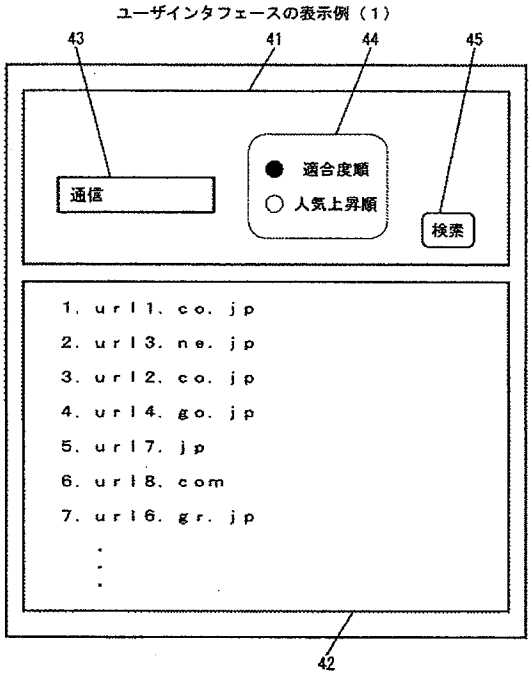
文書ID	記録年月日 2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	—	—	—	—
url2.co.jp	708	28	228	142
url3.ne.jp	949	336	954	706
url4.go.jp	582	556	218	151
url6.gr.jp	40	59	346	57
url7.jp	298	823	371	553
url8.com	773	904	357	888

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
—	—	—	—	—
324	244	997	826	679
277	337	851	358	220
908	245	367	193	185
699	777	72	943	546
833	846	455	378	454
390	286	415	241	608

[Drawing 12]

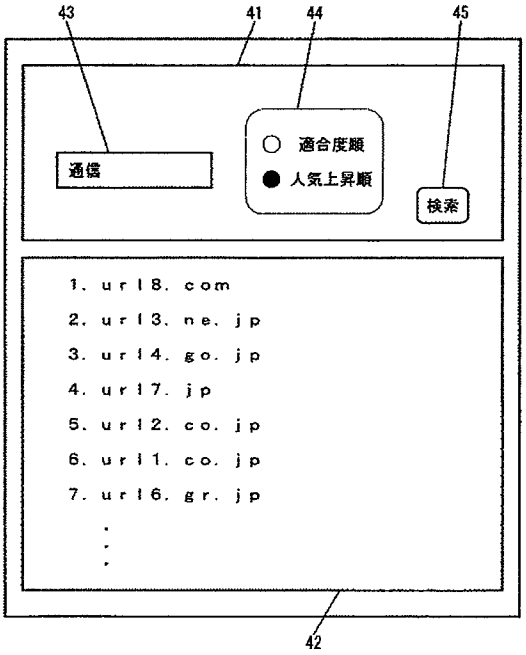


[Drawing 13]



[Drawing 14]

ユーザインタフェースの表示例（2）



[Translation done.]

(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-227165
(P2004-227165A)
(43)公開日　平成16年8月12日(2004.8.12)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	220C	5B075
	G06F 17/30	170A	

審査請求 未請求　請求項の数 12　OL　（全 16 頁）

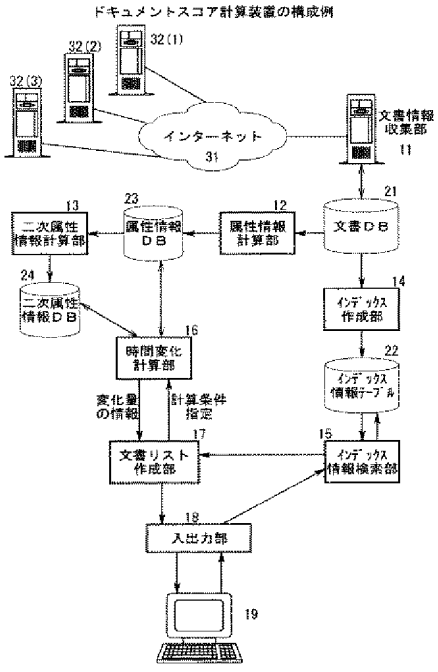
(21)出願番号　特願2003-12517(P2003-12517)
(22)出願日　平成15年1月21日(2003.1.21)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(74)代理人　100072718
弁理士　古谷　史旺
(72)発明者　大森　信行
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　井上　孝史
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　竹野　浩
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】ドキュメントスコア計算方法及び装置並びにプログラム

(57)【要約】
【課題】本発明は大量の文書を検索する場合にユーザが必要とする文書により近い結果を得ることが可能なドキュメントスコア計算方法及び装置並びにプログラムを提供することを目的とする。
【解決手段】更新される可能性のある複数の文書の情報を収集し、収集された文書群の中から特定の文書を抽出するために利用される得点を文書毎に割り当てるためのドキュメントスコア計算方法において、各文書からその特性を表す属性情報を抽出し、文書毎に２つ以上の時点で抽出された属性情報を同時に保持し、文書毎に保持されている複数の属性情報に基づいて属性情報の経時変化を計算し、求められた属性情報の経時変化を得点の計算に反映することを特徴とする。
【選択図】　図１



【特許請求の範囲】
【請求項1】
更新される可能性のある複数の文書の情報を収集し、収集された文書群の中から特定の文書を抽出するために利用される得点を文書毎に割り当てるためのドキュメントスコア計算方法において、
各文書からその特性を表す属性情報を抽出し、
文書毎に２つ以上の時点で抽出された属性情報を同時に保持し、
文書毎に保持されている複数の属性情報に基づいて、属性情報の経時変化を計算し、
求められた属性情報の経時変化を得点の計算に反映する
ことを特徴とするドキュメントスコア計算方法。
【請求項2】
請求項１のドキュメントスコア計算方法において、
少なくともキーワードを含む指定された検索条件に基づいて文書毎に適合度を算出し、
求められた適合度と前記属性情報の経時変化との両者を反映した得点を文書毎に算出し、
前記得点の大きさの順番に並べて複数の文書の情報を出力する
ことを特徴とするドキュメントスコア計算方法。
【請求項3】
請求項１又は請求項２のドキュメントスコア計算方法において、各文書から属性情報を抽出する場合には、
目的の文書自身に含まれている情報によって定まる一次属性情報と、目的の文書以外の他文書に含まれている情報のうち目的の文書と他文書との関連性を表す二次属性情報とをそれぞれ抽出する
ことを特徴とするドキュメントスコア計算方法。
【請求項4】
請求項１又は請求項２のドキュメントスコア計算方法において、文書情報の収集の終了もしくは収集した文書情報の保存を契機として属性情報の抽出処理を開始することを特徴とするドキュメントスコア計算方法。
【請求項5】
更新される可能性のある複数の文書の情報を収集し、収集された文書群の中から特定の文書を抽出するために利用される得点を文書毎に割り当てるためのドキュメントスコア計算装置において、
各文書からその特性を表す属性情報を抽出する属性情報抽出手段と、
文書毎に２つ以上の時点で抽出された属性情報を同時に保持する属性情報保持手段と、
前記属性情報保持手段に文書毎に保持されている複数の属性情報に基づいて、属性情報の経時変化を計算する属性変化検出手段と、
求められた属性情報の経時変化を得点の計算に反映する得点計算手段と
を設けたことを特徴とするドキュメントスコア計算装置。
【請求項6】
請求項５のドキュメントスコア計算装置において、前記得点計算手段は、少なくともキーワードを含む指定された検索条件に基づいて文書毎に適合度を算出し、求められた適合度と前記属性情報の経時変化との両者を反映した得点を文書毎に算出し、前記得点の大きさの順番に並べて複数の文書の情報を出力することを特徴とするドキュメントスコア計算装置。
【請求項7】
請求項５又は請求項６のドキュメントスコア計算装置において、前記属性情報抽出手段は、目的の文書自身に含まれている情報によって定まる一次属性情報と、目的の文書以外の他文書に含まれている情報のうち目的の文書と他文書との関連性を表す二次属性情報とをそれぞれ抽出することを特徴とするドキュメントスコア計算装置。
【請求項8】

請求項５又は請求項６のドキュメントスコア計算装置において、前記属性情報抽出手段は、文書情報の収集の終了もしくは収集した文書情報の保存を契機として属性情報の抽出処理を開始することを特徴とするドキュメントスコア計算装置。
【請求項9】
更新される可能性のある複数の文書の情報を収集し、収集された文書群の中から特定の文書を抽出するために利用される得点を文書毎に割り当てるためのコンピュータで実行可能なプログラムにおいて、
各文書からその特性を表す属性情報を抽出する属性情報抽出手順と、
文書毎に２つ以上の時点で抽出された属性情報を同時に保持する属性情報保持手順と、
前記属性情報保持手段に文書毎に保持されている複数の属性情報に基づいて、属性情報の経時変化を計算する属性変化検出手順と、
求められた属性情報の経時変化を得点の計算に反映する得点計算手順と
を設けたことを特徴とするプログラム。
【請求項10】
請求項９のプログラムにおいて、前記得点計算手順では、少なくともキーワードを含む指定された検索条件に基づいて文書毎に適合度を算出し、求められた適合度と前記属性情報の経時変化との両者を反映した得点を文書毎に算出し、前記得点の大きさの順番に並べて複数の文書の情報を出力することを特徴とするプログラム。
【請求項11】
請求項９又は請求項１０のプログラムにおいて、前記属性情報抽出手順では、目的の文書自身に含まれている情報によって定まる一次属性情報と、目的の文書以外の他文書に含まれている情報のうち目的の文書と他文書との関連性を表す二次属性情報とをそれぞれ抽出することを特徴とするプログラム。
【請求項12】
請求項９又は請求項１０のプログラムにおいて、前記属性情報抽出手順では、文書情報の収集の終了もしくは収集した文書情報の保存を契機として属性情報の抽出処理を開始することを特徴とするプログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、例えばインターネット上などに存在する多数の文書の中からユーザが参照すべき文書の候補を自動的に抽出するような場合に利用されるドキュメントスコア計算方法及び装置並びにプログラムに関する。
【０００２】
【従来の技術】
従来の一般的なテキスト検索システムの技術については非特許文献１に開示されている。また、Ｗｅｂから情報を収集して検索を行う従来のシステムについては非特許文献２に開示されている。
例えば、インターネットやＬＡＮのようなネットワーク上には様々なコンピュータが接続され、膨大な文書情報がアクセス可能な状態で存在している。これらの膨大な文書の中には目的とする文書が存在する可能性が高い。しかし、文書の数が膨大であるし、それぞれの文書が存在する場所もそれぞれ異なるので、これらの膨大な文書情報の中からユーザが必要とする特定の文書を探し出すのは難しい。
【０００３】
従来より、例えばインターネット上では目的の文書を検索するための検索サービスが提供されている。このような検索サービスを提供するシステム、すなわち検索エンジンにおいては、インターネット上で様々な場所に存在するＷＷＷ（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ）ページの文書情報を予め収集してデータベースに保持しておき、ユーザが入力したキーワードなどの検索条件と一致する文書をデータベースから抽出し、適合度の大きい順番で文書リストを一覧表示する。
【非特許文献１】

井上他：ＩｎｆｏＢｅｅ　テキスト情報検索技術，ＮＴＴ　Ｒ&Ｄ，ｖｏｌ．４６，Ｎｏ．１０，１９９７，ｐｐ．９３－９８
【非特許文献２】
ＭｃＢｒｙａｎ：ＧＥＮＶＬ　ａｎｄ　ＷＷＷＷ：　Ｔｏｏｌｓ　ｆｏｒ　Ｔａｍｉｎｇ　ｔｈｅ　Ｗｅｂ，　Ｐｒｏｃ．　ｏｆ　ｔｈｅ　ｆｉｒｓｔ　Ｉｎｔｅｒｎａｔｉｏｎａｌ　ＷＷＷ　ｃｏｎｆｅｒｅｎｃｅ，　１９９４
【０００４】
【発明が解決しようとする課題】
上記のような検索サービスを提供するシステムにおいては、例えば一定の間隔で周期的に文書情報の収集を行い、更新されてた文書情報については最新の文書情報に更新し、古い文書情報は廃棄する。そして、収集された最新の文章情報だけに基づいて検索を実行している。
【０００５】
しかしながら、従来の検索サービスを利用して目的の文書を検索する場合には、ユーザの必要としない文書が検索結果として抽出される場合も多いのが実情である。
すなわち、インターネット上には信頼性の高い文書，信頼性の低い文書，利用価値の高い文書，利用価値の低い文書など様々な文書が混在しているため、キーワードの適合性を識別するだけでは、信頼性の低い文書や利用価値の低い文書が検索結果として抽出されるのを避けることはできない。
【０００６】
本発明は、大量の文書を検索する場合にユーザが必要とする文書により近い結果を得ることが可能なドキュメントスコア計算方法及び装置並びにプログラムを提供することを目的とする。
【０００７】
【課題を解決するための手段】
請求項１は、更新される可能性のある複数の文書の情報を収集し、収集された文書群の中から特定の文書を抽出するために利用される得点を文書毎に割り当てるためのドキュメントスコア計算方法において、各文書からその特性を表す属性情報を抽出し、文書毎に２つ以上の時点で抽出された属性情報を同時に保持し、文書毎に保持されている複数の属性情報に基づいて、属性情報の経時変化を計算し、求められた属性情報の経時変化を得点の計算に反映することを特徴とする。
【０００８】
インターネット上などで目的の文書を検索する場合には、指定されたキーワードとの適合性だけでなく、各文書の特性、例えば文書サイズ，単語数，タグ数，リンク数，被リンク数などを反映することにより、信頼性の低い文書や利用価値の低い文書を排除し、検索結果をユーザが必要とする文書に近づけることが可能になる。
【０００９】
しかし、ある１時点の文書の特性を参照するだけでは、希望する検索結果が得られない場合も多い。例えば、過去のある時点までは利用価値の高い文書であったが、現在は利用価値が低くなったような文書も存在するので、この文書を検索結果として抽出するのは好ましくない。
請求項１においては、文書毎に保持されている複数時点の属性情報（文書サイズ，単語数，タグ数，リンク数，被リンク数など）に基づいて、属性情報の経時変化を計算するので、属性情報の経時変化を反映した検索結果を得ることができる。
【００１０】
例えば、第１の文書の利用価値が高い場合には、第２の文書から第１の文書にアクセスする（又は参照する）ためのリンクが形成される。このリンクは第２の文書上に存在し、第１の文書にとっては被リンクとみなすことができる。
第１の文書の利用価値が高い場合には、多数の文書が第１の文書に対してリンクを形成するので、第１の文書における被リンク数が多くなる。

【００１１】
また、第１の文書の利用価値が高い時には、第１の文書における被リンク数が増大する可能性が高く、第１の文書の利用価値が低くなった時には、第１の文書における被リンク数が減少する可能性が高い。
このため、ユーザが必要とする文書により近い検索結果を得ることができ、検索制度が向上する。
【００１２】
請求項２は、請求項１のドキュメントスコア計算方法において、少なくともキーワードを含む指定された検索条件に基づいて文書毎に適合度を算出し、求められた適合度と前記属性情報の経時変化との両者を反映した得点を文書毎に算出し、前記得点の大きさの順番に並べて複数の文書の情報を出力することを特徴とする。
請求項２においては、適合度と前記属性情報の経時変化との両者を反映した得点の大きさの順番に並べて複数の文書の情報を出力するので、ユーザが必要とする文書により近いと考えられる文書から順番に並んだ情報が得られる。
【００１３】
請求項３は、請求項１又は請求項２のドキュメントスコア計算方法において、各文書から属性情報を抽出する場合には、目的の文書自身に含まれている情報によって定まる一次属性情報と、目的の文書以外の他文書に含まれている情報のうち目的の文書と他文書との関連性を表す二次属性情報とをそれぞれ抽出することを特徴とする。
【００１４】
請求項３においては、一次属性情報及び二次属性情報を利用して各文書の得点を計算できるので、様々なユーザの検索条件に適した検索を実行できる。一次属性情報としては、文書サイズ，単語数，タグ数，リンク数などが考えられる。二次属性情報としては、被リンク数などが考えられる。
【００１５】
請求項４は、請求項１又は請求項２のドキュメントスコア計算方法において、文書情報の収集の終了もしくは収集した文書情報の保存を契機として属性情報の抽出処理を開始することを特徴とする。
請求項４においては、文書情報を収集する度に新たな属性情報を得ることができる。
【００１６】
請求項５は、更新される可能性のある複数の文書の情報を収集し、収集された文書群の中から特定の文書を抽出するために利用される得点を文書毎に割り当てるためのドキュメントスコア計算装置において、各文書からその特性を表す属性情報を抽出する属性情報抽出手段と、文書毎に２つ以上の時点で抽出された属性情報を同時に保持する属性情報保持手段と、前記属性情報保持手段に文書毎に保持されている複数の属性情報に基づいて、属性情報の経時変化を計算する属性変化検出手段と、求められた属性情報の経時変化を得点の計算に反映する得点計算手段とを設けたことを特徴とする。
【００１７】
請求項５においては、請求項１と同様の結果が得られる。
請求項６は、請求項５のドキュメントスコア計算装置において、前記得点計算手段は、少なくともキーワードを含む指定された検索条件に基づいて文書毎に適合度を算出し、求められた適合度と前記属性情報の経時変化との両者を反映した得点を文書毎に算出し、前記得点の大きさの順番に並べて複数の文書の情報を出力することを特徴とする。
【００１８】
請求項６においては、請求項２と同様の結果が得られる。
請求項７は、請求項５又は請求項６のドキュメントスコア計算装置において、前記属性情報抽出手段は、目的の文書自身に含まれている情報によって定まる一次属性情報と、目的の文書以外の他文書に含まれている情報のうち目的の文書と他文書との関連性を表す二次属性情報とをそれぞれ抽出することを特徴とする。
【００１９】

請求項７においては、請求項３と同様の結果が得られる。
請求項８は、請求項５又は請求項６のドキュメントスコア計算装置において、前記属性情報抽出手段は、文書情報の収集の終了もしくは収集した文書情報の保存を契機として属性情報の抽出処理を開始することを特徴とする。
請求項８においては、請求項４と同様の結果が得られる。
【００２０】
請求項９は、更新される可能性のある複数の文書の情報を収集し、収集された文書群の中から特定の文書を抽出するために利用される得点を文書毎に割り当てるためのコンピュータで実行可能なプログラムにおいて、各文書からその特性を表す属性情報を抽出する属性情報抽出手順と、文書毎に２つ以上の時点で抽出された属性情報を同時に保持する属性情報保持手順と、前記属性情報保持手段に文書毎に保持されている複数の属性情報に基づいて、属性情報の経時変化を計算する属性変化検出手順と、求められた属性情報の経時変化を得点の計算に反映する得点計算手順とを設けたことを特徴とする。
【００２１】
請求項９のプログラムを所定のコンピュータで実行することにより、請求項１と同様の結果が得られる。
請求項１０は、請求項９のプログラムにおいて、前記得点計算手順では、少なくともキーワードを含む指定された検索条件に基づいて文書毎に適合度を算出し、求められた適合度と前記属性情報の経時変化との両者を反映した得点を文書毎に算出し、前記得点の大きさの順番に並べて複数の文書の情報を出力することを特徴とする。
【００２２】
請求項１０のプログラムを所定のコンピュータで実行することにより、請求項２と同様の結果が得られる。
請求項１１は、請求項９又は請求項１０のプログラムにおいて、前記属性情報抽出手順では、目的の文書自身に含まれている情報によって定まる一次属性情報と、目的の文書以外の他文書に含まれている情報のうち目的の文書と他文書との関連性を表す二次属性情報とをそれぞれ抽出することを特徴とする。
【００２３】
請求項１１のプログラムを所定のコンピュータで実行することにより、請求項３と同様の結果が得られる。
請求項１２は、請求項９又は請求項１０のプログラムにおいて、前記属性情報抽出手順では、文書情報の収集の終了もしくは収集した文書情報の保存を契機として属性情報の抽出処理を開始することを特徴とする。
【００２４】
請求項１２のプログラムを所定のコンピュータで実行することにより、請求項４と同様の結果が得られる。
【００２５】
【発明の実施の形態】
本発明のドキュメントスコア計算方法及び装置並びにプログラムの１つの実施の形態について図１～図１４を参照して説明する。この形態は全ての請求項に対応する。
【００２６】
図１はドキュメントスコア計算装置の構成例を示すブロック図である。図２は時間変化計算処理の手順を示すフローチャートである。図３はドキュメントリスト作成処理の手順を示すフローチャートである。図４は属性情報ＤＢ（データベース：以下同様）の構成を示す模式図である。図５は二次属性情報ＤＢの構成を示す模式図である。
【００２７】
図６は文書サイズテーブルの構成例を示す模式図である。図７はタグ数テーブルの構成例を示す模式図である。図８はリンク数テーブルの構成例を示す模式図である。図９は更新日時テーブルの構成例を示す模式図である。図１０は被リンク数テーブルの構成例を示す模式図である。

図１１は文書間関連度テーブル（１）の構成例を示す模式図である。図１２はユーザインタフェースの構成例を示す正面図である。図１３はユーザインタフェースの表示例（１）を示す正面図である。図１４はユーザインタフェースの表示例（２）を示す正面図である。

【００２８】

この形態では、請求項５の属性情報抽出手段，属性情報保持手段，属性変化検出手段及び得点計算手段は、それぞれ属性情報計算部１２（二次属性情報計算部１３），属性情報ＤＢ２３（二次属性情報ＤＢ２４），時間変化計算部１６及び文書リスト作成部１７に相当する。

この形態では、インターネット上に存在する文書群を対象として目的の文書を検索する場合を想定している。勿論、例えばＬＡＮで接続された他のコンピュータ上に存在する文書群を検索対象にすることも可能である。

【００２９】

図１に示すドキュメントスコア計算装置は、文書情報収集部１１，属性情報計算部１２，二次属性情報計算部１３，インデックス作成部１４，インデックス情報検索部１５，時間変化計算部１６，文書リスト作成部１７，入出力部１８，ユーザ端末１９，文書ＤＢ２１，インデックス情報テーブル２２，属性情報ＤＢ２３及び二次属性情報ＤＢ２４を備えている。

【００３０】

文書情報収集部１１は、インターネット３１と接続されたコンピュータであり、インターネット３１に接続されている任意の文書サーバ３２（１），３２（２），３２（３），・・・からアクセス可能な全ての文書の情報を定期的に自動収集する。文書情報収集部１１が収集した文書の情報は、文書ＤＢ２１に保存される。一般的な検索サービスを提供するシステムにおいては、インターネット上の同じＵＲＬのサイトから新しい文書の情報を収集する度にデータベースの内容を更新し、最新の文書情報だけを保持するが、図１の文書ＤＢ２１は収集された文書情報を順次追加登録する。

【００３１】

従って、文書情報収集部１１が文書情報の収集を繰り返すと、文書ＤＢ２１上には同じサイトから複数の時点で（例えば一日おきに）それぞれ収集された文書情報が同時に存在することになる。なお、文書ＤＢ２１に空きがなくなった場合には、最も古い時点で収集された情報から順番に削除すればよい。

【００３２】

属性情報計算部１２及び二次属性情報計算部１３は、文書情報収集部１１が新たな文書情報の収集を完了する度に、あるいは収集された文書情報を文書ＤＢ２１に追加する度に、それを契機として属性情報の計算を行う。

属性情報とは、各文書情報の特性を表す情報である。この例では、文書の更新日時，文書内のリンク数，文書内のタグ数，文書サイズ及び文書の被リンク数を属性情報として取得する。

【００３３】

被リンク数とは、該当する文書を参照している他の文書の数である。また、文書の更新日時，文書内のリンク数，文書内のタグ数及び文書サイズは、該当する文書内の情報によって特定される。一方、被リンク数は該当する文書を参照している他の文書に埋め込まれたリンクなどの情報によって決定される。

そこで、この例では文書の更新日時，文書内のリンク数，文書内のタグ数及び文書サイズを一次属性情報として区分し、被リンク数のように他の文書の情報によって定まる情報を二次属性情報として区分している。

【００３４】

属性情報計算部１２は、新たに文書情報が収集され文書ＤＢ２１に蓄積される度に、その情報に基づいて文書毎に一次属性情報を計算する。属性情報計算部１２が計算を実施する度に、その計算結果、すなわち一次属性情報は属性情報ＤＢ２３に追加登録される。

従って、属性情報計算部１２が計算を繰り返すと、属性情報ＤＢ２３には、それぞれの文書について互いに異なる時点で収集された文書情報に関する一次属性情報が同時に保持される。

【００３５】

二次属性情報計算部１３は、属性情報計算部１２が計算を実施する度に、属性情報ＤＢ２３に記録された一次属性情報及び文書ＤＢ２１の内容に基づいて、二次属性情報を計算する。

例えば、１つの文書（第１の文書）の中に１つのリンク情報が含まれていることを検出する度に、そのリンク情報の参照先である第２の文書に関する被リンク数に１を加算すればよい。

【００３６】

二次属性情報計算部１３が計算を実施する度に、その結果、すなわち二次属性情報が二次属性情報ＤＢ２４に追加登録される。

従って、二次属性情報計算部１３が計算を繰り返すと、二次属性情報ＤＢ２４には、それぞれの文書について互いに異なる時点で収集された文書情報に関する二次属性情報が同時に保持される。

【００３７】

実際には、属性情報ＤＢ２３には図４に示すように更新日時テーブルＴ１１，リンク数テーブルＴ１２，タグ数テーブルＴ１３及び文書サイズテーブルＴ１４が設けられている。更新日時テーブルＴ１１，リンク数テーブルＴ１２，タグ数テーブルＴ１３及び文書サイズテーブルＴ１４は、それぞれ各文書の更新日時，リンク数，タグ数及び文書サイズを一次属性情報として保持している。

【００３８】

更新日時テーブルＴ１１，リンク数テーブルＴ１２，タグ数テーブルＴ１３及び文書サイズテーブルＴ１４の構成の具体例がそれぞれ図９，図８，図７及び図６に示されている。図６に示すように、文書サイズテーブルＴ１４には記録年月日毎に、各文書の文書サイズ、すなわち文書ファイルのバイト数が記録されている。また、各文書を特定するための文書ＩＤとしては、その所在を表すＵＲＬを用いている。

【００３９】

例えば、文書ＩＤが「ｕｒｌ１．ｃｏ．ｊｐ」の文書については、２００２年８月２９日に記録された文書サイズが８７１バイトであり、２００２年８月３０日に記録された文書サイズが７７３バイトであり、２００２年８月３１日に記録された文書サイズが１３６バイトである。この例では、１日おきにその時点の各文書の文書サイズを取得し、それを文書サイズテーブルＴ１４に追加登録している。

【００４０】

同様に、タグ数テーブルＴ１３には記録年月日毎に各文書に含まれているタグの数が記録され、リンク数テーブルＴ１２には記録年月日毎に各文書に含まれているリンクの数が記録され、更新日時テーブルＴ１１には記録年月日毎に各文書の更新日時が記録されている。

なお、更新日時テーブルＴ１１における各文書の更新日時は、ある基準日時（例えばデータ収集日時）に対する更新日時までの秒数を表している。例えば、基準日時が２００２年８月２９日の０時０分０秒である場合に、文書の更新日時が２００２年８月２９日の０時９分８秒であったとすると、記録される更新日時（更新日時－基準日時）は５６８秒になる。

【００４１】

一方、二次属性情報ＤＢ２４には図５に示すように被リンク数テーブルＴ２１及び複数の文書間関連度テーブルＴ２２１，Ｔ２２２，Ｔ２２３，・・・が設けられている。

被リンク数テーブルＴ２１の具体例は図１０に示されており、文書間関連度テーブル（１）Ｔ２２１の具体例は図１１に示されている。

【００４２】

図１０に示すように、被リンク数テーブルＴ２１には記録年月日毎に、各文書に対する他の文書からの被リンク数が記録されている。また、各文書を特定するための文書ＩＤとしては、その所在を表すＵＲＬを用いている。
各文書間関連度テーブルＴ２２１，Ｔ２２２，Ｔ２２３，・・・は、それぞれの文書について、他の文書との関連度を表す情報を保持している。例えば、図１１に示す文書間関連度テーブル（１）Ｔ２２１は１番目の文書（文書ＩＤ：ｕｒｌ１．ｃｏ．ｊｐ）と他の各文書（文書ＩＤ：ｕｒｌ２．ｃｏ．ｊｐ：ｕｒｌ３．ｎｅ．ｊｐ：ｕｒｌ４，ｇｏ．ｊｐ・・・）との関連度を表す情報をそれぞれ保持している。

【００４３】
同様に、文書間関連度テーブルＴ２２２は２番目の文書（文書ＩＤ：ｕｒｌ２．ｃｏ．ｊｐ）と他の各文書との関連度を表す情報をそれぞれ保持し、文書間関連度テーブルＴ２２３は３番目の文書（文書ＩＤ：ｕｒｌ３．ｎｅ．ｊｐ）と他の各文書との関連度を表す情報をそれぞれ保持している。

【００４４】
時間変化計算部１６は、属性情報ＤＢ２３に保持されている複数時点の一次属性情報及び二次属性情報ＤＢ２４に保持されている複数時点の二次属性情報に基づいて、各一次属性情報の経時変化及び二次属性情報の経時変化を文書毎に計算する。
例えば、同じ文書について２つの時点で記録された２つの属性情報の差分を計算することにより、２つの時点の間における属性の経時変化を求めることができる。

【００４５】
実際には、インターネット上のＨＴＭＬ文書を検索対象とする場合が多い。このような文書を処理する場合には、文書サイズや更新日時はその文書自体から取得することができる。また、タグ数やリンク数については、文書のテキストを構文解析することにより取得できる。また、各文書について形態素解析処理を行えば各文書の単語数を属性情報として取得することもできる。

【００４６】
なお、図６～図１１の例では各テーブルにおける属性情報の取得時刻を記録年月日として表しているが、日付又は時刻あるいはある時点からの経過日時として表しても良い。単位についても、秒，分，時などの数値を用いることができる。また、図６～図１１の例では１日１回、２４時間周期で定期的に情報を収集した場合を想定しているが、周期の長さについては任意に定めることができる。また、必ずしも一定の周期で情報を収集する必要はない。

【００４７】
図１の時間変化計算部１６によって実行される時間変化計算処理の内容について、図２を参照しながら説明する。
ステップＳ１１では、計算条件を決定する情報を取得する。この情報には、計算対象文書の文書ＩＤ，対象属性，対象期間の開始時刻，対象期間の終了時刻及び計算方法が含まれる。

【００４８】
これらの情報は、ユーザ端末１９を操作するオペレータからの入力によって特定される。例えば、属性がページ間（文書間）の関連度の場合には、関連度の計算対象となる文書の文書ＩＤ（ＵＲＬ）を入力する必要がある。
ステップＳ１２では、ステップＳ１１で取得した計算条件の対象属性に該当するテーブルを属性情報ＤＢ２３上のテーブル群（図４参照）又は二次属性情報ＤＢ２４上のテーブル群（図５参照）から選択する。

【００４９】
ステップＳ１３では、Ｓ１２で選択されたテーブル上で、Ｓ１１で取得した計算条件の文書ＩＤと一致する行を計算対象として選択する。
ステップＳ１４では、Ｓ１２で選択されたテーブル上のＳ１３で選択された行から、計算条件として指定された対象期間の開始時刻と終了時刻とに含まれる全ての列を選択し、そ

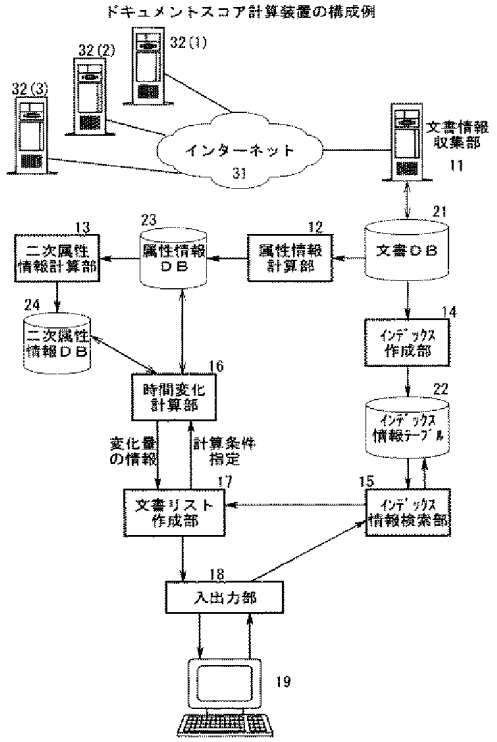
れらの属性値を記録年月日（収集日時）とともに取得する。
【００５０】
例えば、計算条件として
対象属性：被リンク数
対象期間の開始時刻：２００２．０８．２９
対象期間の終了時刻：２００２．０９．０２
が指定された場合には、図１０に示される被リンク数テーブルから次のような５つの時点の情報が抽出される。
【００５１】
２００２．０８．２９：３
２００２．０８．３０：１
２００２．０８．３１：４
２００２．０９．０１：１
２００２．０９．０２：９
ステップＳ１５では、Ｓ１４でテーブルから取得した属性値に基づき、指定された計算方法で演算を実行する。
【００５２】
例えば、計算条件として
対象属性：被リンク数
対象期間の開始時刻：２００２．０８．２９
対象期間の終了時刻：２００２．０９．０２
計算方法：属性値の増加数
が指定された場合には、
開始時刻　２００２．０８．２９：属性値　３
終了時刻　２００２．０９．０２：属性値　９
から（９－３）＝６が変化量として求められる。
【００５３】
また、例えば、計算条件として
対象属性：被リンク数
対象期間の開始時刻：２００２．０８．２９
対象期間の終了時刻：２００２．０９．０２
計算方法：属性値の平均
が指定された場合には、
２００２．０８．２９：３
２００２．０８．３０：１
２００２．０８．３１：４
２００２．０９．０１：１
２００２．０９．０２：９
から（（３＋１＋４＋１＋９）／５）＝３．６が変化量として求められる。
【００５４】
図１に示す装置が検索サービスを提供するために設けられたインデックス作成部１４，インデックス情報テーブル２２，インデックス情報検索部１５，文書リスト作成部１７及び入出力部１８については、基本的な動作は非特許文献１に記載された従来技術と同様である。
すなわち、インデックス作成部１４は文書ＤＢ２１に保存されている文書群のデータに基づいて検索に必要なインデックス情報を作成する。このインデックス情報はインデックス情報テーブル２２に保存される。
【００５５】
入出力部１８は、ユーザ端末１９から入力される検索条件、例えば検索すべきキーワードなどをインデックス情報検索部１５に送信する。インデックス情報検索部１５は、入出力

部１８から受け取った検索条件に一致する文書の情報をインデックス情報テーブル２２から取り出して文書リスト作成部１７に送信する。
文書リスト作成部１７は、指定された検索条件に一致した文書群の情報を適合度の順番に並べ替えて一覧として入出力部１８に送出する。この結果がユーザ端末１９の画面上に表示される。
【００５６】
但し、図１に示す装置においては更に次に示すような特徴的な動作を行う。すなわち、検索エンジンの検索精度を改善するために、属性値の時間変化を反映した結果を出力する。実際には、文書リスト作成部１７が、検索条件に適合した文書群を得点の順番に並べ替える。一般的には得点として適合度を利用するが、図１の文書リスト作成部１７は、適合度と属性値の時間変化値の両方を反映した得点を文書毎に算出し、この得点を利用して文書群の情報を並べ替える。
【００５７】
具体的な文書リスト作成部１７の動作は図３に示すとおりである。
ステップＳ２１では、ユーザ端末１９からの入力によって指定された検索条件と一致する文書群の情報を抽出する。実際には、文書リスト作成部１７はインデックス情報検索部１５から検索結果である文書群の情報を受け取る。これらの情報には、各文書を識別するための文書ＩＤであるＵＲＬや、指定された検索条件との適合度が含まれている。
【００５８】
ステップＳ２２では、検索条件と一致する文書群の中で未処理のものの中から１つの文書情報を選択する。
ステップＳ２３では、選択した文書の文書ＩＤ並びに計算条件、すなわち対象属性，対象期間の開始時刻，対象期間の終了時刻及び計算方法を時間変化計算部１６に与える。
【００５９】
なお、この計算条件については予め定められた条件を適用することができる。また、１つの検索結果に関しては同じ計算条件が用いられる。但し、実際の検索条件は時間の経過に伴って変化する可能性がある。
例えば、被リンク数、すなわち他の文書からリンクされている場合の該当する他の文書の数は、その文書の人気度合いを表す指標として利用できる。従って、例えば最近１週間で被リンク数が上昇している文書を検索することには大きな意義がある。この場合には、開始時刻の現時点から１週間前であり、終了時刻は現時点になるのでその条件は日々変化する。
【００６０】
また、対象となる属性や時間変化を計算する期間を検索条件に応じて適応的に変更することも可能である。また、ユーザ端末１９を操作するオペレータの指示に応じて計算条件を変更することもできる。
計算条件を与えることにより、時間変化計算部１６は前述のような処理を実行し、文書毎に変化量を算出する。文書リスト作成部１７は、時間変化計算部１６が出力する文書毎の変化量を受け取り保存する。
【００６１】
未処理文書が無くなると、文書リスト作成部１７の動作はステップＳ２４からＳ２５に進む。
ステップＳ２５では、時間変化計算部１６から受け取った文書毎の時間変化量の値と、インデックス情報検索部１５から受け取った適合度の値との両方を予め定めた計算式に従って計算し、対象となる文書毎にその得点を求める。
【００６２】
ステップＳ２５で利用する計算式については、２つの値ｘ（時間変化量），ｙ（適合度）を用いるので、２つの値ｘ，ｙの重み付き代数和（ｆ（ｘ，ｙ）＝ａｘ＋ｂｙ）や２つの値の積などを用いることが考えられる。
ステップＳ２６では、インデックス情報検索部１５の検索結果である文書群の情報を得点

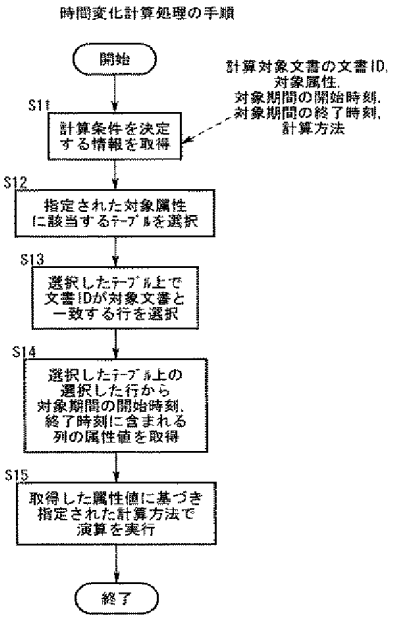
の大きい順番で並べ替えた結果をドキュメントリストとして作成し出力する。
【００６３】
従って、ユーザが指定したキーワードなどとの適合性だけでなく、被リンク数のような属性値の時間変化を反映する形で検索結果を出力することができる。このため、検索精度が向上する。
図１に示す装置を実現する場合には、ユーザ端末１９上に例えば図１２，図１３及び図１４に示すようなユーザインタフェースを設けるのが望ましい。このユーザインタフェースは、図１２～図１４に示すよう状態でユーザ端末１９の画面上に表示される。
【００６４】
このユーザインタフェースには、図１２に示すように検索条件指定部４１及び検索結果表示部４２が設けてある。また、検索条件指定部４１には、キーワード入力部４３，検索結果表示順序指定部４４及び検索実行ボタン４５が設けてある。
キーワード入力部４３を操作することにより、検索対象となる任意のキーワードを入力することができる。また、この例では検索結果表示順序指定部４４を操作することにより、２種類の表示順序、すなわち「適合度順」及び「人気上昇順」を選択することができる。
【００６５】
検索結果表示部４２には、キーワード入力部４３で指定された検索条件と一致する文書群の各文書ＩＤ（ＵＲＬ）が、検索結果表示順序指定部４４の指定に従って並べられて一覧表示される。
図１３に示す例では、キーワード入力部４３で「通信」のキーワードを指定し、検索結果表示順序指定部４４で「適合度順」を指定した場合を想定している。従って、検索結果表示部４２には検索結果の文書群が適合度順に並べられた状態で一覧表示されている。
【００６６】
図１４に示す例では、キーワード入力部４３で「通信」のキーワードを指定し、検索結果表示順序指定部４４で「人気上昇順」を指定した場合を想定している。従って、検索結果表示部４２には検索結果の文書群が人気上昇順に並べられた状態で一覧表示されている。
【００６７】
図１４に示すように人気上昇順で結果を出力する場合には、時間変化計算部１６が計算する属性として被リンク数を選択し、例えば最近１週間の被リンク数の増加量を反映した得点を計算すれば良い。
この場合、計算条件として例えば次のような情報を与えればよい。
対象属性：被リンク数
開始時刻：２００２．０８．３０
終了時刻：２００２．０９．０６
計算方法：属性値の増分
そして、図１３と図１４との違いから分かるように、同じキーワードを用いて検索を行う場合であっても、検索結果表示順序指定部４４の指定に応じて異なる結果が得られる。
【００６８】
なお、図１に示すような装置については、専用のハードウェアで実現することもできるし、コンピュータ上でプログラムを実行して実現することもできる。
【００６９】
【発明の効果】
本発明によれば、複数時点の属性情報を用いて属性情報の変化を調べ、それを検索結果に反映することができるので、検索精度の改善に効果がある。
【図面の簡単な説明】
【図１】ドキュメントスコア計算装置の構成例を示すブロック図である。
【図２】時間変化計算処理の手順を示すフローチャートである。
【図３】ドキュメントリスト作成処理の手順を示すフローチャートである。
【図４】属性情報ＤＢの構成を示す模式図である。
【図５】二次属性情報ＤＢの構成を示す模式図である。

【図６】文書サイズテーブルの構成例を示す模式図である。
【図７】タグ数テーブルの構成例を示す模式図である。
【図８】リンク数テーブルの構成例を示す模式図である。
【図９】更新日時テーブルの構成例を示す模式図である。
【図１０】被リンク数テーブルの構成例を示す模式図である。
【図１１】文書間関連度テーブル（１）の構成例を示す模式図である。
【図１２】ユーザインタフェースの構成例を示す正面図である。
【図１３】ユーザインタフェースの表示例（１）を示す正面図である。
【図１４】ユーザインタフェースの表示例（２）を示す正面図である。
【符号の説明】
１１　文書情報収集部
１２　属性情報計算部
１３　二次属性情報計算部
１４　インデックス作成部
１５　インデックス情報検索部
１６　時間変化計算部
１７　文書リスト作成部
１８　入出力部
１９　ユーザ端末
２１　文書ＤＢ
２２　インデックス情報テーブル
２３　属性情報ＤＢ
２４　二次属性情報ＤＢ
３１　インターネット
３２　文書サーバ

【図1】

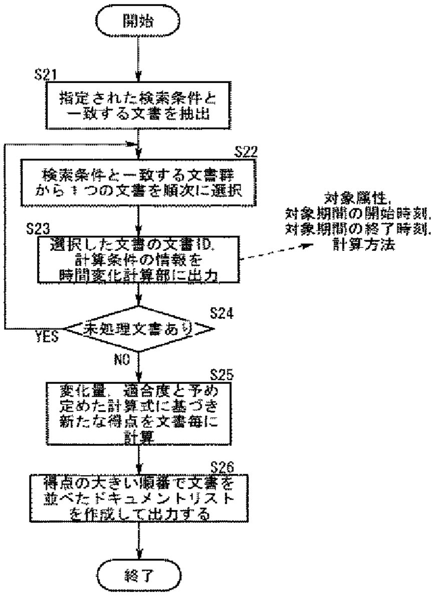


【図2】



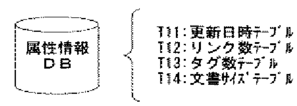
【図3】

ドキュメントリスト作成処理の手順



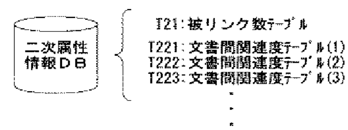
【図4】

属性情報ＤＢの構成



【図5】

二次属性情報ＤＢの構成



【図6】

文書サイズテーブル

文書ID	記録年月日								
	2002.08.29	2002.08.30	2002.08.31	2002.09.01	2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
url1.co.jp	871	773	136	34	17	411	55	263	853
url2.co.jp	874	732	629	719	300	138	215	403	874
url3.ne.jp	333	633	890	777	95	298	276	536	247
url4.go.jp	978	249	92	174	751	942	395	7	309
url6.gr.jp	96	814	259	421	118	839	738	491	884
url7.jp	190	966	115	718	134	106	20	62	852
url8.com	216	51	801	185	382	887	875	503	578

【図7】

タグ数テーブル

文書ID	記録年月日								
	2002.08.29	2002.08.30	2002.08.31	2002.09.01	2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
url1.co.jp	32	88	3	31	56	68	82	93	97
url2.co.jp	53	18	9	28	72	42	54	71	67
url3.ne.jp	25	23	28	16	26	27	29	29	37
url4.go.jp	・	・	・	・	・	・	・	・	・
url6.gr.jp	・	・	・	・	・	・	・	・	・
url7.jp	・	・	・	・	・	・	・	・	・
url8.com									

【図8】

リンク数テーブル

文書ID	記録年月日								
	2002.08.29	2002.08.30	2002.08.31	2002.09.01	2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
url1.co.jp	0	1	1	1	2	2	1	0	3
url2.co.jp	3	8	2	2	7	4	4	1	3
url3.ne.jp	5	2	2	1	2	2	2	2	3
url4.go.jp	・	・	・	・	・	・	・	・	・
url6.gr.jp	・	・	・	・	・	・	・	・	・
url7.jp	・	・	・	・	・	・	・	・	・
url8.com									

【図9】

更新日時テーブル

文書ID	記録年月日								
	2002.08.29	2002.08.30	2002.08.31	2002.09.01	2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
url1.co.jp	568	568	568	263040	356521	356521	356521	356521	356521
url2.co.jp	・	・	・	・	・	・	・	・	・
url3.ne.jp	・	・	・	・	・	・	・	・	・
url4.go.jp	・	・	・	・	・	・	・	・	・
url6.gr.jp	・	・	・	・	・	・	・	・	・
url7.jp	・	・	・	・	・	・	・	・	・
url8.com									

【図10】

被リンク数ﾃｰﾌﾞﾙ

文書ID	記録年月日			
	2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	3	1	4	1
url2.co.jp	3	9	2	0
url3.ne.jp	5	3	8	8
url4.go.jp	.	.	.	.
url6.gr.jp	.	.	.	.
url7.jp	.	.	.	.
url8.com				

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
9	6	8	3	0
5	7	6	3	1
6	7	9	9	7
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

【図11】

文書間関連度ﾃｰﾌﾞﾙ(1)

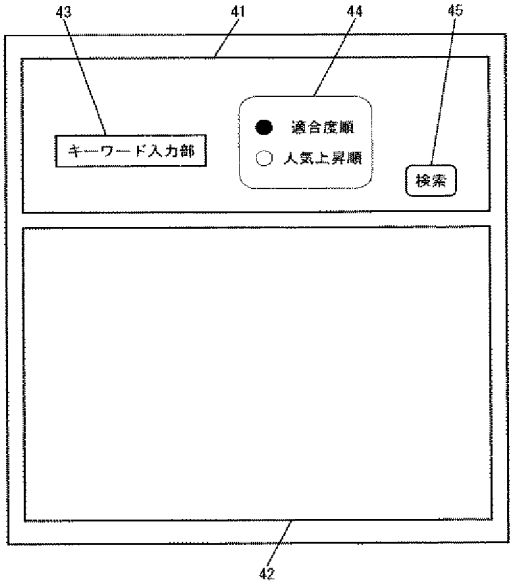
関連度計算対象：url1.co.jp

文書ID	記録年月日			
	2002.08.29	2002.08.30	2002.08.31	2002.09.01
url1.co.jp	—	—	—	—
url2.co.jp	708	28	228	142
url3.ne.jp	949	336	954	706
url4.go.jp	582	556	218	151
url6.gr.jp	40	59	346	57
url7.jp	298	823	371	553
url8.com	773	904	357	888

2002.09.02	2002.09.03	2002.09.04	2002.09.05	2002.09.06
—	—	—	—	—
324	244	997	826	679
277	337	951	358	220
908	245	367	193	185
699	777	72	943	546
633	846	455	378	454
390	286	415	241	608

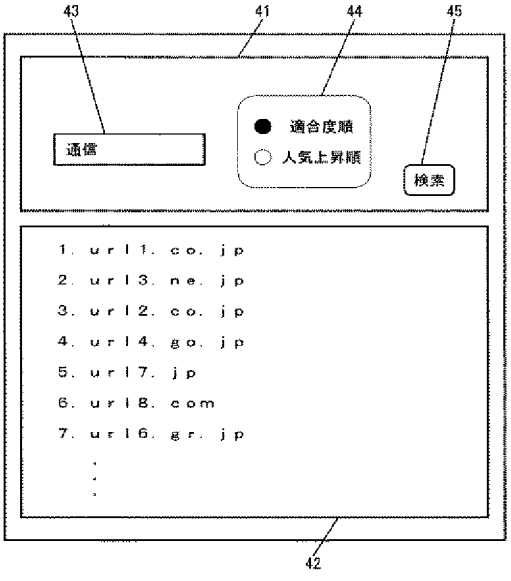
【図12】

ユーザインタフェースの構成例



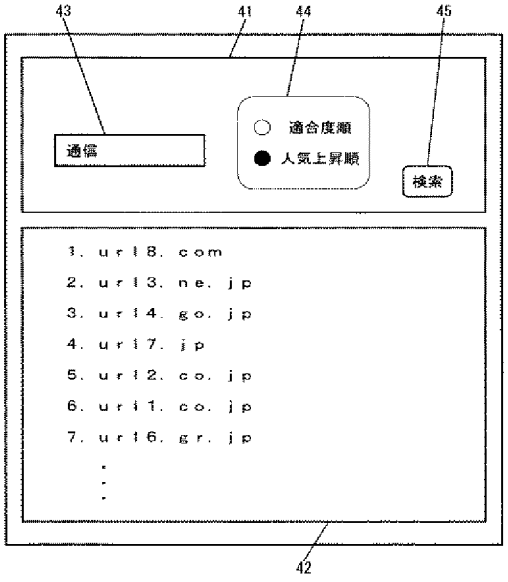
【図13】

ユーザインタフェースの表示例（１）



【図14】

ユーザインタフェースの表示例（２）



(72)発明者　茨木　久
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
Ｆターム(参考) 5B075 ND03　NK02　NK10　NS10　PP28　PQ74　PR03　QP05　QS01　UU06

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 August 2001 (30.08.2001)

PCT

(10) International Publication Number
WO 01/63479 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/US01/40173

(22) International Filing Date: 22 February 2001 (22.02.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/183,971 22 February 2000 (22.02.2000) US
60/201,839 3 May 2000 (03.05.2000) US

(71) Applicant: **METACARTA, INC.** [US/US]; P.O. Box 397207, Cambridge, MA 02139 (US).

(72) Inventors: **FRANK, John, R.**; P.O. Box 397207, Cambridge, MA 02139 (US). **RAUCH, Erik, M.**; 40 Circle Drive, Park Ridge, NJ 07656-1103 (US). **DONOGHUE, Karen**; 122 Lake Street, Arlington, MA 02474 (US).

(74) Agent: **PRAHL, Eric, L.**; Fish & Richardson P.C., 225 Franklin Street, Boston, MA 02110-2804 (US).

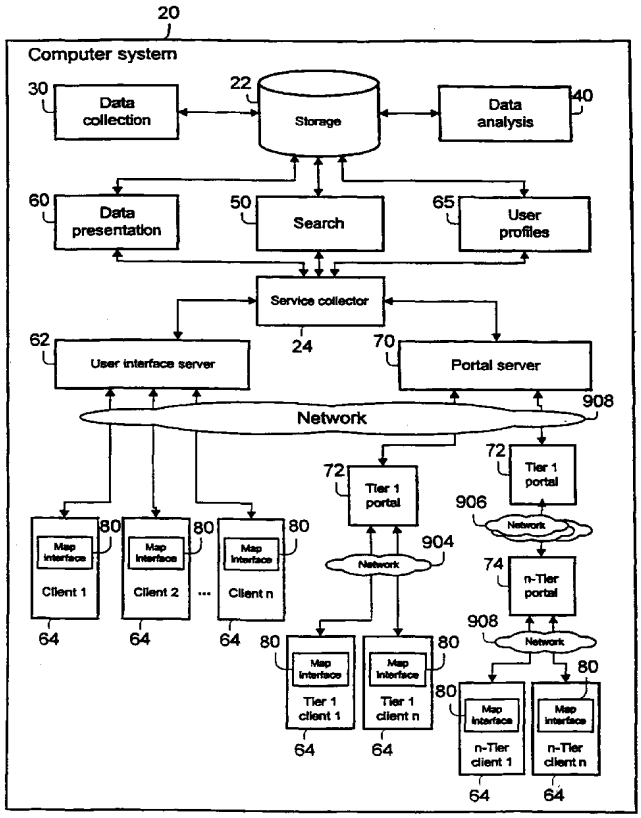
(81) Designated States *(national)*: CA, JP.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SPATIALLY CODING AND DISPLAYING INFORMATION



(57) Abstract: A computer system (20) includes storage system (22) which contains information in the form of documents, along with a spatial information about the documents. The computer system (20) also includes subsystems for data collection (30), data analysis (40), search (50), data presentation (60), and portal services (70). The computer system (20) further includes a map interface (80). Through the map interface (80), users can query the storage (22) and view a representation of the query results arranged on a map.

WO 01/63479 A1

Spatially Coding and Displaying Information

Under 35 U.S.C. §119(e)(1), this application claims benefit of prior U.S. Provisional Applications No. 60/183,971, entitled "Metacarta: Map-based Information Search Engine and Catalog," filed February 22, 2000; and No. 60/201,839, entitled "Method and System for Associating Information with Physical Objects and Locations; and Methods of Expanding a Database," filed May 3, 2000, both of which are incorporated herein by reference.

TECHNICAL FIELD

This invention relates to computer systems, and more particularly to spatial databases, document databases, search engines, and data visualization.

BACKGROUND

There are many tools available for organizing and accessing documents through different interfaces that help users find information. Some of these tools allow users to search for documents matching specific criteria, such as containing specified keywords. Some of these tools present information about geographic regions or spatial domains, such as driving directions presented on a map.

These tools are available on private computer systems and are sometimes made available over public networks, such as the Internet. Users can use these tools to gather information.

SUMMARY OF THE INVENTION

In a computer system that presents a map interface to a user, the invention enables a user, among other things, to pose a query via the map interface and to be able to inspect a representation of the query results arranged on the map as icons. The map and the icons are responsive to further user actions, including changes to the scope of the map, changes to the terms of the query, or closer examination of a subset of the results.

The targets of the query are documents. Examples of documents include text-based computer files, as well as files that are partially text-based, files containing spatial information, and computer entities that can be accessed via a document-like interface.

Documents can contain other documents and may have other interfaces besides their document-like interfaces. Every document has an address. In the case of world wide web documents, this address is commonly a URL.

The documents exist on computer systems arrayed across a computer network, such as a private network or the Internet. The documents may be hyperlinked, that is, may contain references (hyperlinks) to an address of another document. Copies of the documents may be stored in the page repository.

A spatial recognizer process examines documents for spatial information content. When the spatial recognizer determines that a document has spatial information content, the document is added to a spatial document collection.

A document ranking process assigns a spatial relevance score to each document in the spatial document collection. The spatial relevance score is a measure of the degree to which the document relates to the spatial location mentioned in its spatial information content. In cases where the document has more than one instance of spatial information content, the document is scored against each instance.

The spatial-keyword document indexer examines each document in the spatial document collection and represents it in an spatial-keyword document index data structure. The spatial-keyword document indexer indexes a document both by keywords and by at least one instance of spatial information content. The spatial-keyword document index enables unusually fast responses by the computer system to queries that combine spatial criteria with keyword criteria.

The crawler extends the collection of known documents by examining the hyperlinks contained in the known documents. When a hyperlink references a previously unknown document, the crawler adds the unknown document to the collection of known documents and examines them, in turn, for new hyperlinks to follow.

The crawler may prioritize the hyperlinks it follows based in part on spatial relevance scores.

The computer system includes a metasearcher process for initializing the collection of known documents. This initializing step is known as bootstrapping and is known in the art. The metasearcher queries predetermined search engines known to store information about other computer systems and document sources, such as search engine web sites on the

Internet. The human administrators of the metasearcher provide it with a collection of known spatial locations. The metasearcher formulates queries based on these spatial locations and directs the queries to the search engines. After each query, the results are compared to the collection of known documents and are added if new.

However, it is common for search engines to cap the maximum number of results returnable to a single query. The metasearcher is able to respond to a results cap by issuing follow-on queries which are progressively more spatially focused. An example of a progressively more spatially focused series might be "New York state," "New York, NY," "Times Square, New York, NY," etc. By progressively narrowing the scope of its queries, the metasearcher reduces the number of results until the results number fits within the cap. The progressive spatial focus produces information more closely matched to a specific spatial location, as well as a more exhaustive sample of the results available from a given search engine. At the same time, the generality of early queries casts a net as broad as possible, so as not to miss any results. As a result, the documents found by the metasearcher form a diverse yet highly spatially-qualified sample for the crawler to start from.

In general, in one aspect, the invention is an interface program stored on a computer-readable medium for causing a computer system with a display device to perform a set of functions. The functions are accepting search criteria from a user including a free text entry query and a domain identifier identifying a domain; in response to accepting the search criteria from the user, retrieving a plurality of record identifiers each of which identifies a corresponding record which: (1) has associated therewith a location identifier that locates it at a specific location within the domain identified by the domain identifier; and (2) contains information that is responsive to the free text entry query; displaying a representation of the domain on the display device; and displaying on the display device a plurality of icons as representations of the records identified by the plurality of record identifiers. For each of the plurality of record identifiers, a corresponding one of the plurality of icons is displayed within the representation of the domain that is being displayed on the display device. The corresponding icon for each of the plurality of record identifiers is positioned within the representation of the domain at a coordinate within the domain that corresponds to the location identifier for the corresponding record.

Preferred embodiments include one or more of the following features. The domain is a geographical region and the representation is a multi-dimensional map of the geographical region. More specifically, the representation is a two-dimensional map of the geographical region. The step of accepting input further includes accepting a designation by the user of a designated category, wherein each of the records corresponding to the plurality of retrieved record identifiers also includes information that falls within the designated category. The step of accepting the designation by the user of a category includes presenting to the user a list of predefined categories and accepting as the designated category a selection by the user from that list. The interface program also is for causing the computer to perform the further functions of, after displaying the corresponding icon for each of the plurality of record identifiers, accepting further search criteria from the user. The further search criteria are selected from the group of search criteria types consisting of a domain identifier input type, a free text entry query input type, and a category type. It also causes the computer, in response to accepting the further search criteria from the user, to perform the functions of: (1) retrieving a subset of the plurality of record identifiers, wherein the subset of the plurality of record identifiers identifies all record identifiers among the plurality of record identifiers that fall within the further search criteria; (2) displaying a two-dimensional map of a revised geographical region on the display device that is responsive to the further search criteria; (3) for each of the record identifiers of the subset of plurality of record identifiers, displaying a corresponding icon within the displayed map, wherein the corresponding icon for each of the record identifiers of the subset of the plurality of record identifiers is positioned within the displayed map at a coordinate that corresponds to the location identifier for the corresponding record; and (4) storing as a filter the first-mentioned search criteria in combination with the further search criteria, wherein the stored filter is retrievable for use by the user in specifying a future search through the interface. The first-mentioned search criteria in combination with the further search criteria is an ordered sequence of inputs and the stored filter is the ordered sequence of inputs preserving the order of the sequence of inputs.

Preferred embodiments may also include one or more of the following features. The interface program also causes the computer to perform the further functions of: presenting to the user via the display device a map; and enabling the user to input the domain identifier as

part of the search criteria by interacting with the displayed map. The plurality of icons include an icon of a first icon class and an icon of a second icon class; and the icon of the first icon class has first visual characteristics and the icon of the second icon class has second visual characteristics that are different from the visual characteristics associated with the first icon class. At least some of the records identified by the plurality of record identifiers are of a first type and at least some of the other records identified by the plurality of record identifiers are of a second type and the records of the first type are displayed using the icon of the first icon class and records of the second type are displayed using the icon of the second icon class. At least one of the icons of the plurality of icons represents multiple of the records identified by the plurality of record identifiers, wherein each of the multiple of the records having a location identifier that locates that record within a neighborhood about a central location.

Also in preferred embodiments, the interface program causes the computer to perform the further functions of: accepting a change of scale request from the user; in response to accepting the change of scale request, consolidating at least some of the plurality of icons with each other to form a second plurality of icons that is fewer in number than the number of icons in the first-mentioned plurality of icons; and in response to accepting the change of scale request, redisplaying the domain using a decreased scale and also using the second plurality of icons to identify the locations of the records identified by the plurality of record identifiers. In addition, it also causes the computer to perform the further functions of: accepting from the user a specification of an electronic note which has an associated location within the map; and displaying a sticky-note icon on the map at position that corresponds to the associated location. The electronic note comprises a web page with its own externally accessible address which enables people to electronically access its contents through that address.

In general, in another aspect, the invention is a method that performs the functions described above.

In general, in yet another aspect, the invention is a database system stored on a computer-readable medium for causing a computer system to perform the functions of: accepting search criteria including at least one of: (1) text; (2) a domain identifier identifying a domain; and (3) a filter identifier identifying a filter; and retrieving a plurality of record

identifiers each of which identifies a corresponding record which has associated therewith the text, domain identifier, or layer identifier of the search criteria, where the retrieving is performed with a spatial-keyword document index.

In general, in still another aspect, the invention is a method of retrieving a plurality of record identifiers each of which identifies a corresponding record which has associated therewith at least one of a text, a domain identifier, and a layer identifier as specified by search criteria, wherein such retrieving is performed with a spatial-keyword document index.

Preferred embodiments include one or more of the following features. The spatial-keyword document index includes a spatial index tree extended to reference documents and a plurality of trees with the same structure as the spatial index tree but trimmed for particular lexicon entries and filters. The plurality of record identifiers are retrieved in a spatial-keyword document index tree and the branching structure of the tree is analyzed to identify geographic phenomena. A geographic phenomena is a spatial-keyword document index tree branching structure in which more than a predetermined fraction of the branches share more than predetermined number of parent nodes.

In general, in still yet another aspect, the invention is a program stored on a computer-readable medium for causing a computer system to perform the functions of: loading documents referred to by document addresses; parsing those documents for more document addresses to load; and parsing those documents for possible spatial identifiers.

Preferred embodiments may include one or more of the following features. The program also causes the computer system to perform the further function of analyzing the possible spatial identifiers to determine a location in a domain. A portion of the document addresses are gathered by a metasearcher process, which queries other computer systems using text that refers to spatial domains. The program also causes the computer system to perform the further function of computing a relevance score for each of the possible spatial identifiers found in each document. The relevance score includes one or more of: (1) the possible spatial identifier's position in the document; (2) the number of other possible spatial identifiers in the document; (3) whether the possible spatial identifier is in a sentence or is free standing; and (4) the formatted emphasis of the characters in the possible spatial identifier. The program also causes the computer system to perform the further function of sorting the document addresses by relevance score before loading.

In general in still another aspect, the invention is a program stored on a computer-readable medium for causing a computer system to perform the functions of: loading documents referred to by document addresses; parsing those documents for more document addresses to load; parsing those documents for possible spatial identifiers; and analyzing the possible spatial identifiers to determine a location in a domain.

The details of one or more embodiments of the invention are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

DESCRIPTION OF DRAWINGS

FIG. 1 schematically shows an overall arrangement of a computer system according to an embodiment of the invention;

FIG. 2 schematically represents an arrangement of controls on a map interface according to an embodiment of the invention;

FIG. 3 is an explanatory diagram of storage entities and entities in a data collection process;

FIG. 4. is an explanatory diagram of entities in a data analysis process;

FIG. 5. is an explanatory diagram of entities in a search process;

FIG. 6. is an explanatory diagram of steps in a process for building a spatial-keyword indexer; and

FIG. 7. is an explanatory diagram of steps in a spatial indexer process.

DETAILED DESCRIPTION

In general, with reference to Fig. 1, the computer system 20 includes a storage 22 system which contains information in the form of documents, along with spatial information about the documents. The computer system 20 also includes subsystems for data collection 30, data analysis 40, search 50, data presentation 60, and portal services 70. The computer system 20 further includes a map interface 80 presented to a user through a variety of clients. Through the map interface 80, the user can query the storage 22 and can view a representation of the query results arranged on a map.

DOCUMENTS

The targets of a user query are documents. Examples of documents include text-based computer files, as well as files that are partially text-based, non-text files, files containing spatial information, and computer entities that can be accessed via a document-like interface. Documents can contain other documents and may include other interfaces besides their document-like interfaces. Every document has an address. In the case of World Wide Web documents, this address is commonly a URL. As is the case with URL's, a portion of the address may include instructions or parameters that are passed to the computer server process that serves the document.

The documents exist on computer systems arrayed across a computer network, such as a private network or the Internet. The documents may be hyperlinked, that is, may contain an address of another document. Copies of the documents may be stored in the page repository 222 (Fig. 3).

GENERAL USER INTERFACE

With reference to Fig. 2, the map interface 80 is presented to the user on a computing device having a user interface. The user interface may be graphical (GUI), voice-based, or text-only. Each feature of the GUI will be reproduced in a voice-based or text-only user interface, to the extent possible.

As is common in the art, the GUI includes a pointer symbol responsive to the user's manipulation of a pointing device such as a mouse, a touch-sensitive area, or a combination of directional buttons. The pointer symbol is superimposed on the GUI contents. The GUI is also responsive to a click event generated by the user. The click is usually associated with the user's manipulation of a button on or near the pointing device, but may be activated in other ways, depending on the computing device and its operating system. The client process receives click events and the position of the pointer symbol from the operating system of the computing device.

The map interface 80 includes a map 805. The map 805 is a representation, often in part, of at least one spatial domain. A spatial domain is any space with a location metric known to the spatial recognizer 48. In one embodiment, the surface of the Earth is a spatial domain under the 2-dimensional location metric of latitude and longitude – henceforth, the

"standard geographic domain." In another embodiment, the "GPS domain" is defined by a volume around the surface of the Earth under the 3-dimensional metric of GPS (global positioning satellite) data.

A metric on a space need not identify spatial point locations. A document may be identified as being near a spatial point location. For example, a document could be identified as "near exit 19 off I-80 in Pennsylvania." A document could refer to an extended region like Plum Island state park.

The map 805 uses a scale in representing the domain. The scale indicates what subset of the domain will be displayed in the map 805. There is usually a range of scales appropriate to a given domain. By choosing a smaller scale, the user can examine a smaller portion of the entire domain in exchange for more detail per unit.

Several domains may exist along one spatial continuum. For instance, in one embodiment, the map 805 initially displays a representation of the whole surface of the Earth under the standard geographic domain. The map 805 might then be changed to display only one continent. This map 805 still displays a portion of the domain, but the scale has changed. However, the scale might change to the point that the map 805 displays only a close-up of a concert hall. At that point, the map interface 80 can change the domain to that of the concert hall, where locations can be expressed by section, row, and seat number, for instance. Domains can intersect or overlap, therefore, along a continuum of scale.

When the domain has geographic meaning, the map 805 may include standard geographic map features such as streets and waterways. Data for geographic features is available from the U.S. Census Bureau, the U.S. Geographic Survey, and companies such as GDT, of 11 Lafayette Street, Lebanon, NH, or NavTech, of 10400 W. Higgins Road, Rosemont, IL. The map can include spatial landmark features appropriate only to certain scales within the same domain, such as street names that do not appear until the scale is sufficient to allow individual streets to be represented.

The map 805 need not represent a domain that physically exists; the map 805 may represent something that is itself a representation, such as a virtual layout of a planned housing development. Still more abstractly, the map 805 may represent entities in a spatial layout where the spatial dimensions do not correspond to physical spatial dimensions. For

instance, the domain may be a genealogical tree laid out on a plane, wherein one axis of the plane represents the linear progress of time.

A domain location is a location in the space that the map 805 represents. The domain location is usefully distinguished from the display location, which describes the placement of elements as displayed by the map 805. The measure of distance between domain locations depends on the domain, whereas the distance between display locations is measured in pixels of the computing device hosting the client 64.

A domain frame is the subset of the overall domain (possibly including the entire domain) displayed by a state of the map 805.

CHANGING THE MAP VIEW

The user can adjust the view displayed by the map 805 in several ways.

The user can change the scale of the map 805 by a click on the zoom bar 891. The zoom bar 891 visually represents a plurality of scales that the map interface 80 is capable of displaying in the map 805. The scales displayed by the zoom bar 891 in any given state may be a subset of the total scales the map interface 80 is capable of displaying in the map 805. This subset may adapt to a change in state, including a change in scale. For instance, in a geographic context, the first state of the map interface 80 may display the entire globe in the map 805. In this first state, the zoom bar 891 may display scales ranging from global to street level, for instance. If the user narrows the scope of the displayed region to a street address corresponding to a concert hall, the zoom bar 891 might display scales ranging from seating sections to individual seats inside the hall.

The user can move the center of the map 805 by a click on the map border 892. The map border 892 surrounds the map 805.

By interacting with the map mode controls 830, the user can specify how the map interface 80 should respond to clicks on the map 805. The map mode controls 830 include controls for pan 832, zoom 834, and post a note 836. The pan 832 control and the zoom 834 control each have states including an "on" state and an "off" state. When the pan 832 control is in its "on" state, a click on the map 805 instructs the map interface 80 to re-center the map 805 around the location represented by the click. Likewise, when the zoom 834 control is in

its "on" state, a click on the map 805 instructs the map interface 80 to zoom the map 805 in around the location represented by the click. The post a note 836 control is described in the section on electronic notes, below.

The map interface 80 includes spatial criteria entry controls 806. The spatial criteria entry controls 806 include a data entry control 808, a submission control 809, and a prompt for spatial criteria 807. The prompt for spatial criteria 807 instructs the user as to the purpose of the data entry control 808. The prompt for spatial criteria 807 may include a static instruction or may be dynamically responsive to user interaction, such as movement of the pointer symbol over the data entry control 808. The prompt for spatial criteria 807 may include sound. The user invokes the submission control 809 to notify the client process 64 that the data in the data entry control 808 is complete.

Examples of spatial criteria include geographic measurements such as latitude, longitude, or altitude; postal address information; or, returning to the concert hall example, row and seat number. Spatial criteria also include criteria that are indirectly spatial, i.e., criteria that do not describe spatial attributes but specify an entity which does have spatial attributes. An example of indirectly spatial criteria is the tracking number of a package. The tracking number might not be spatially descriptive in itself, but the package at any given instant might have a last known location that could be represented on the map 805.

The map interface 80 includes keyword entry controls 801. The keyword entry controls 801 include a data entry control 803, a submission control 804, and a prompt for keywords 802. The prompt for keywords 802 instructs the user as to the purpose of the data entry control 803. As with the prompt for spatial criteria 807, the prompt for keywords 802 may include a static instruction or may be dynamically responsive to user interaction, and may include sound. The role of the submission control 804 within the keyword entry controls 801 is identical to the role of the submission control 809 within the spatial criteria entry controls 806.

Note that part or all of the spatial criteria entry controls 806 and the keyword entry controls 801 may use the same interface components. For instance, if the user enters text "shoes near to Cambridge, MA," the system may consider this both spatial criteria and keyword criteria.

Examples of keywords include any word of interest to the user, or simply a string pattern. The computer system 20 compares the data in the data entry control 803 against the contents of the documents in storage 22. There are no predetermined restrictions on the keywords that may match a document.

The user may enter any text he/she desires in the text entry tools. The computer system 20 will parse entries to get possible domain changing commands and keyword queries. Keyword queries can be of any form. The computer system 20 does not restrict the entries to predefined categories. Instead, the computer system 20 attempts to match the query text against text found in all documents in the corpus.

One way to match the query text is to split it into separate strings divided by white space, where white space is commonly defined in the art as tabs, spaces, carriage returns, and other characters generally referred to by the regular expression character "\s". Each of these separate strings can then be searched for in the documents.

The text contents of the documents can be similarly divided into separate strings divided by white space. Thus, if the text entered by the user match any strings in the document corpus, the computer system 20 can retrieve results.

This "free text entry query" allows much more versatile searching than searching by predetermined categories.

ICONS

The map interface 80 may include one or more icons 810 superimposed upon the map 805. Icons 810 need not be present in the map interface 80 when the client 64 is initially presented to the user. After the user has submitted a query, though, the map interface 80 may use icons 810 to represent documents in storage 22 that satisfy the query criteria to a degree determined by the search 50 process.

The display placement of an icon 810 represents a correlation between its documents and the corresponding domain location. Specifically, for a given icon 810 having a domain location, and for each document associated with the icon 810, the subsystem for data analysis 20 must have determined that the document relates to the domain location. The subsystem for data analysis 20 might determine such a relation from a user's inputting that location for

the document. Note that a document can relate to more than one domain location, and thus would be represented by more than one icon 810.

An individual icon 810 belongs to an icon class. Icons 810 of the same icon class share visual characteristics that may include shape, color, size, indexing scheme (Roman numerals versus letters, e.g.), or animated behavior. An icon face 818 is an interface element of the map interface 80 satisfying the requirements of an icon class. In one embodiment, the client process 64 runs on a computer equipped with a monitor having a pixel size of approximately 0.28 mm, which is approximately the industry standard for desktop computers at the time the invention was made. For this pixel size, typical icons would be 15 to 20 pixels in diameter.

Note that there may be more than one way to satisfy the requirements of an icon class, so an icon class may have more than one icon face 818. For an example, see icon subclasses, below.

Visual similarities conferred by icon class may be used to represent topical similarities among the documents the icon 810 represents. For instance, documents affiliated with restaurant menus might be represented by icons 810 sharing a fork-and-knife shape. The fork-and-knife shape would be a property of the icon class.

Different colors, shapes, tints, and animated motions of the icons 810 might represent different features of the documents represented by the icons 810.

A class of icons may share the same geometric shape but have different colors, or different shades of the same color. The different shades might represent the several different properties of the documents represented by the icons. Different properties of the documents include the time elapsed since the document was created, the time elapsed since the document was introduced to the system, a relevance measure of the document, the size of the document.

Another feature of the icon class is the icon subclass. Two icon classes may be subclasses of a third class if they share the characteristics affiliated with the third class but vary at least one other characteristic in a consistent and meaningful way. For instance, the icon class for restaurants might have subclasses for quality, as measured by a certain newspaper's restaurant reviews. All icons 810 in the icon subclasses for restaurant quality would have a fork-and-knife shape in common, but icons 810 would be colored green for

good reviews, red for poor reviews, or yellow for mixed-opinion reviews. They could even be divided in pie-chart fashion to show a percentage of each type of review. Thus, broad visual similarities can be used to imply broad topical similarities on one level, while visual sub-variation can be used to imply topical sub-variation on a second level. The icon legend 812 can inform the user of such conventions.

If icon class B is a subclass of icon class A, then icon class A is a parent class of icon class B.

Many domain locations have multiple documents referring to that location. To illustrate this to the user, the icon 810 used at that location in the map interface might be of a different size, color, or shape from other icons 810. For example, an icon 810 may be made to appear "stacked" as if a few of the icons 810 where placed nearly on top of each other. For another example, the icon 810 might appear to have parts of different icons 810 spliced together.

In a preferred embodiment, different colored icons 810 represent different layers of documents; varying shapes represent varying numbers of documents; varying shades represent varying relevance numbers for the underlying documents. The relevance of a set of documents referring to a given domain location might be computed by averaging or summing the relevance of the individual documents.

An icon 810 may represent one location in a domain or several neighboring locations. The number of locations depends upon several factors, including the scale of the map 805. When multiple icons 810 have display locations within a tolerance determined by the computer system 20, the map interface 80 consolidates icons 810 to increase visual clarity. Should the user change the scale of the map by zooming it, the map interface 80 recalculates whether to consolidate icons 810. The tolerance beyond which icon consolidation occurs may vary. The primary factor in the decision to consolidate is whether icons 810 are overlapping. For many icons 810, a good test of overlap is whether the display locations are closer than two times the average diameter of the icon faces 818. Other factors in the decision to consolidate include visual characteristics of the icon faces 818, visual characteristics of the map 805, characteristics of the domain, characteristics of the documents, and the number and variety of icons currently present in the display.

A consolidated icon 810 may represent multiple spatial domains. For instance, consider the standard geographic domain that includes Washington, DC, and another domain for Ford's Theater under the concert-hall seating metric. At certain display scales, documents that represent the Lincoln Memorial might be consolidated into the same icon as documents describing the specific seat in Ford's Theater in which Lincoln was shot. In this example, the Lincoln Memorial document might be affiliated with the standard geographic domain. The Ford's Theater document might be affiliated with a domain specific to Ford's Theater, but in this example is may be displayed on the standard geographic because the entire Ford's Theater domain can be mapped onto a fairly small region, relative to the size of the domain requested by the user.

An icon 810 may also represent multiple topical categories among its documents, regardless of whether the icon 810 is consolidated. In this case, the icon face 818 may be altered to reflect the multiplicity of topics.

The icon legend 812 is another element of the map interface 80. The icon legend 812 relates an icon 810 to the documents it represents. The icon legend 812 comprises a listing of documents. The listing may be grouped or ordered in a variety of ways.

Icons 810 are listed in the icon legend 812 according to an order compiled by the search 50 process.

A non-consolidated icon 810 represents a single display location. The order of its documents as listed in the icon legend 812 is based on a relevance ranking compiled by the search 50 process. The relevance ranking scores each document against the user's query criteria.

A consolidated icon 810 may represent a plurality of domain locations. A consolidated icon 810 may represent a plurality of icon classes. The different icon classes may entail different topical categories. The icon legend 812 may differentiate the document listings according to these topical categories: for instance, by grouping them by category; by adding a field to each entry in the list, specifying the category; or by adding a visual emphasis. The visual emphasis may include a change in typeface, a change in color, or the presence of an icon type affiliated with the category. Several effects can be combined, such as grouping by category in combination with a variation in background color between adjacent groups.

FILTERS

The map interface 80 includes two groups of controls for managing filters, a general filter display 850 and a user-specific filter display 860.

With reference to Fig. 3, a filter selects a subset of the corpus of documents in the page repository 222. Filters are defined recursively: a filter is a list of elements, where each element can be either a keyword string, a set of spatial criteria, a human-compiled list of documents, a domain frame, or another filter. The elements may be defined in a sequence allowing the user to select a collection of documents. The sequence of filters may be combined with the Boolean AND operator to produce an intersected document set that is the same for any order of the filters. Two sets of filters may be combined with the Boolean OR operator. When viewing a set of documents in a map 805, the user may change the map view to display a subset of this document set, which may be different than if the user performed the filtering operation after changing the map view. Thus, every user query defines a filter, because it contains either keywords, spatial criteria, a change to the domain frame, or several in combination. The initial state of the map interface 80 – even if the user has not yet interacted with it – defines a filter, since the map 805 has at least a domain frame associated with it. Similarly, because a non-empty map 805 defines a filter, zooming or panning the map 805 always defines a new filter based on the previous filter plus the new domain frame. Each group of icons 810 defines its own unique filter: namely, the filter defined by the current state of the map 805, but with the resulting documents restricted to those associated with at least one of the icons 810 in the group. In this way, a click on an icon 810 can define a filter, since a single icon 810 is simply a group of one.

The general filter display 850 includes filters created for the user. The user-specific filter display 860 includes filters created by the user. The two sets of controls, 850 and 860, can be disjoint or can share controls in the map interface 80.

The general filter display 850 includes general 852 filters, search history 854 filters, and inferred 856 filters. A general 852 filter is a filter predefined by the computing system 20. This includes filters handpicked by human editors to be of general interest to the user population, as well as filters selected algorithmically for having a high frequency of recurrence among the usage patterns of the user population. A search history 854 filter is a filter the current user has applied in the current or previous session possibly without

explicitly instructing the system to remember it. By providing easy access to search history 854 filters, the system allows the user to reapply a filter that he/she created earlier but neglected to add to the user-specific filter display 860.

An inferred 856 filter is a filter selected algorithmically based on the usage patterns of the current user.

A data-mined 857 filter is a filter created algorithmically by a procedure that analyzes the content and hyperlinks of documents in the page repository 222 to create a set of documents sharing a property. The property may be determined heuristically, e.g. "all documents appearing to relate to cooking recipes." The algorithm to construct such a filter might include the use of Bayesian learning, statistical analysis, and ontologies of words and phrases.

The user-specific filter display 860 is not shown by certain states of the map interface 80. For example, if the computer system 20 cannot determine the correct user profile to apply to the current user, or if a security measure associated with the profile has not been satisfied, the user-specific filter display 860 may be hidden or disabled.

When displayed and active, the user-specific filter display 860 includes filters associated with a user profile. The user can add, modify, or delete these filters, and can assign them to user-defined groups.

Filters that the user can add to the user-specific filter display 860 include: a filter in the general filter display 850; the filter defined by the current state of the map 805; the filter defined by a group of icons 810, which the user can specify by using the pointer symbol; a filter combined from at least two existing filters; and a modified filter which the user chooses to save under a new name.

The modifications that the user can apply to a filter in the user-specific filter display 860 include: renaming the filter; adding, deleting, or reordering elements in its list; and changing the icon class associated with the filter or defining a new icon class for the filter. Properties of the icon class that the user can edit include: its name, its icon face 818, its parent icon class, a textual summary of the document, and any properties displayed in the icon class legend 817.

ELECTRONIC NOTES

A note document is a document associated with a domain location. It may also be associated with a user profile, or it may exist anonymously. An electronic sticky-note 870 is a representation of a note document displayed on the map 805 in a display location corresponding to a domain location associated with the note document. The note document can contain any form of information that a document in storage 22 can contain. For instance, the note document may contain text, graphics, sound, video, hyperlinks, or a combination thereof. The note document can have its own URL and act as a web page.

The post a note 836 control changes the state of the map interface 80 such that a subsequent click on the map 805 will create a new note document. The note document will be associated with a domain location corresponding to the display location that was clicked, and an electronic sticky-note 870 will appear at said display location and be associated with the domain location represented by that display location.

In one embodiment, having put the map interface 80 in the appropriate state, the user can move document content from outside the client process onto the map 805, thereby initiating a note document creation. The content can be moved by drag-and-drop or copy-and-paste, among other methods appropriate to the computing environment and the media type. For instance, the document content could be a media stream which the computer system 20 begins recording. The content becomes part of the new note document, and the note document is given at least one externally accessible address such as a URL. With the map interface 80 in the appropriate state, the user can create web pages, for example, with one rapid action. In this embodiment, the mechanisms allowing the user to drag-and-drop or copy-and-paste content are provided by the operating system. The terms "drag-and-drop" and "copy-and-paste" are well known in the art.

Several other features of these note documents require description. Users can specify calendar dates and/or times when a document is not to be served to the public, or will expire altogether. When a note document expires, it may be deleted from storage or prevented from appearing in the interface. This allows users to post time-sensitive information at geographic locations. Short lifetime note documents might be used to make an animated icon on the map interface. Such an icon could follow a moving object or a user's approximate path through the domain.

Users can digitally sign note documents to help ensure their authenticity to other users. Public key cryptography, like PGP, is standard in the art and can be used to affect this. The audience of a document can be limited using this same type of public key cryptography or by requiring users to login with a private password that authenticates their identity. The creator of the note document can determine the list of registered users permitted to see a particular note document. Alternatively, the creator could distribute the encryption key needed to open a note document. This allows users to publish note documents to a subscription list.

Users can host their own note documents on private computer systems. Such private computer systems may be licensed copies of part or all of the computer system 20. Such a privately held note document might be protected by security measures. The creator of such a note document can create additional note documents in other instances of the computer system 20, which may be owned by other people or companies. These additional note documents could provide pointers to one or many note documents on the creator's private computer system. These additional note documents might contain a summary of the original note document. Users of one instance of the computer system 20 may have access to certain other instances of the computer system 20. This access is determined by the owner of each instance. This allow many instances of the computer system 20 to participate the hosting and distribution of geographically-located note documents.

Since any media type can be easily put in to a note document, it is easy for the owners of an instance of the computer system 20 to create note documents from data from other computer systems under their control. For example, a store owner can copy their inventory database into note documents in their instance of the computer system 20. This conversion of a store database to geographically-located note documents makes it easy to serve the inventory information to other users interested in the stores physical area.

A user can upload or create a collection of note documents in one action, such as dragging and dropping a folder of documents into the map interface. If the documents contain location information, they can be automatically posted in the map interface. If not, the user can be prompted to select locations for each document.

Such a collection of note documents will be grouped in a filter in the user-specific filter display 860. Examples of such grouped note documents include a collection of

photographs taken on a vacation, a collection of sound recordings taken around a city, a set of data gathered from various sensors, a sequence of events for an newspaper article, or a set of descriptions for a trail guide. A collection might have colored lines connecting the various icons on the map 805, thereby indicating a path that could be followed by a user in the domain.

Such a collection could be created for a user by a service or device. For example, a user's camera might include a GPS or other spatial locating device that imprints each picture with a location stamp. Uploading the pictures is then quite simple: the stamps locate each picture on the map 805. A service might do this on a user's behalf. For example, a hospital might annotate a user's medical record with locations of where the user was treated and post them as a private note document collection for the user and other care providers.

The user can post a note document containing dynamic software such as a discussion board, order entry tools, telephone connect service, or other software-backed tool. A note document posted at the location of a vending machine might have an order entry tool connected to the vending machine that allows users to use a credit card or other payment mechanism to purchase items from the machine. This allows users get physical items without paying cash or even carrying a credit card.

A note document posted at a store might contain a discussion board with text and other media entry tools allowing the general public to engage in a discussion at that location. Such message boards might receive text messaging from portable phones and broadcast them to users viewing the discussion board.

A note document might contain a tool, which, when clicked, causes a user's phone to dial into a service. Such a note document might be posted at a restaurant or theater where telephone reservations are required.

COMMUNITY FEEDBACK

The map interface 80 can use the community feedback 880 control to show the user information gathered from the behavior of other users. Features of the community feedback 880 control include domain usage feedback 882, word-domain suggestion 884, and word-word suggestion 886.

When the user views a spatial domain, domain usage feedback 882 tells the user how many people have viewed that domain or part of that domain in the recent past. For example, "23 people have viewed this region in the last 18 minutes."

When a user views a spatial domain, word-domain suggestion 884 can tell the user keywords that are relevant to this domain. These words can be gathered by analyzing documents that refer to this region to find the words that occur most in that domain. These words may also be gathered by recording the keywords that other users have entered when viewing this region. The most commonly searched for words can be presented to the user.

When a user enters a keyword query, word-word suggestion 886 can tell the user additional keywords that relate to the keyword(s) just entered. These additional keyword suggestions come from a thesaurus that may be built by recording the sequence of queries entered by other users. If many users enter the same keywords together or in a single session, then those keywords can be considered related. For example, if many users search for "chocolate" and then search for "chocolatier" the computer system 20 can suggest to the next user who enters "chocolate" to try a keyword query for "chocolatier." This suggestion helps users find what they want.

DATA COLLECTION

The computer system 20 includes a data collection 30 process for gathering new documents. With reference to Fig. 3, the data collection 30 process includes a crawler 36 process, a page queue 34, and a metasearcher 32 process.

CRAWLER AND PAGE QUEUE

The crawler 36 loads a document over a network, saves it to the page repository 222, and scans it for hyperlinks. By repeatedly following these hyperlinks, much of a networked system of documents can be discovered and saved to the page repository 222. The crawler 36 gathers documents into the computer system 20 in this manner. In one embodiment, these documents are World Wide Web pages available on the Internet. In this case, downloading pages can be done using any of the various Internet protocols, including the HyperText Transfer Protocol (http), the File Transfer Protocol (ftp), gopher, news, wais, and others.

The page queue 34 stores document addresses. The crawler 36, the pioneer 48, and the metasearcher 32 add document addresses. The page queue 34 comprises a database table, the page queue table 340.

The crawler 36 gets document addresses to crawl from the page queue 34. When the crawler 36 loads a previously unknown document, it passes the document to the pioneer 48 process. The pioneer 48 parses the content of the document for hyperlinks to new documents. The pioneer 48 adds any addresses referenced by such hyperlinks to the page queue 34.

The crawler 36 makes use of the fact that the probability of being spatially relevant is correlated with linkage; in other words, pages linked to a spatially relevant page have a greater probability than average of being spatially relevant. Each crawled URL is assigned a spatial relevance. Considering spatial relevance helps the crawler 36 use time and other resources efficiently.

The crawler first crawls pages linked from those pages with spatial relevance greater than a predetermined threshold. After a page has been downloaded and its spatial relevance calculated, its spatial relevance level 342 field can be recalibrated to reflect the actual relevance we found.

METASEARCHER

The metasearcher 32 initializes the collection of known documents. This initializing step is called "seeding" or "bootstrapping." The computer system may have to be seeded for each domain. For example, separate bootstrapping operations may be used for United States postal addresses and French postal addresses.

The metasearcher queries search engines known to store information appropriate to the domain, such as search engine web sites on the Internet. The human administrators of the metasearcher provide it with a collection of known spatial locations appropriate to the domain. The metasearcher formulates queries based on these spatial locations and directs the queries to the search engines. The results are compared to the collection of known documents and are added if new.

A crawling is complete when all discoverable documents on the network have been found. In practice, this rarely happens over large document collections unless the collections

are extremely static. Thus, since a complete crawling is rarely likely, the speed of the crawl is an important design concern. The speed of crawling is limited by the speed at which new pages are discovered through links on previously downloaded pages. A good way to accelerate this crawling is to query existing search engines that have already crawled at least part of the document collection, which could be the Web. The results given by these search engines are used to bootstrap the data collection 30 process.

In one embodiment, the metasearcher 32 bootstraps its knowledge of the geography of the United States. The process for this bootstrapping comprises six steps. Other domains may require different processes.

The steps are a system of levels intended to gather the most useful spatial URLs from existing search engines. Since search engines commonly limit the number of results returned to a single query, searches might not return all the results that we would like to gather. For instance, in a geographic query, this happens with town names like "Boston, MA." In such cases, it is useful to specify other words in the query, such as all the street names in that town.

Major search engines include AltaVista, Fast, Lycos, MetaCrawler, DogPile, NorthernLight. Each engine has a maximum number of results that they will return for a query, even if they have more pages that meet the query. If a metasearch query overflows this number, the metasearcher 32 adds words to the query to squeeze out more URLs.

In step 1, the metasearcher 32 queries the search engines with just the town names, e.g. "boston" "cambridge" "new york" "madison" "san antonio".

In step 2, for any town name that resulted in the maximum number of results for that engine, the metasearcher 32 re-queries the search engine with the town and the state, e.g. "boston, ma" "boston mass" "boston massachusetts" "cambridge, ma" etc... "new york, ny" etc ... "madison nj" ... "madison ny" ...

In step 3, the metasearcher 32 switches to a second table, which has more information. The second table includes all the streets in every town in the USA. For any town-state pair that overflows on a particular engine, the metasearcher 32 queries for every street, e.g. "highland somerville" "hancock somerville" "elm somerville" etc.

In step 4, the metasearcher 32 adds in state names with the street names, e.g. "highland somerville ma" "hancock somerville ma" "elm somerville ma" etc.

In step 5, the metasearcher 32 adds in street types, e.g. "highland ave somerville" "highland avenue somerville" ... "hancock st somerville" ... "elm st somerville" etc.

In step 6, the metasearcher 32 adds in street types and state names, e.g. "highland ave somerville ma" "highland avenue somerville ma" "highland avenue somerville massachusetts" etc. Few places reach this level.

The page queue table 340 includes a spatial relevance level 342, which helps constrain the crawler 36 to documents that are spatially relevant. When the metasearcher 32 gathers a document, the document is given a level of "0."

DATA ANALYSIS

With reference to Fig 4., the computer system 20 includes a data analysis 40 process for extracting information and meta-information from documents. Data analysis 40 includes a spatial recognizer 42 process, a spatial coder 43 process, a keyword parser 44 process, an indexer 46 process, a spatial document ranking 45 process, and a pioneer 48 process. The role of the pioneer 48 process is described in the section for data collection 30. In the data analysis section, we will repeatedly cite the example of the standard geographic domain for the USA, identified by the standard latitude/longitude but also by postal system addresses, localities, and phone numbers.

SPATIAL RECOGNIZER

As new documents are saved in the page repository 222, the spatial recognizer 42 opens each document and scans the content. It searches for patterns that resemble parts of spatial identifiers. For example, in the standard geographic domain for the USA, patterns include street addresses of the USA postal system, localities, and phone numbers.

In step 422, the spatial recognizer 42 finds candidate spatial data in unstructured text. Candidate spatial data, is called a PSI, for possible spatial identifier.

In step 424, the spatial recognizer 42 parses the text of the candidate spatial data to determine its structure, thereby forming a PSI. We break addresses into a standard set of fields used by the US postal system. Similar formats exist for other postal systems, which would be represented as other domains. The constituent parts of the PSI are identified. Not all may be present in a given document; for localities and phone numbers, only town, state, and possibly ZIP and ZIP+4 are used. The constituent parts include:

House number
Street prefix (e.g. East, South)
Street name
Street suffix (e.g. East, South)
Street type (e.g. Street, Turnpike, Square)
Town
State
Zip
4-digit zip extension

PSIs are stored in the spatial lexicon 224 for further analysis. The table for these possible spatial identifiers (PSIs), which in this case is mapped against the standard geographic domain, includes fields for latitude and longitude. Regardless of domain, the table may include fields for spatial coding confidence, number of documents located at this place, status of spatial coding, and sum of relevances of documents located at this place.

The relevance scorer 426 assigns a relevance score to the document.

The relevance scorer 426 includes a multiple spatial references partitioner 4262 process. Many documents have multiple spatial references. It might be the case that all the spatial identifiers are relevant to the whole document. An example is a web page listing branch locations of a store chain. However, it can be the case instead that each spatial identifier is only relevant to a proper subset of the page. An example of this is a page giving short reviews of a number of restaurants. Such a page is a multi-part document.

Multi-part documents present a problem when searching the document collection by keyword. Were the document to be keyword indexed as a whole, a word in one part of the document would be indexed as though it were relevant to addresses in a different part of the document, when in fact the word may not be relevant to that part.

To detect multi-part documents, the multiple spatial references partitioner 4262 invokes the multi-part cluster measurement 42625 process. The multi-part cluster measurement 42625 process first rejects any document with fewer than some number of addresses (usually 5) or which is shorter than some number of words (perhaps 200). The multi-part cluster measurement 42625 process computes an array containing the fractional positions of each PSI in the page. For instance, an address that begins at the 200th word in a 1000-word document is at fractional position 0.2. We then apply a clustering statistic such as the Gini coefficient to produce a clustering score that expresses how concentrated the

addresses are on the page. Documents with low clustering score (indicating that the addresses are evenly spread out) are likely to be multi-part documents. The threshold for the maximum clustering score is determined empirically and may vary for each domain.

The multiple spatial references partitioner 4262 partitions the document into segments that contain one PSI each, using the PSIs as boundaries, as follows. The nth segment, containing PSI n, begins at the word following the end of PSI n-1, and ends at the word before PSI n+1. For n=1, the segment begins at the first word. For the last PSI on the page, it ends at the end of the page.

Each segment then has the title portion of the document added to it. The tag recognizer 442 provides one way of determining the title portion of a document.

The segment is stored in the page repository 222 to be separately indexed. The unsegmented page is retained, so that when a segment is found as a search result, the full document can be returned, with an anchor placed at the beginning of the segment so that the document can be scrolled to the segment before presenting it to the user.

SPATIAL CODER

To further analyze the PSIs, the spatial coder 43 process runs several processes that associate domain locations with various identifiers in the document content. In the standard geographic domain, we can associate latitude/longitude points or bounding polygons with identifiers; this process is known as geocoding. If no latitude/longitude can be matched to a PSI, the spatial coder 43 marks it misrecognized. Otherwise, the spatial coder 43 turns the PSI into a known spatial identifier, or KSI. This completes the entry in the spatial lexicon 224 described above.

The spatial coder 43 for the standard geographic domain for the USA includes an address encoder 432, a locality encoder 434, and a phone number encoder 436.

With reference again to the standard geographic domain for the USA, addresses are considered the best match. Thus, if a page has addresses in it, simple place names like "Cambridge, MA" and phone numbers are not used to spatially code the page. A page can have multiple KSIs, but that reduces its spatial relevance (see spatial document ranking 45), so we look primarily for pages with only a few highly focused KSIs. A focused KSI means that the spatial coder 43 associates a small area in "lat/long space" (space identified by

latitude and longitude) with high certainty. Thus, for example, a phone number associates with a region the size of a telephone exchange, which is at least several square miles, but a postal address associates with a "rooftop" sized region usually represented by a point in the middle of the hypothetical rooftop. If a phone number and an address in a document both agree on the location of the page, we can improve the ranking of the document (spatial document ranking 45).

Address encoder 432: Postal addresses in the USA and other countries can be associated with small geographic regions, usually the size of a building. Standard geocoding procedures approximate this by a point. Given a PSI like this, for instance:

77 massachusetts ave|cambridge|ma|02139

the associated lat/long can be discovered by feeding the text string into any standard address geocoding product. Examples include Etak's Eaglecoder, Sagent's GeoStan, and ESRI's ArcINFO geocoding plug-in. The output of Etak's Eaglecoder looks like this:

<command line interface> jrf@raag:~$ mc/lib/etak/rie -b
<input text of PSI> 77 massachusetts ave|cambridge|ma|02139
<output of geocoder> 77 MASSACHUSETTS AVE,CAMBRIDGE,MA,02139,42.358968,-071.093997

The third line of the output contains lat/long information to associate with this address. Thus, this PSI can be converted into a KSI.

Locality encoder 434: Place names, like "Boston, MA" and "Washington Monument," are listed by the US Census along with the latitude longitude of the center of the place. This makes it easy to geocode them. The locality encoder 434, similar to the address encoder 432, searches for candidate strings that could be town and state names. The locality encoder 434 differs, however, in that it looks up the town name in a database of all known towns in the United States 2262, and rejects the town name if it does not appear.

Phone number encoder 436: The phone number encoder 436 converts phone numbers to geographic locations by looking up the area code and exchange in a phone-to-place table 2266. The phone-to-place table 2266 maps area code-exchange pairs to town name-state name pairs. This pair is then treated as a locality name, except that its relevance score is lowered by a small constant number (determined heuristically) to reflect the fact that towns obtained in this way are somewhat less valuable than towns that have been mentioned by name. A single telephone company central office may cover multiple towns, especially in

suburban locations; there is a chance that the phone number is actually located in a neighboring town.

SPATIAL MEANING INFERENCE

The spatial coder 43 includes a spatial meaning inference 438 process, or SMI 438, which can perform a special type of spatial coding. The SMI 438 can deduce a spatial relevance for terms (words and phrases) based not on a semantic interpretation but on statistical properties of appropriate portions of the spatial-keyword document index 505.

Certain words and phrases correspond to geographic locations but are not recorded by any existing geocoding services. To discover these geographic relations, the SMI 438 statistically analyzes the correlation of candidate words and phrases with KSIs. The SMI 438 uses the premise that if a phrase occurs mostly in documents with addresses in the same place, then the phrase is probably also about that place. For example, "the big apple" occurs on many pages with the words "New York, NY" and addresses in New York City. The SMI 438 can deduce that "the big apple" is also about New York City.

The SMI 438 deduces spatial relevance as follows. The spatial-keyword document index 505 contains a tree for each indexed term, i.e. each term in word lexicon 225. For each word in a given string, the SMI 438 examines the tree associated with that word. The examination includes invoking the imbalance measurer 439 to measure a degree of imbalance in the structure of the tree, which, since it is a trimmed version of the spatial document index 503, may have significant imbalance as a result of trimming. The imbalance measurer 439 is described below. Broadly speaking, and as will be described in more detail, if enough terms in the string have trees which have similar imbalances, the SMI 438 associates the string with the spatial regions described by the imbalanced portions of said trees.

Returning to an earlier example, each word in the phrase "the big apple" appears in many documents. Performing a search over a spatial-keyword document index 505 for that phrase without specifying a bounding box will find a large "peak" in the number of documents near New York City. This is evidenced by the degree of imbalance in the trimmed result tree. The tree resulting from the intersection of these three words has many branches in the latitude-longitude region covering New York City. This tells us that pages with these three words next to each other are probably referring to this lat/long region.

We call such words and phrases "geographic phenomena."

A tree address is defined as follows. Given a spatial-keyword document index 505, any node or leaf in the index trees can be identified by a set of values indicating the sequence of child node numbers that must be traversed to reach that node. For example, in a binary tree, the tree address 0110 specifies the node found by starting at the root node and going to the first child's second childs' second child's first child. In a 16-way tree, the tree address written in hexadecimal as "0x4f8" specifies the node found by starting at the root node and going to the fifth child's sixteenth child's ninth child.

To measure the "peakiness" of a particular phrase without using a spatial-keyword document index 505, the imbalance measurer 439 first computes a "standard peakiness" of average words and then compares candidates to that. In one embodiment, the imbalance measurer 439 computes the standard peakiness by picking a random sampling of words and, for each of those words, computes the 2-dimensional variance of the points referred to by documents that contain the word. Documents that are particularly relevant to a word can be given extra weight in computing the variance, e.g. a highly relevant document can be scaled linearly so that it appears to represent multiple documents at that location. Given this random set of variances, the imbalance measurer 439 computes the average variance. The average variance can be used as a baseline to detect a geographically relevant phrase or word. Any word or phrase with a variance much smaller than the baseline is a geographic phenomenon.

Use of the spatial-keyword document index 505 simplifies the SMI 438 dramatically. Since the trees in the spatial-keyword document index 505 already span all the documents known to the computer system 20, the SMI 438 can detect a geographic phenomenon simply by considering the set of tree addresses of leaves in a trimmed result tree. For example, given a candidate word or phrase, the SMI 438 queries the spatial-keyword document index 505 to get the trimmed result tree for this word or phrase and performs the following operation on this list of addresses.

From the tree, the SMI 438 creates a list of the tree addresses of every leaf. Starting at the first digit in all the addresses, the SMI 438 finds the most common branch number at this level (i.e., for this digit). The branch indexed by this digit is called a "candidate fork"

because it is a fork of the tree, pointing in the direction of the candidate location. The SMI 438 computes the fraction of the addresses that follow the candidate fork at that level.

At the next level, the SMI 438 considers all addresses that took the candidate fork in the last level and once again finds the most common fork direction, using it as the next fork direction. The SMI 438 again computes the fraction of addresses still following the candidate fork.

The SMI 438 repeats this until the percentage of addresses still following the candidate fork falls below a predetermined threshold adjustable by the operators of the computer system 20. The particular threshold may be adjusted for each domain. Adjusting the threshold adjusts the quality of matches that are considered. It is set empirically.

For example, for simplicity of explanation consider a binary tree whose nodes divide a domain space into rectangles, and consider these four addresses that fork together for several levels:

1011110101011111
1011101011101010
1011101011101111
1011101011101101

Level 1: forked 1 = 100%
Level 2: forked 0 = 100%
Level 3: forked 1 = 100%
Level 4: forked 1 = 100%
Level 5: forked 1 = 100%
Level 6: forked 0 = 75%
Level 7: forked 1 = 75%
Level 8: forked 0 = 75%
Level 9: forked 1 = 75%
Level 10: forked 1 = 75%
Level 11: forked 1 = 75%
Level 12: forked 0 = 75%
Level 13: forked 1 = 75%
Level 14: forked 1 = 50%
Level 15: forked 0 = 25% -- below 50% threshold.

These tree addresses suggest that the word is 100% relevant to a region defined by the rectangle in the spatial index tree 502 by the address 10111, and 75% relevant to the rectangle 10111010111.

If a particular word is rare, i.e. occurs only a few times in the entire page repository 222, but its appearances are highly correlated with geographic identifiers in the same place, then that word might be associable to a point location. For example, the word "EVOO" is the name of a restaurant in Somerville, MA, USA. The word "EVOO" appears only a few times in the entire corpus. Most of these times it appears on a page with the address for the restaurant. The other times, it appears on pages reviewing the restaurant. Given the strong correlation of "EVOO" with the restaurant's address, we can geocode the word "EVOO" with the same latitude/longitude point. This enables us to geocode the other pages with that same point. The latitude/longitude point is transmitted from one page to the other pages through the word link "EVOO."

Note that the spatial meaning inference 438 process is not usually able to associate a phrase with a location as focused as a point. Bounding polygons are a more common result. The main purpose of geocoding these phrases is to improve the ranking of documents, discussed in the section on spatial document ranking 45.

KEYWORD PARSER

Non-geographic search terms (keywords) are identified as follows. As the documents are saved to the page repository 222, a keyword parser 44 process opens each document and scans its keywords. These keywords are stored in a database table called word_instances 227, which includes the fields: wordID 2272, docID 2274, and word-doc relevance float 2276. The word_instances 227 table associates a given keyword with a set of documents containing it.

The WordID is a number that replaces the string of characters in the word. This reduces storage requirements and allows us to treat a phrase like "the big apple" as a single database entry. The word lexicon 225 is a database table that acts as the dictionary of all words and their corresponding WordIDs. The word lexicon 225 table includes the fields: word 22621; wordid 22623; and word_occurrences 22625.

The keyword parser 44 includes a tag recognizer 442 for parsing documents that contain tagged text such as SGML or the related standards HTML and XML. Tag recognizers for various document standards are well known in the computing art and can even be a feature of the operating system.

Methods standard in the art may be used to index a document for phrase searching, this allows a user to issue a query for a set of words close together or immediately adjacent in documents.

SPATIAL DOCUMENT RANKING

Given the potentially vast amount of information, document ranking is very important. Results relevant to the user's query must not be overwhelmed by irrelevant results, or the system will be useless.

The spatial document ranking 45 process produces a ranking of documents that includes evaluations of document-to-place relevance 452, document-to-word relevance 454, and abstract quality 456. Evaluations are combined into a floating point number indicating the relevance of each document to the query.

The document-to-place relevance 452 score indicates a document's relevance to a domain location, where the domain location is described by a PSI or KSI within the document. The following is a method of considering the relevance of one SI (spatial identifier, which might be a PSI or a KSI) to one document. It is possible to compute this for several different SIs in the same document. These SIs can be combined if they all refer to the same geographic region. For example, a document might have an address and a phone number that we can geocode. If the address is to a point nested inside the phone number's area, then we can improve the geographic relevance of the document to that address. The boost in relevance might be affected by handcrafted weights chosen for the different circumstances in which multiple SIs can combine on a page. This improvement is secondary to the relevance computed by the following method.

DOCUMENT-TO-PLACE RELEVANCE

The document-to-place relevance 452 score includes the following scores: position in page 4521, distance from end 4523, number of other SIs 4525, in sentence 4527, and emphasis 4529. (See Appendix A)

The position in page 4521 score is a heuristic function, calibrated from large numbers of observations of SIs. It assigns a score on the premise that SIs appearing earlier in a document are likely to be more relevant. Distance may be measured in characters or bytes. SIs that appear "above the fold" (on screen when a page is first loaded, without having to scroll) are considered most relevant.

The distance from end 4523 score gives the document-to-place relevance 452 score a slight boost if the SI occurs at the footer of the document; this partially counteracts the low score assigned to it by the position heuristic.

The number of other SIs 4525 score is a heuristic function that dilutes the relevance of a SI based on how many other SIs are in the same document. Documents with large numbers of addresses tend to be lists, where any individual address has a low probability of being relevant to the document.

The in sentence 4527 score gives a slight boost to SIs that are free-standing, as opposed to being mentioned in a sentence.

The emphasis 4529 score reflects the degree of emphasis of the SI text, including being in boldface, large type, or in the page's title. This score takes the form of a decimal number where 1.0 is assumed to be standard (neither de-emphasized or emphasized); lower numbers indicate lack of emphasis (such as small text) and higher numbers indicate prominence.

DOCUMENT-TO-WORD RELEVANCE

The document-to-word relevance 454 score indicates the relevance of a particular word to a particular document that contains it. Means for measuring the relevance of a word to a document are well known in the art. For instance, see S.E. Robertson and K. Sparck Jones, "Simple, proven approaches to text retrieval," University of Cambridge Computer Laboratory technical report, May 1997.

Phrase searching may also affect document relevance. This type of relevance is typically computed on-the-fly at the time of a user's query for a particular phrase. There are methods standard in the art for computing this type of relevance.

ABSTRACT QUALITY

The abstract quality 456 score represents document value independent of a given word or place. There are several ways to measure this, including the number of pages that link to the document, the number of times people click on the document when it is served as a search result, and the number of other documents that refer to the same words and places – that is, if it is a document like many others, its abstract value might be considered low, independent of the particular words it contains.

The abstract quality 456 score include components for network connectedness 4562 and a manual updates 4564. Network connectedness 4562 is computed from the probability that the page will be chosen by a random crawl of the web. This probability is then mapped to a score. The particular mapping chosen depends on the size of the document collection in the page repository 222, since the probability of finding any given document is inversely proportional to the collection size.

The manual updates 4564 score is designed to incorporate the input of human editors. The editors can craft rules that adjust the abstract quality 456 of particular documents. For example, they can weight all documents within a particular site as better than other documents simply by increasing their document quality measures. They might do this with a site that itself is a careful product of human editors, such as Zagat.com.

The abstract quality 456 score is stored in an abstract_document_quality 228 table, which includes the fields doc_id 2281 and document_quality 2283. The doc_id 2281 field is a foreign key referencing the doc_id 2221 field in the page repository 222.

INDEXER

The indexer 46 analyzes documents to prepare data structures that accelerate the search 50 process. The indexer 46 includes a spatial indexer 462, spatial-keyword indexer 465, and a tree degree converter 466.

SPATIAL INDEXER

With reference to Fig. 7, the spatial indexer 462 creates a spatial index 502 and a spatial document index 503 for a domain space. The spatial index 502 is a binary tree. The spatial document index 503 is a tree that is based on the spatial index 502, but may be of a higher degree than 2 (the degree of all binary trees).

The spatial indexer 462 in step 4621 gathers a collection of all domain locations referenced by a document in the page repository 222, then creates a root node for the spatial index 502 tree in step 4622. The spatial indexer 462 passes the root node and the collection to step 4624, which marks the beginning of the recursive spatial indexing subroutine (or RSIS) 4620.

In step 4624, the RSIS 4620 receives a node and a collection. The RSIS 4620 examines the collection in step 4625 to determine whether the collection contains more than one element. If it does not, the RSIS 4620 associates the current node with the one element's domain location in step 46295 and goes to step 4629, returning control to the routine that invoked it. Otherwise, the RSIS 4620 proceeds to step 4626, where the RSIS 4620 spatially divides the collection along spatial divider D into collections L and R, such that L and R are as equal in number as possible. If the domain space is a plane, the spatial divider D is a line in the plane. If the domain space is in three dimensions, the spatial divider D is a plane through 3-space. In general, if the domain space is of X dimensions, the spatial division is a manifold of dimension X minus one. The RSIS 4620 in step 4626 also stores the criteria for the spatial divider D in node N. Thus, each node contains criteria that divide a master collection of locations into two sub-collections.

The RSIS 4620 in step 4626 also creates a left node and a right node on the node passed to step 4624. This creates a fork in the binary tree that will act as an index. The tree as a whole becomes the spatial index 502.

The RSIS 4620 becomes recursive by invoking itself on each of the sub-collections. Specifically, in step 4627 the RSIS 4620 passes sub-collection L and the current left node to step 4624, while in step 4628 the RSIS 4620 passes sub-collection R and the current right node to step 4624. The RSIS 4620 repeats until every collection has been divided into collections of single elements, which are associated with childless nodes. All other nodes have division criteria and two nodes descending from them.

After the spatial indexer 462 builds the spatial index 502 tree, which indexes the points referred to in a corpus of documents, the spatial indexer 462 builds the spatial document index 503 by extending a copy of the spatial index 502 tree to cover multiple documents that refer to the same spatial point. The spatial indexer 462 invokes a tree degree

converter 466 to make a version of the spatial index 502 that is represented in a tree of degree k.

The extension of the spatial index 502 produces new branches that no longer reflect spatial divisions but instead reflect partitions of the documents referring to that point. In particular, instead of the nodes including criteria that define spatial divisions within the domain (as the nodes inherited from the spatial index 502 continue to do), the nodes added after the extension include criteria for branching within the space of the docID 2221 numbers of the documents. Partitioning based on a key value (such as the docID 2221) of a database table is standard in the art. Such a partitioning produces a k-way tree on the documents using their docID 2221 numbers as a key.

DEGREE K

An important optimization of an index tree of degree k involves the selection of k. A k-way branching structure must be chosen before building or storing the trees. K could be as low as two and as high as a few thousand or tens of thousands, depending on the number of documents and possibly the underlying computing platform. A tree of degree k can index $(k)^L$ documents in L levels.

A large value for k makes it faster and more storage efficient to deal with keywords that appear in only a few documents. If the number of rare words in the page repository 222 is large, a large value for k is more storage efficient than a smaller one. However, a smaller value of k can be more search efficient, as it allows a traversing process (in response to a query) to ignore branches of the tree that fail its constraints.

The selection of k is an empirical process that may be performed for every set of documents, in the page repository 222, to be indexed. It is influenced by hardware limits, such as the number of bits handled by a single processor instruction and the number of blocks loaded by the disk drive. The most important factor in choosing k is the word-frequency distribution. Keyword lexicons for web pages, for instance, show a huge number of words that appear in only one or two documents, but more common words appear in many documents. These common words produce a "fat tailed" distribution. The exact shape of the distribution for a particular set of documents determines the optimal k. Given a value for k,

it is a simple calculation to count the number of bytes used to store the word trees for a particular lexicon and set of documents.

TREE DEGREE CONVERTER

The tree degree converter 466 is a function that accepts parameters including a binary tree and an integer k, and returns as its output a tree of degree k incorporating the structure and data of the binary tree. Methods for this conversion are known in the computing art.

SPATIAL-KEYWORD INDEXER

The spatial-keyword indexer 465 builds a spatial-keyword document index 505 responsive to queries for documents. The queries can have keyword criteria, spatial criteria, or both.

The spatial-keyword indexer 465 gathers all domain locations referenced by documents in the page repository 222.

The spatial-keyword indexer 465 uses the spatial document index 503 generated by the spatial indexer 462. The spatial document index 503 is a k-way tree on this list of documents. The spatial-keyword indexer 465 copies the spatial document index 503 to create a keyword tree 506 for every keyword. For each keyword tree 506, the spatial-keyword indexer 465 trims away all documents that do not contain that particular keyword. If, after the document trimming, the subtree depending from a node of the keyword tree 506 does not contain a document, the spatial-keyword indexer 465 removes that node (and therefore its subtree).

The spatial-keyword indexer 465 creates for each keyword a minimal keyword tree 506 that relates the keyword to the corpus of documents in the page repository 222. Furthermore, the spatial-keyword indexer 465 ensures that one branching structure is common to all keyword trees as well as to the spatial document index 503 tree.

SEARCH

With reference to Fig. 5, the search 50 process responds to queries with a set of documents ranked by relevance.

A lexical tree 508 is any copy of the spatial document index 503 tree, possibly trimmed. Thus, every keyword tree 506 is a lexical tree 508, as is the spatial document index 503 tree itself. Also, any filter can be expressed as a lexical tree 508, since a filter determines a set of documents, and any set of documents determines a trimming of the spatial document index 503 tree. Thus, lexical trees 508 can be built to index arbitrarily complex sets of documents.

The search 50 process uses the spatial document index 503 and spatial-keyword document index 505 to find documents that refer to a given set of domain locations or regions, and documents related to a given set of keywords exisiting in the word lexicon 225. The search 50 process can also find documents using a lexical tree 508, such as might represent a filter. Thus, the search 50 process can respond to queries that seek documents according to spatial domain criteria, keyword criteria, filters, or any combination thereof. Furthermore, the search 50 process can invoke the document ranker 56 process to rank the result set of documents by relevance to the query terms.

The search 50 process answers queries via the procedure in Fig. 6. A query includes at least one of the following: a bounding region specifying a closed shape (typically a polygon in two dimensions), words, phrases, and layers. The bounding region can be the domain frame from the map interface 80.

For each element in the query, the search 50 process loads a copy of the appropriate tree, determined as follows. If a bounding region is specified, step 703 loads the spatial document index 503. If keywords are specified, step 702 load the spatial-keyword document index 505 tree for each keyword. If a phrase is specified and the phrase is not a single entry in the word lexicon 225, then step 702 loads each word's spatial-keyword document index 505. If a phrase is specified and is a single entry in the word lexicon 225, then step 702 need only load that phrase's spatial-keyword document index 505. If a layer is specified, its name identifies the appropriate lexical tree 508, which is loaded by step 702.

The search 50 process counts the number of leafs of each of these trees. At step 703, the search 50 process estimates the approximate number of leafs in the spatial document index 503 bounded by the query bounding region, by multiplying the area of the bounding

region by the average density of points in the corpus. At step 704, these numbers are used to order the trees in a list, with the smallest tree first.

At step 705, this smallest tree is re-labeled as the result tree and will be trimmed to generate the final result tree. For each node that exists in the partially trimmed result tree, the search 50 process checks all the trees to see if they also contain that node. In steps 708 and 712, the search 50 process checks the trees in list order. If any tree lacks that node, the search 50 process stops checking and in step 709 deletes the subtree below that node in the result tree. (See Appendix B) Steps 710 and 711 traverse the tree. The search 50 process continues checking all the nodes in the result tree until only leaf nodes remain. These leaf nodes represent the result set of documents. Step 713 returns the result tree.

The leafs of spatial-keyword document index 503 trees have word relevances and lists of the positions and contextual emphasis of the words in each document. The spatial document index 503 has spatial relevances for each document. The lexical tree 508 for each layer may have an abstract document quality 456 for some documents. These relevances are combined for each document in the result set. The combination procedure might be averaging, summing, or a weighted average.

A second process might compute adjustments to the document relevances by considering the emphasis and proximity of multiple query words within the documents. This standard procedure simply gives higher relevance to documents in which the query words appear closer together.

The final result list of documents might be sorted for return to the user. The sorting procedure might extract only a portion of the documents with the highest relevance.

DOCUMENT RANKER

Document ranker 56 combines various relevance scores for each document in a result set and sorts the documents by this combined relevance. The combination function may be an averaging or a weighted sum or some other combining function tailored to the various relevance scores used. The document ranker 56 may take streams of sorted result sets from several database systems and merge sort them to produce a new result set.

ICON RANKER

The icon ranker 57 receives a sorted list of results from the document ranker 56. To present this list to the user who requested the documents, the icon ranker 57 aggregates overlapping icons according to the manner described in the section on Icons. This list of aggregated icons is presented to the user with sublists next to each icon 810. These sublists identify the documents aggregated into that icon 810.

The icon ranker 57 groups documents into icons 810 as follows. The icon ranker 57 takes the first document from the sorted result list and makes it the first icon 810 in the icon list. For each subsequent document having a tentative display location in the result list, the icon ranker 57 examines whether an icon 810 situated at the tentative display location would collide with any icon 810 already in the icon list. If a collision occurs, the icon ranker 57 associates the colliding document with the existing icon. If no collision occurs, the icon ranker 57 adds an icon 810 to the icon list and associates the current document with said icon 810. This procedure may terminate whenever the number of icons reaches the lesser of a maximum number determined by the user or a predetermined number that is a customizable operating parameter of the computer system 20.

If a document is topically affiliated with a particular icon class, the icon ranker 57 assigns an icon face 818 from said icon class to the icon 810 that will represent the document. If multiple icon classes are affiliated with documents represented by a single icon 810, the icon ranker 57 may select one of the said icon classes to assign to said icon 810 or may assign a new icon class built to reflect said multiple icon classes.

USER PROFILES

The user profiles 65 process manages information specific to user accounts. The information may include descriptions of how users have interacted with the computer system 20 in the past. Other elements that might be recorded include default location to display to the user when beginning an interaction, set of previously collected layers, set of previously posted note documents, previous searches, and previous click patterns or behavior. Part or all of this information may be made directly viewable and editable by the user.

The user profiles 65 process also allows a user to log into the computer system 20 with a user name and possibly a password. The user name identifies the user with a user

account, as is common in the art. The map interface 80 can include account login entry controls 861, including a prompt for account login 862, a data entry control 863, and a submission control 864.

DATA PRESENTATION

The data presentation 60 process manages the state of the map interface 80 for each user session. As the user changes the state of the map interface 80 – for instance, by issuing queries, selecting controls, and generally utilizing the interface tools – the data presentation 60 system keeps track of these changes and their sequence. This recorded history enables querying within previous result sets. For example, a user can query for documents referring to "shoes" in "cambridge, ma," and in a subsequent interaction, the user can filter this set of documents further by requesting only those documents that contain the word "store." This results in a list of documents containing "shoes" and "store" and referring to "cambridge, ma." The user could then zoom out to see a larger region with these document still displayed in the map. To see new documents that might fit the keyword query in this larger domain, the user can re-issue the query.

Similarly, the user could combine a set of documents with another set of documents selected by a different query.

Any number of subsequent filter operations or result set combinations can be performed, limited only by the storage resources of the computer system 20 or, optionally, by parameters built into the computer system 20, as for performance reasons. The data presentation 60 system keeps track of filter operations by a given user so that the computer system 20 can present the correct set of documents to the user at each subsequent interaction.

SERVICE COLLECTOR

The service collector 24 includes a proxy through which the user interface server 62 and the portal server 70 communicate with the processes for data presentation 60, search 50, and user profiles 65.

PORTAL SERVER

The computer system 20 includes a portal server 70 process. The portal server 70 offers at least some of the services of the computer system 20 through remote procedure calls

and other network protocols. This allows the services, data, and tools of the computer system 20 to be delivered through public portal systems or directly to individuals. Examples of companies offering public portal systems include Yahoo! Inc. of 3420 Central Expressway, Santa Clara, CA, and Sprint PCS of PO Box 8077, London, KY.

ALTERNATE EMBODIMENTS

A number of embodiments of the invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. Accordingly, other embodiments are within the scope of the following claims.

APPENDIX B

to test the existence of a node address N in a tree T:

```
if (T is a keyword/layer tree) {
 nodeQ = pointer to root node of T;
 foreach $step in node address N {
  next nodeQ = pointer to child number $step from previous nodeQ;
  if ( nodeQ is a valid child ) {
    continue foreach loop;
  } else {
    exit with return value = "false";
  }
 }
 return "true";  # the loop exited without hitting a nonexistent child
}

if (T is the spatial tree) {
 in the following, polygonP is the bounding region given by the user;
 nodeQ = pointer to root node of T;
 foreach $step in node address N {
  next nodeQ = pointer to child number $step from previous nodeQ;
  if ( region below nodeQ overlaps polygonP) {
  } else {
    exit with return value = "false";
  }
 }
 return "true";  # the loop exited without hitting a division outside
                 # the query's bounding region
}
```

APPENDIX A

psuedocode excerpt for assigning relevance of document to place

```
# Heuristically-determined parameters:

# $emphasis_bonus_modifier determines the importance of the emphasis bit.

# $sentence_penalty_modifier determines the importance of the in_sentence
# bit

# $sp_full_point: the position after which the sentence penalty fully
# applies

# $sp_transition_point: the position after which the sentence penalty
# starts to apply; it goes from 0 at this position to
# $sentence_penalty_modifier at $sp_full_point

# $end_bonus_size: maximum number of characters from the end of the
# document at which the end-of-document bonus applies

# $end_bonus_max: the maximum relevance value for which the end-of-
# document bonus applies

# $end_bonus_multiplier determines the weight of the end-of-document bonus


# Start with the position heuristic function. This is a nonincreasing
# function which is normalized to 1 for position 0. It decreases slowly up
# to some position p_f which is the average position of the "fold", that
# is, the place where the end of the visible area of a typical document
# occurs when it is first displayed to a user. For positions
```

```
# greater than p_f it decreases more quickly, but levels off for large
# positions. The exact form is determined heuristically by manually
# assigning a score to a large number of instances of PSIs in typical
# documents and fitting a function to these scores.

$relevance = &position_function($pos);


# Bonus for being bold, large font, in title, etc. $emphasis is a
# heuristic function of the PSI which was assigned on how emphasized it
# is.

$emphasis_bonus = $emphasis_bonus_modifier * $emphasis;

# Penalty for being in a sentence, e.g. "We would like to announce the
# availability of several of our products through the Hopkinton Drug
# Store, 52 Main Street, Hopkinton, MA 01748."
# No penalty is assigned for PSIs in the first $sp_transition_point
# characters, going up to the full penalty after $sp_full_point
# characters.

if ($pos > $sp_full_point) {
 $sentence_penalty = $sentence_penalty_modifier * in_sentence;
} else {
 if ($pos > $sentence_penalty_transition_point) {
  $sentence_penalty = $in_sentence * $sentence_penalty_modifier *
                      (($pos-$sp_transition_point)/
                    $sp_full_point-$sp_transition_point);
 } else {
  $sentence_penalty = 0.0;
 }
```

```
}

$relevance += $emphasis_bonus - $sentence_penalty;


# Bonus for being at end of document for long documents. It is
# proportional to
# how low the relevance already is, so that already highly
# scoring PSIs don't receive a bonus for being at the end.
# This is before the number of PSIs function so that it will be
# depressed by that function (and the last PSI in a big list won't
# score too high.)

if ($size - $pos < $end_bonus_size && $relevance < $end_bonus_max) {
  $relevance += ($end_bonus_max - $relevance) * $end_bonus_multiplier;
}


# Now depress the above score based on how many other PSIs
# appear on the page.
# num_psi_function($num) is a function which determines how much less
# valuable a PSI is when it occurs together with other PSIs.
# It is nonincreasing, and is one for $num = 1; it decreases
# quickly for small $num, and more slowly for large $num.
# This function is determined heuristically as described above for the
# position function.

$relevance *= &num_psi_function($num);
```

WHAT IS CLAIMED IS:

1. An interface program stored on a computer-readable medium for causing a computer system with a display device to perform the functions of:

accepting search criteria from a user including a free text entry query and a domain identifier identifying a domain;

in response to accepting said search criteria from the user, retrieving a plurality of record identifiers each of which identifies a corresponding record which: (1) has associated therewith a location identifier that locates it at a specific location within the domain identified by the domain identifier; and (2) contains information that is responsive to the free text entry query;

displaying a representation of said domain on the display device; and

displaying on the display device a plurality of icons as representations of the records identified by said plurality of record identifiers, wherein for each of said plurality of record identifiers, a corresponding one of the plurality of icons is displayed within said representation of the domain that is being displayed on the display device, the corresponding icon for each of said plurality of record identifiers being positioned within the representation of the domain at a coordinate within the domain that corresponds to the location identifier for the corresponding record.

2. The interface program of claim 1 wherein the domain is a geographical region and said representation is a multi-dimensional map of the geographical region.

3. The interface program of claim 1 wherein said representation is a two-dimensional map of the geographical region.

4. The interface program of claim 2 wherein accepting input further comprises accepting a designation by the user of a designated category and wherein each of the records corresponding to the plurality of retrieved record identifiers also includes information that falls within the designated category.

5. The interface program of claim 4 wherein accepting said designation by the user of a category comprises presenting to the user a list of predefined categories and accepting as the designated category a selection by the user from that list.

6. The interface program of claim 3 for causing the computer to perform the further functions of:

after displaying the corresponding icon for each of the plurality of record identifiers, accepting further search criteria from the user, said further search criteria selected from the group of search criteria types consisting of a domain identifier input type, a free text entry query input type, and a category type;

in response to accepting said further search criteria from the user, retrieving a subset of said plurality of record identifiers, wherein said subset of said plurality of record identifiers identifies all record identifiers among said plurality of record identifiers that fall within said further search criteria;

displaying a two-dimensional map of a revised geographical region on the display device that is responsive to said further search criteria; and

for each of the record identifiers of said subset of plurality of record identifiers, displaying a corresponding icon within said displayed map, the corresponding icon for each of the record identifiers of said subset of said plurality of record identifiers being positioned within the displayed map at a coordinate that corresponds to the location identifier for the corresponding record.

7. The interface program of claim 6 for causing the computer to perform the further functions of storing as a filter the first-mentioned search criteria in combination with said further search criteria, wherein said stored filter is retrievable for use by the user in specifying a future search through the interface.

8. The interface program of claim 7 wherein the first-mentioned search criteria in combination with said further search criteria is an ordered sequence of inputs and wherein the stored filter is the ordered sequence of inputs preserving the order of the sequence of inputs.

9. The interface program of claim 3 for causing the computer to perform the further functions of:

presenting to the user via the display device a map; and

enabling the user to input said domain identifier as part of the search criteria by interacting with the displayed map.

10. The interface program of claim 3 wherein said plurality of icons include an icon of a first icon class and an icon of a second icon class, wherein the icon of the first icon class has first visual characteristics and the icon of the second icon class has second visual characteristics that are different from the visual characteristics associated with the first icon class.

11. The interface program of claim 10 wherein at least some of the records identified by said plurality of record identifiers are of a first type and at least some of the other records identified by said plurality of record identifiers are of a second type and wherein records of the first type are displayed using the icon of the first icon class and records of the second type are displayed using the icon of the second icon class.

12. The interface program of claim 3 wherein at least one of the icons of the plurality of icons represents multiple of the records identified by said plurality of record identifiers, each of said multiple of the records having a location identifier that locates that record within a neighborhood about a central location.

13. The interface program of claim 3 for causing the computer to perform the further functions of:

accepting a change of scale request from the user;

in response to accepting said change of scale request, consolidating at least some of said plurality of icons with each other to form a second plurality of icons that is fewer in number than the number of icons in said first-mentioned plurality of icons; and

in response to accepting said change of scale request, redisplaying said domain using a decreased scale and also using the second plurality of icons to identify the locations of the records identified by said plurality of record identifiers.

14. The interface program of claim 3 for causing the computer to perform the further functions of:

accepting from the user a specification of an electronic note which has an associated location within the map; and

displaying a sticky-note icon on the map at position that corresponds to the associated location.

15. The interface program of claim 14 wherein the electronic note comprises a web page with its own externally accessible address which enables people to electronically access its contents through that address.

16. A database system stored on a computer-readable medium for causing a computer system to perform the functions of:

accepting search criteria including at least one of: (1) text; (2) a domain identifier identifying a domain; and (3) a filter identifier identifying a filter; and

retrieving a plurality of record identifiers each of which identifies a corresponding record which has associated therewith the text, domain identifier, or layer identifier of the search criteria, where the retrieving is performed with a spatial-keyword document index.

17. A method of retrieving a plurality of record identifiers each of which identifies a corresponding record which has associated therewith at least one of a text, a domain identifier, and a layer identifier as specified by search criteria, wherein such retrieving is performed with a spatial-keyword document index.

18. The method of claim 17 wherein the spatial-keyword document index comprises a spatial index tree extended to reference documents and a plurality of trees with the same structure as said spatial index tree but trimmed for particular lexicon entries and filters.

19. The method of claim 17 wherein said plurality of record identifiers are retrieved in a spatial-keyword document index tree and the branching structure of said tree is analyzed to identify geographic phenomena.

20. The method of claim 19 wherein a geographic phenomena is a spatial-keyword document index tree branching structure in which more than a predetermined fraction of the branches share more than predetermined number of parent nodes.

21. A program stored on a computer-readable medium for causing a computer system to perform the functions of:

loading documents referred to by document addresses;

parsing those documents for more document addresses to load; and

parsing those documents for possible spatial identifiers.

22. The program of claim 21 for causing the computer system to perform the further function of analyzing the possible spatial identifiers to determine a location in a domain.

23. The program of claim 21 wherein a portion of the document addresses are gathered by a metasearcher process, which queries other computer systems using text that refers to spatial domains.

24. The program of claim 21 for causing the computer system to perform the further function of computing a relevance score for each of the the possible spatial identifiers found in each document.

25. The program of claim 21 wherein the relevance score comprises one or more of:

(1) the possible spatial identifier's position in the document;

(2) the number of other possible spatial identifiers in the document;

(3) whether the possible spatial identifier is in a sentence or is free standing; and

(4) the formatted emphasis of the characters in the possible spatial identifier.

26. The program of claim 21 for causing the computer system to perform the further function of sorting the document addresses by relevance score before loading.

27. A program stored on a computer-readable medium for causing a computer systen to perform the functions of:

loading documents referred to by document addresses;

parsing those documents for more document addresses to load;

parsing those documents for possible spatial identifiers; and

analyzing the possible spatial identifiers to determine a location in a domain.

28. A method of displaying spatially coded information, comprising:

through an automated computer process, gathering documents in a database;

selecting a subset of the documents which can be determined to contain spatial information;

associating at least one spatial identifier with each document in the subset;

indexing the documents, the indexing comprising an index on spatial identifiers and an index on keywords;

providing a computer interface through which a user can submit a query comprising spatial information;

responding to the query with a result set comprising documents; and

displaying the result set to the user through the computer interface.

29. The method of claim 28, wherein the result set, when it contains more than one element, comprises a plurality of groups organized by spatial proximity, each group containing at least one document of the result set.

30. The method of claim 29 wherein the plurality of groups is ordered according to a predetermined function on groups representing relevance to the criteria.

31. The method of claim 29, wherein the content of each group is ordered according to a predetermined function on elements representing relevance to the criteria.

32. The method of claim 28, wherein the criteria include keywords.

33. A method for populating a spatial document database with hyperlinked documents containing spatial information, the method comprising:

providing a destination database containing potential sources of gatherable documents;

providing a history database of known sources where documents have been gathered;

providing a crawler computer process which can follow a hyperlink in a document to access a potential source of gatherable documents specified by the hyperlink;

bootstrapping the crawler;

iterating the crawler over the destination database, including the steps of:

moving a potential source of gatherable documents from the destination database to the history database;

inspecting the potential source for gatherable documents;

storing any such gatherable documents in the spatial document database; and

adding to the destination database all potential sources of gatherable documents which are referenced by a hyperlink in the gatherable documents.

34. The method of claim 33, wherein the bootstrapping comprises

providing a plurality of locations of known interest;

providing the destination database with a plurality of metasources, each metasource being a source of potential sources of gatherable documents, and each metasource responding to queries by the computer process with a result set comprising potential sources of gatherable documents; and

priming the destination database by repeatedly running a primer process comprising:

formulating a query based on at least one of the locations of known interest:

directing the query to a metasource;

receiving a result set for the query; and

inserting each element of the result set into the destination database.

35. A method for scoring the relevance to a query of a document containing a known spatial identifier, the method comprising:

scoring the document for its relevance to a place specified in the query;

scoring the document for its relevance to a word specified in the query;

scoring the document for its quality; and

combining the scores to form a single score.

36. A method for indexing a plurality of documents to enable queries comprising keywords and spatial information, the method comprising:

initializing a master spatial tree of predetermined degree as a computer data structure, such that: (1) each leaf node of the master spatial tree represents a document; (2) each non-leaf node of the master spatial tree represents a range of space; (3) a root node of the master spatial tree represents a range of space encompassing any space that a valid query may refer to; and (4) the collection of non-leaf child nodes, relative to a parent node, define a partition on the range of space represented by the parent;

recursively adding a child node to the tree.

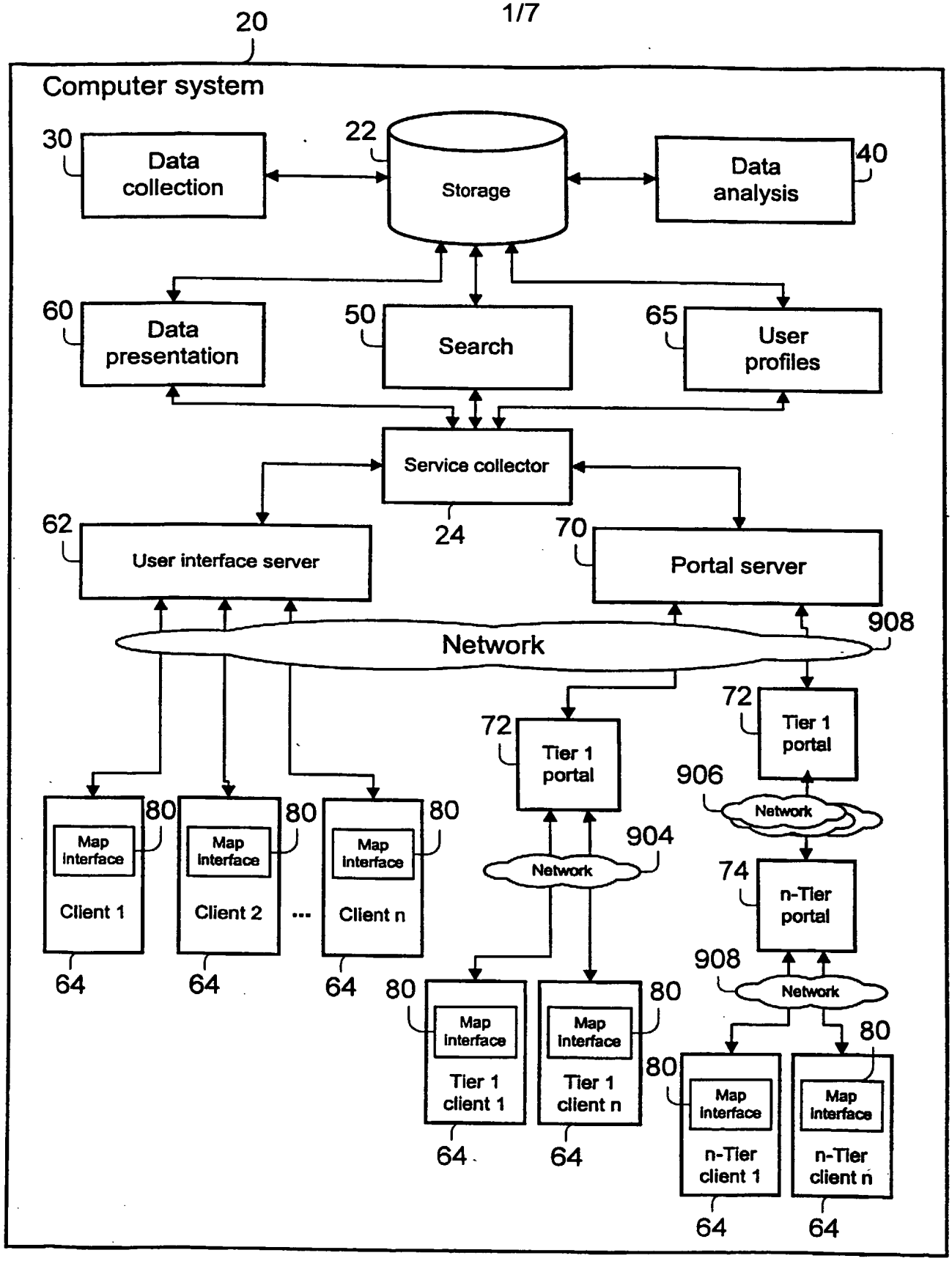

20
Computer system
30
Data collection
22
Storage
40
Data analysis
60
Data presentation
50
Search
65
User profiles
Service collector
24
62
User interface server
70
Portal server
908
Network
72
Tier 1 portal
72
Tier 1 portal
906
Network
904
Network
74
n-Tier portal
908
Network
80
Map Interface
Client 1
Client 2
Client n
64
Tier 1 client 1
Tier 1 client n
n-Tier client 1
n-Tier client n

FIG. 1

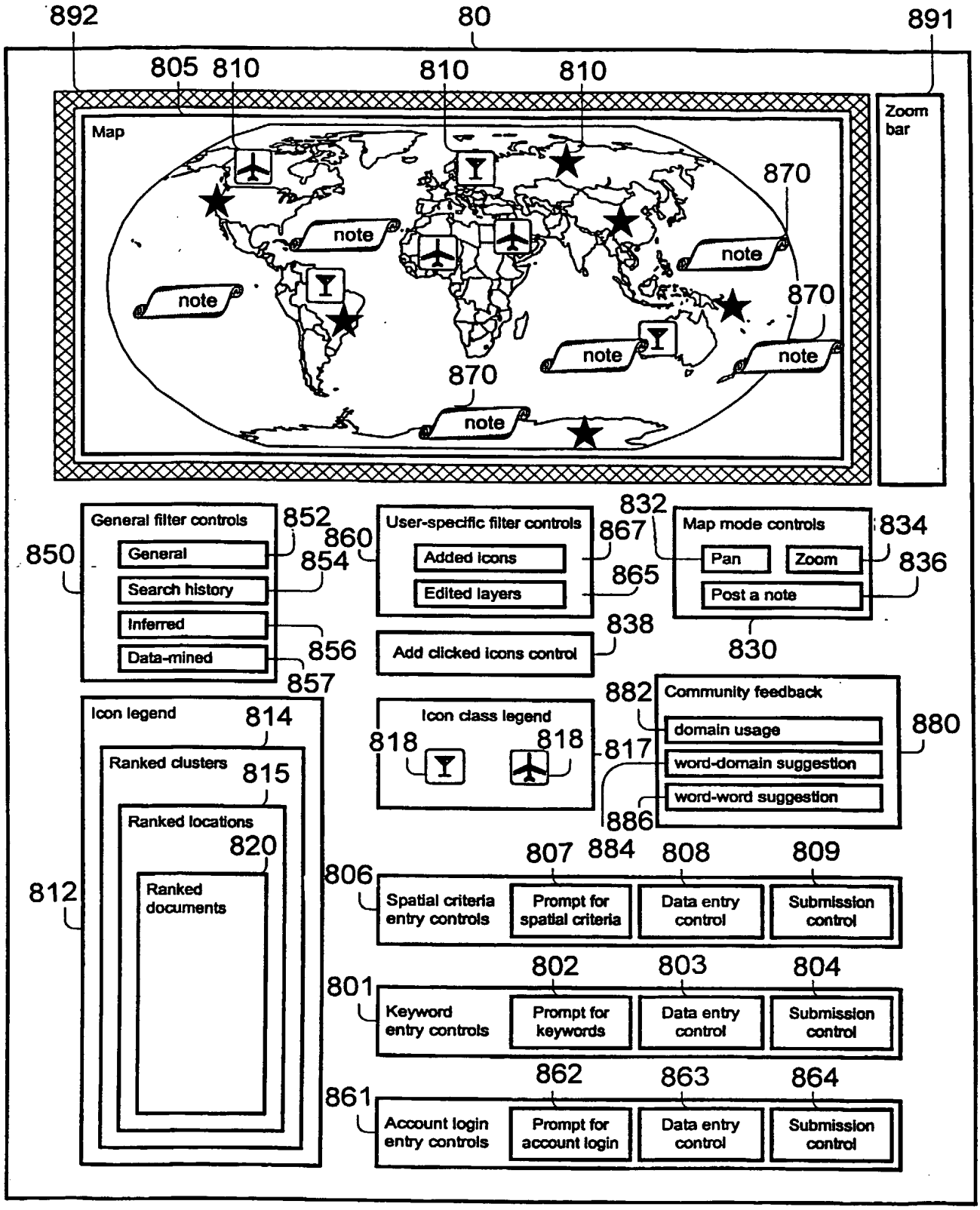

892
80
891
805
810
810
810
Map
Zoom bar
870
note
note
note
870
note
note
870
note
850
General filter controls
852
General
Search history
854
Inferred
Data-mined
856
857
860
User-specific filter controls
Added icons
Edited layers
865
867
832
Map mode controls
834
Pan
Zoom
836
Post a note
830
838
Add clicked icons control
Icon legend
814
Ranked clusters
815
Ranked locations
820
812
Ranked documents
Icon class legend
818
818
882
Community feedback
880
domain usage
817
word-domain suggestion
word-word suggestion
886
807
884
808
809
806
Spatial criteria entry controls
Prompt for spatial criteria
Data entry control
Submission control
802
803
804
801
Keyword entry controls
Prompt for keywords
Data entry control
Submission control
862
863
864
861
Account login entry controls
Prompt for account login
Data entry control
Submission control

FIG. 2

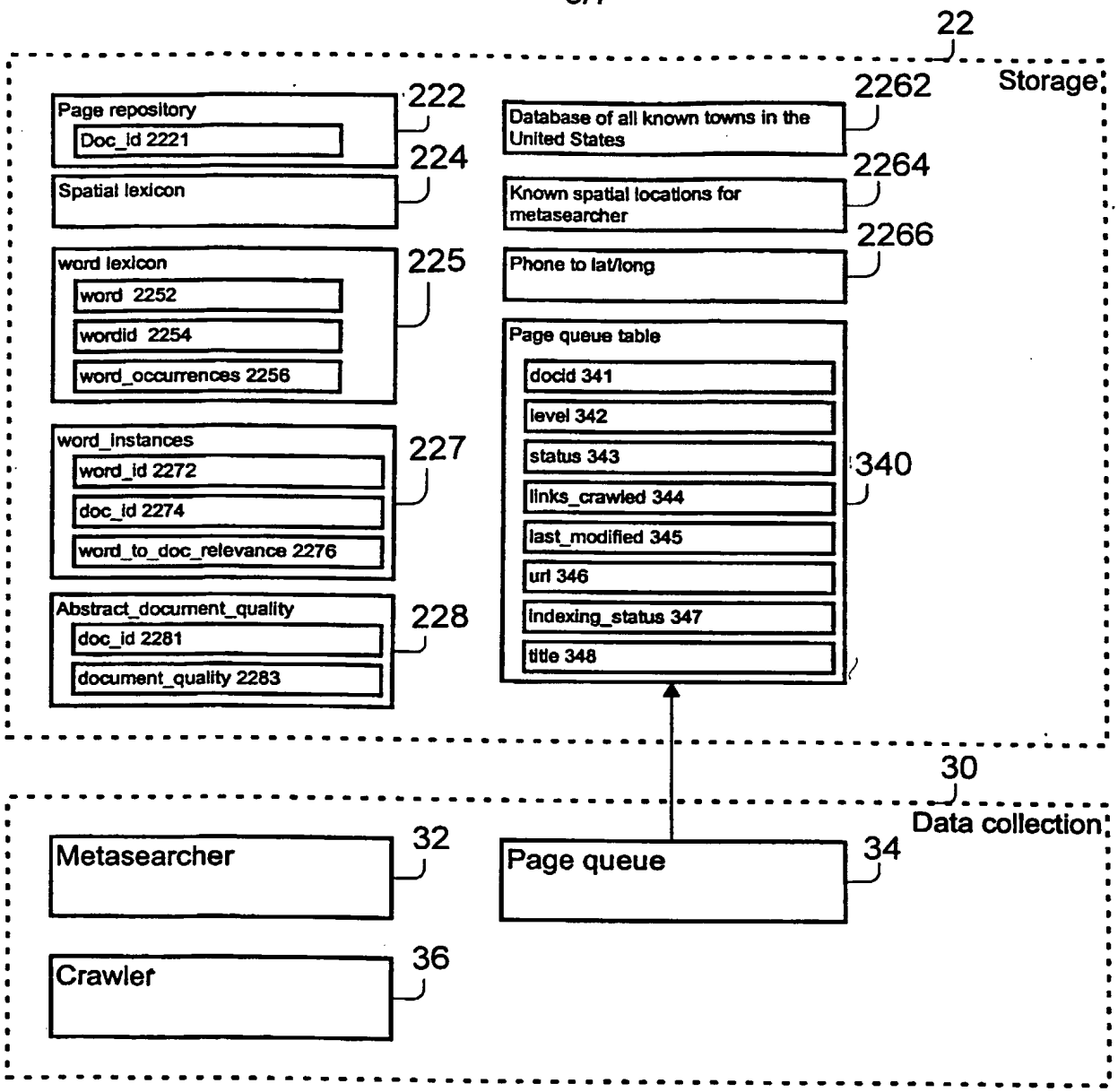

22
Storage
Page repository
Doc_id 2221
222
Spatial lexicon
224
word lexicon
word 2252
wordid 2254
word_occurrences 2256
225
word_instances
word_id 2272
doc_id 2274
word_to_doc_relevance 2276
227
Abstract_document_quality
doc_id 2281
document_quality 2283
228
Database of all known towns in the United States
2262
Known spatial locations for metasearcher
2264
Phone to lat/long
2266
Page queue table
docid 341
level 342
status 343
links_crawled 344
last_modified 345
url 346
indexing_status 347
title 348
340
30
Data collection
Metasearcher
32
Page queue
34
Crawler
36

FIG. 3

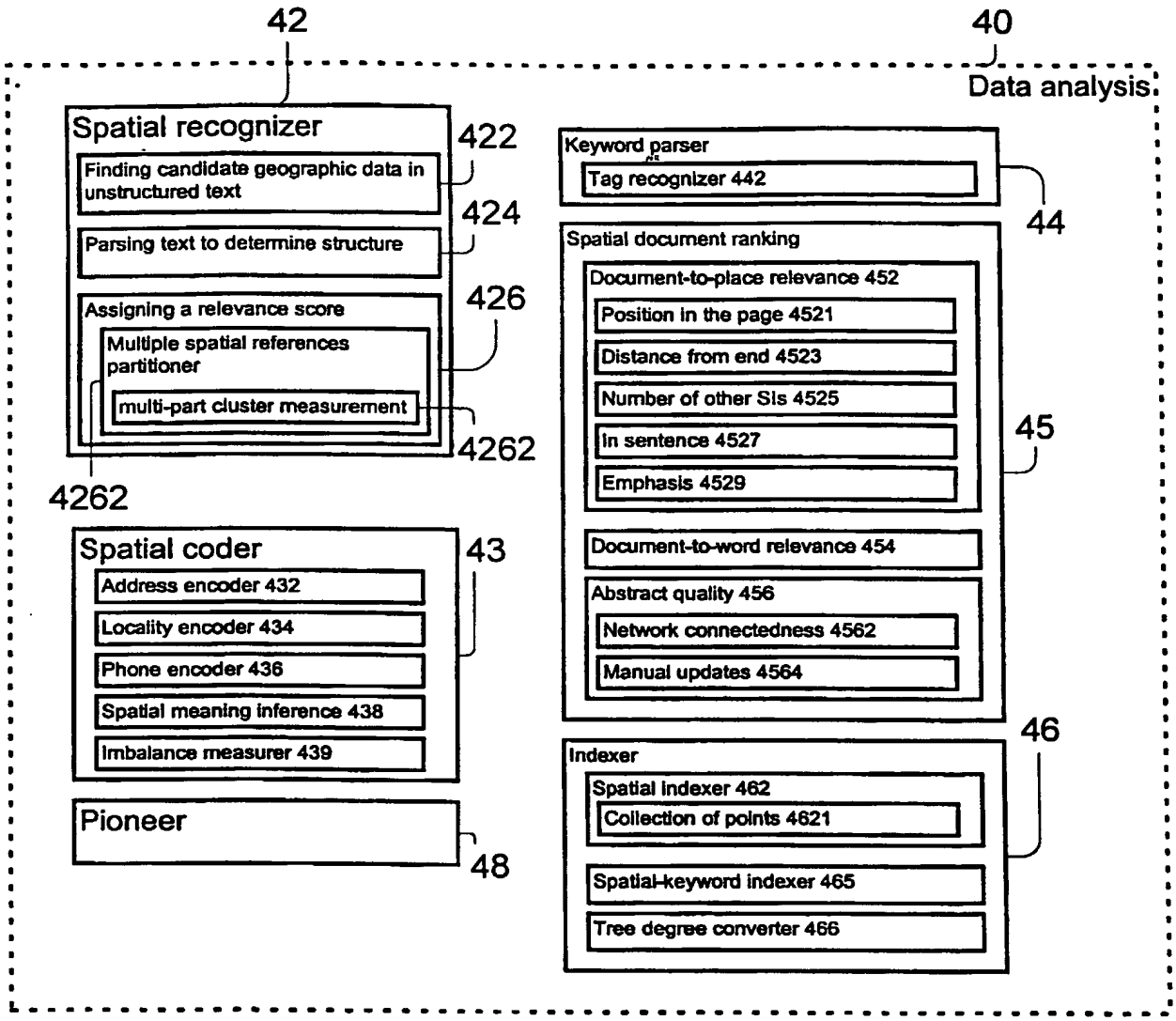

42
40
Data analysis
Spatial recognizer
422
Finding candidate geographic data in unstructured text
424
Parsing text to determine structure
426
Assigning a relevance score
Multiple spatial references partitioner
multi-part cluster measurement
4262
4262
Spatial coder
43
Address encoder 432
Locality encoder 434
Phone encoder 436
Spatial meaning inference 438
Imbalance measurer 439
Pioneer
48
Keyword parser
Tag recognizer 442
44
Spatial document ranking
Document-to-place relevance 452
Position in the page 4521
Distance from end 4523
Number of other SIs 4525
In sentence 4527
Emphasis 4529
45
Document-to-word relevance 454
Abstract quality 456
Network connectedness 4562
Manual updates 4564
46
Indexer
Spatial indexer 462
Collection of points 4621
Spatial-keyword indexer 465
Tree degree converter 466

FIG. 4

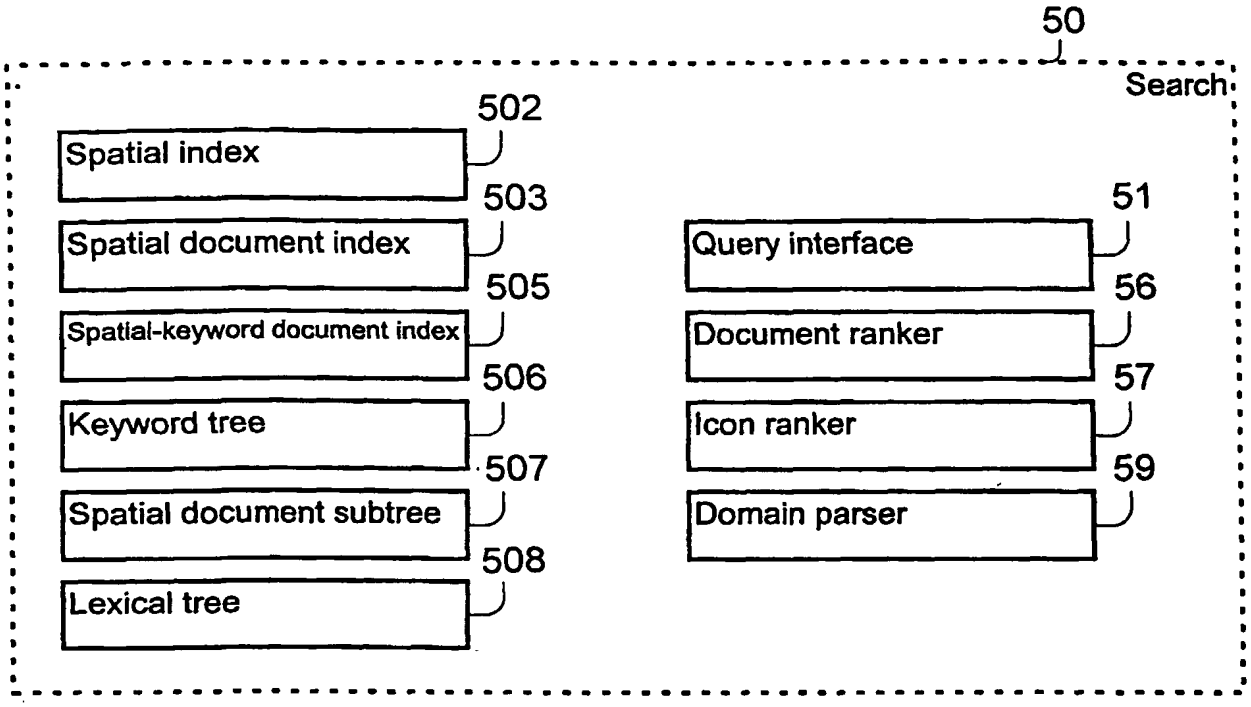

50
Search
502
Spatial index
503
Spatial document index
505
Spatial-keyword document index
506
Keyword tree
507
Spatial document subtree
508
Lexical tree
51
Query interface
56
Document ranker
57
Icon ranker
59
Domain parser

FIG. 5

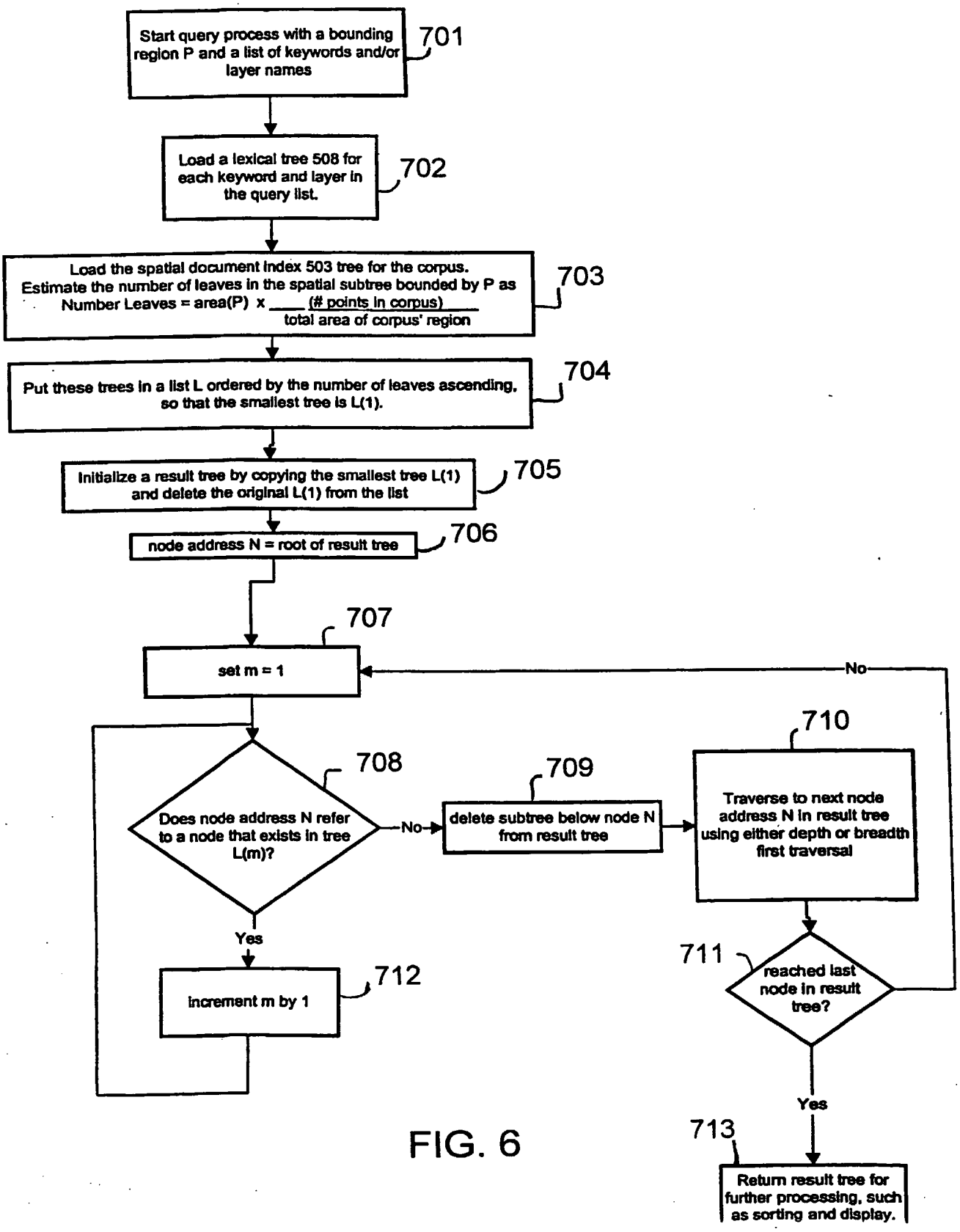

Start query process with a bounding region P and a list of keywords and/or layer names
701
Load a lexical tree 508 for each keyword and layer in the query list.
702
Load the spatial document index 503 tree for the corpus. Estimate the number of leaves in the spatial subtree bounded by P as Number Leaves = area(P) x (# points in corpus) / total area of corpus' region
703
Put these trees in a list L ordered by the number of leaves ascending, so that the smallest tree is L(1).
704
Initialize a result tree by copying the smallest tree L(1) and delete the original L(1) from the list
705
node address N = root of result tree
706
707
set m = 1
No
710
708
Does node address N refer to a node that exists in tree L(m)?
No
709
delete subtree below node N from result tree
Traverse to next node address N in result tree using either depth or breadth first traversal
Yes
711
reached last node in result tree?
712
increment m by 1
Yes
713
Return result tree for further processing, such as sorting and display.

FIG. 6

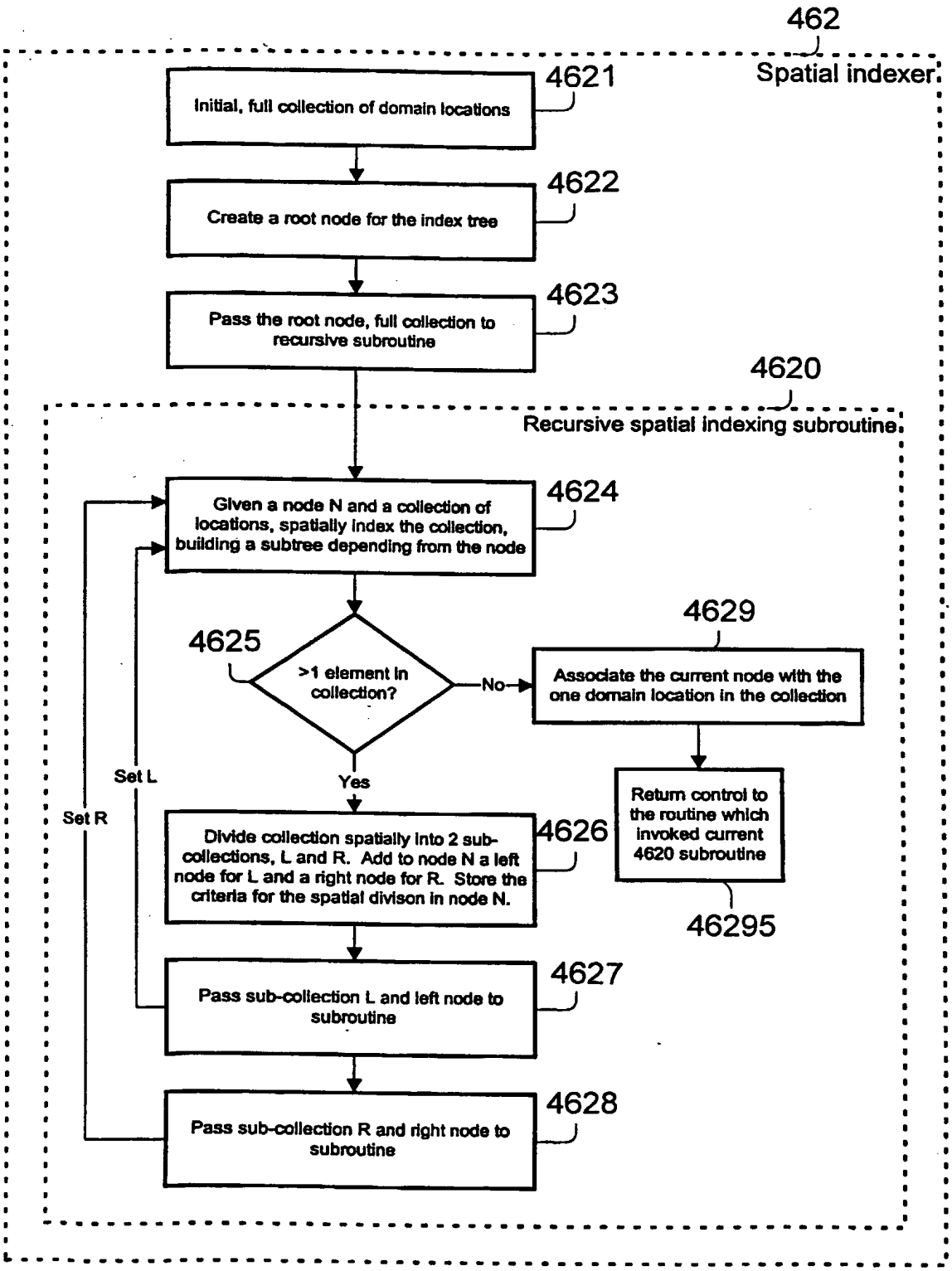

462
Spatial indexer
4621
Initial, full collection of domain locations
4622
Create a root node for the index tree
4623
Pass the root node, full collection to recursive subroutine
4620
Recursive spatial indexing subroutine
4624
Given a node N and a collection of locations, spatially index the collection, building a subtree depending from the node
4625
>1 element in collection?
No
4629
Associate the current node with the one domain location in the collection
Return control to the routine which invoked current 4620 subroutine
46295
Yes
Set L
Set R
4626
Divide collection spatially into 2 sub-collections, L and R. Add to node N a left node for L and a right node for R. Store the criteria for the spatial divison in node N.
4627
Pass sub-collection L and left node to subroutine
4628
Pass sub-collection R and right node to subroutine

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/40173

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 17/30
US CL : 707/4, 104, 1-3, 5-8, 10; 345/348

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 707/4, 104, 1-3, 5-8, 10; 345/348

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,991,781 A (NIELSEN) 23 November 1999 (23.11.1999), ALL.	1-36
A	US 5,920,856 A (SYEDA-MAHMOOD) 06 July 1999 (06.07.1999), ALL.	1-36
A	US 5,802,361 A (WANG et al) 01 September 1998 (01.09.1998), ALL.	1-36

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 June 2001 (20.06.2001)	28 JUN 2001
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Uyen T Le Telephone No. 305-9000

Peggy Harrod

Form PCT/ISA/210 (second sheet) (July 1998)

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]A function, i.e., a free text item inquiry, of the following [computer system / which is the interface program memorized on a computer readable medium, and has a display], A function to receive from a user search criteria containing a domain identifier which identifies a domain, According to reception of said search criteria from a user, are a function to search two or more record identifiers from which each discriminates a corresponding record, and said corresponding record, (1). Position it in a specific position in a domain identified by domain identifier. A function which is a thing including information which has a position identifier relevant to it, and answers (2) free text item inquiry to search said two or more record identifiers, and a function which displays a list of said domains on a display, As a display of a record identified by said two or more record identifiers, Perform a function which displays two or more icons on a display, and said two or more record identifiers of each are received here, One to which it corresponds of two or more icons is displayed in a display of said domain currently displayed on a display, and an icon corresponding to each of two or more of said record identifiers, An interface program being what positioned in a display of a domain of coordinates in a domain corresponding to a corresponding position identifier for records.
[Claim 2]It sets to the interface program according to claim 1, A domain is a geographical field. An interface program, wherein said display is a multi-dimension map of a geographical field.

[Claim 3]it sets to the interface program according to claim 1 -- an interface program, wherein said display is a two-dimensional map of a geographical field.
[Claim 4]it sets to the interface program according to claim 2 -- for a function to receive an input. An interface program, wherein a function to receive specification by a user of a specified category is included and information in a category specified respectively is included in a record corresponding to two or more searched record identifiers.

[Claim 5]it sets to the interface program according to claim 4 -- for a function to receive said specification by a user of a certain category. An interface program, wherein a function to show a user a list of categories defined beforehand, a function to receive selection by a user from the list as a specified category, and ** are contained.

[Claim 6]In the interface program according to claim 3, to a computer, further The following functions, Namely, after displaying an icon corresponding to two or more record identifiers of each, from a user, are further search criteria a function to receive and said further search criteria, Said function chosen from a group of a domain identifier input type, a free text item inquiry input type, and a search-criteria type that comprises a category type, According to reception of said further search criteria from a user, are the function to search a subset of two or more of said record identifiers, and said subset of two or more of said record identifiers, Said function to identify all the record identifiers between said two or more record identifiers in said further search criteria, A function which displays a two-dimensional map of a revised geographical field on a display which answers said further search criteria, As opposed to each record identifier of said subset in two or more record identifiers, Said function positioned in a map with which coordinates corresponding to a position identifier for records to which it is a function which displays an icon [/ in said displayed map], and an icon corresponding to the record identifiers of each of said subset in said two or more record identifiers corresponds were displayed, An interface program making it perform.

[Claim 7]In the interface program according to claim 6, to a computer, further The following functions, Namely, combining said further search criteria, are search criteria expressed first a function memorized as a filter, and said memorized filter, An interface program performing said function which can be searched in order that a user may use for specification of the further search via an interface.

[Claim 8]it sets to the interface program according to claim 7 -- search criteria expressed to the beginning combined with said further search criteria, An interface program, wherein it is a series of inputs set in order and a memorized filter is an input of a series set in order which maintains a series of entry sequenced.

[Claim 9]By having a dialog with a map displayed on a computer as a function to show a user a map via the following functions, i.e., a display, further, in the interface program according to claim 3, An interface program, wherein a user performs a function to enable it to input said domain identifier, as some search criteria.

[Claim 10]it sets to the interface program according to claim 3 -- to said two or more icons. They are contained by an icon of the 1st icon class, and icon of the 2nd icon class, and an icon of the 1st icon class, An interface program, wherein it has the 1st vision characteristics and an icon of the 2nd icon class has the 2nd different vision characteristics from vision characteristics corresponding to the 1st icon class.

[Claim 11]it sets to the interface program according to claim 10 -- at least some records identified by said two or more record identifiers, A record of at least some others which is the 1st type and was identified by said two or more record identifiers, An interface program, wherein it is the 2nd type, and a record of the 1st type is displayed using an icon of the 1st icon class and a record of the 2nd type is displayed using an icon of the 2nd icon class.
[Claim 12]it sets to the interface program according to claim 3 -- at least one icon among two or more icons, An interface program, wherein it expresses a record of a large number identified by said two or more record identifiers and a record of said large number has respectively a position identifier which positions the record in near the middle position circumference.
[Claim 13]A function to receive the following functions, i.e., change of a contraction scale demand from a user, further to a computer in the interface program according to claim 3, In order to form an icon of the 2nd plurality with few numbers than the number of icons in two or more icons stated to said beginning according to change reception of said contraction scale demand, In order to identify a position of a record identified by said two or more record identifiers according to a function which unifies mutually said some of two or more icons at least, and change reception of said contraction scale demand, An interface program performing a function which carries out redisplay of said domain using an icon of the 2nd plurality using a small contraction scale.
[Claim 14]A function to receive from a user specification of an electronic memo which has a position further relevant to a computer in the following functions, i.e., a map, in the interface program according to claim 3, An interface program performing a function which displays pasting MEMOAIKON on a map of a position corresponding to a related position.
[Claim 15]it sets to the interface program according to claim 14 -- to an electronic memo. An interface program which is an address in which external access of itself is possible, and is characterized by containing in the contents a Web page which has an address which accesses possible electronically via the address.
[Claim 16]In database system memorized on a computer readable medium, a function of the following [computer system] -- namely, With a function to receive search criteria in which at least one of (1) text, a domain identifier which identifies (2) domains, and filter identifiers which identify (3) filters is contained. A function to search said record identifier which identifies a corresponding record in which it is two or more record identifiers, and the each related a text, a domain identifier, or a filter identifier of search criteria with it, Database system, wherein it makes it perform and said function to search is performed with a space keyword document index here.
[Claim 17]it is the method of searching two or more record identifiers -- each of two or more of said record identifiers, A method, wherein it has at least one of a text relevant to it specified by search criteria, a domain identifier, and layer identifiers and this search is performed with a

space keyword document index.
[Claim 18]it sets to a method according to claim 17 -- it being extended to a space keyword document index, in order to refer to a document, and said space index tree and two or more trees which have the same structure, but. A method, wherein a space index tree corrected is contained to a specific glossary item and a filter.
[Claim 19]it sets to a method according to claim 17 -- a method, wherein said two or more record identifiers are searched in a space keyword document index tree, and branching structure of said tree is analyzed and a geographical phenomenon is identified.
[Claim 20]it sets to a method according to claim 19 -- a method, wherein a geographical phenomenon is the branching structure of a space keyword document index tree which shares a parent node with more few predetermined branches than a predetermined number.
[Claim 21]it is the program memorized on a computer readable medium -- to a computer system with the following functions, i.e., a function to read a document referred to by a document address. A program performing a function which analyzes the syntax of those documents, and a function which there is no possible space identifier or analyzes the syntax of those documents in order to read many document addresses.
[Claim 22]it sets to the program according to claim 21 -- a program performing a function to analyze a possible space identifier in order to determine the following functions, i.e., a position in a domain, as a computer system further.
[Claim 23]it sets to the program according to claim 21 -- a program collecting some document addresses by a meta search part process, and asking other computer systems a meta search part process using a text which refers to a space domain.
[Claim 24]it sets to the program according to claim 21 -- it was further found out by computer system at the following functional, i.e., each, documents -- each -- a program performing a function which computes a degree-of-association score to a possible space identifier.
[Claim 25]it sets to the program according to claim 21 -- for a degree-of-association score. (1) A position of a possible space identifier in a document, the number of other possible space identifiers in (2) documents, (3) A possible space identifier is [the inside of a sentence, or] whether isolated or not, (4) A program, wherein one or more of emphasis to which formatting of the character in a possible space identifier was carried out, and ** are contained.
[Claim 26]it sets to the program according to claim 21 -- a program performing a function to memorize the following functions to a computer system and to memorize a document address for every degree-of-association score before reading further.
[Claim 27]it is the program memorized on a computer readable medium -- to a computer system with the following functions, i.e., a function to read a document referred to by a document address. In order to determine a position in a domain as a function which analyzes the syntax of those documents in order to read many document addresses, and a function

which there is no possible space identifier or analyzes the syntax of those documents, A program performing a function to analyze a possible space identifier.
[Claim 28]it sets to a method for displaying information coded spatially -- by an automated computer process. A stage of collecting documents in a database, and a stage which chooses a subset of a document it can be judged that includes spacial information, Are at least one space identifier each document in a subset, a stage to which it is made to correspond, and a stage which carries out an indexing to a document, and said indexing, With said stage where an index about an index about a space identifier and a keyword is contained, and a computer interface, a user, A method characterized by containing a stage which answers an inquiry by set, a stage which displays a result set on a user via a computer interface, and ** as a result of containing a document, a stage of providing this computer interface to which an inquiry in which spacial information is included can be contributed, and.
[Claim 29]it sets to a method according to claim 28 -- a result -- a set -- one or more -- when it includes an element, two or more groups constituted by spatial approachability are included -- each of the group -- a result -- a method of a set containing at least one document.
[Claim 30]it sets to a method according to claim 29 -- a method, wherein two or more groups align according to a predetermined function about a group showing a degree of association to a standard.
[Claim 31]it sets to a method according to claim 29 -- a method, wherein contents of each group align according to a predetermined function about a group showing a degree of association to a standard.
[Claim 32]it sets to a method according to claim 28 -- a method, wherein a keyword is contained in a standard.
[Claim 33]In a method for storing in a space document data base a document by which the hyperlink was carried out, including spacial information, A stage of providing a destination database with which a possible source of dispatch of a collectable document is included, A stage of providing a history database of a known source of dispatch with which documents were collected, A stage of providing a round part computer process which can access a possible source of dispatch of a collectable document which followed a hyperlink of a document and was specified by hyperlink, they are a stage of performing bootstrapping of a round part, and a stage which repeats a round part on a destination database -- with a step which moves a possible source of dispatch of a collectable document to a history database from a destination database. A step which investigates a possible source of dispatch to see there is not any collectable document, A step which memorizes all documents in which such collection is possible to a space document data base, A method including said stage containing a step which adds all the possible sources of dispatch of a collectable document referred to by hyperlink in a collectable document to a destination database to repeat.

[Claim 34]A method comprising according to claim 33:
Bootstrapping, A stage of providing two or more positions which are known objects
Are the stage of providing a destination database with two or more sources of meta-dispatch, and each source of meta-dispatch, Said stage which it is a source of dispatch of a possible source of dispatch of a collectable document, and each source of meta-dispatch is a set as a result of including a possible source of dispatch of a collectable document, and answers an inquiry by a computer process.
By carrying out repetition operation of the preparation part process, it is the stage of preparing a destination database, A step which forms an inquiry based on at least one position which is a known object
A step which addresses an inquiry to a source of meta-dispatch.
A step as for which a result for an inquiry receives a set.
A step which inserts each element of a result set in a destination database.
Said stage to include and to prepare.

[Claim 35]A method characterized by comprising the following for carrying out score attachment at goodness of fit to a document containing a known space identifier through which it asks and passes.
A stage which carries out score attachment at a document to goodness of fit of a document to a place specified in an inquiry.
A stage which carries out score attachment at a document to goodness of fit of a document to a word specified in an inquiry.
A stage which carries out score attachment to quality of a document at a document.
A stage used as a single score combining a score.

[Claim 36]In a method for making possible an inquiry which carries out an indexing to two or more documents, and includes a keyword and spacial information, It is a stage which initializes a main space tree of a predetermined degree as a computer-data structure, (1) Each leaf node of a main space tree expresses a document, Each non leaf nodes of (2) main-space tree, the range of space is expressed A root node of (3) main-space tree the range of space which includes arbitrary space which an effective inquiry can refer to, [express and] (4) A method that a set of a non-leaf child node is characterized by including said stage initialized like and a stage of adding a child node to a tree recursively of defining a division on the range of space expressed by parents on the basis of a parent node.

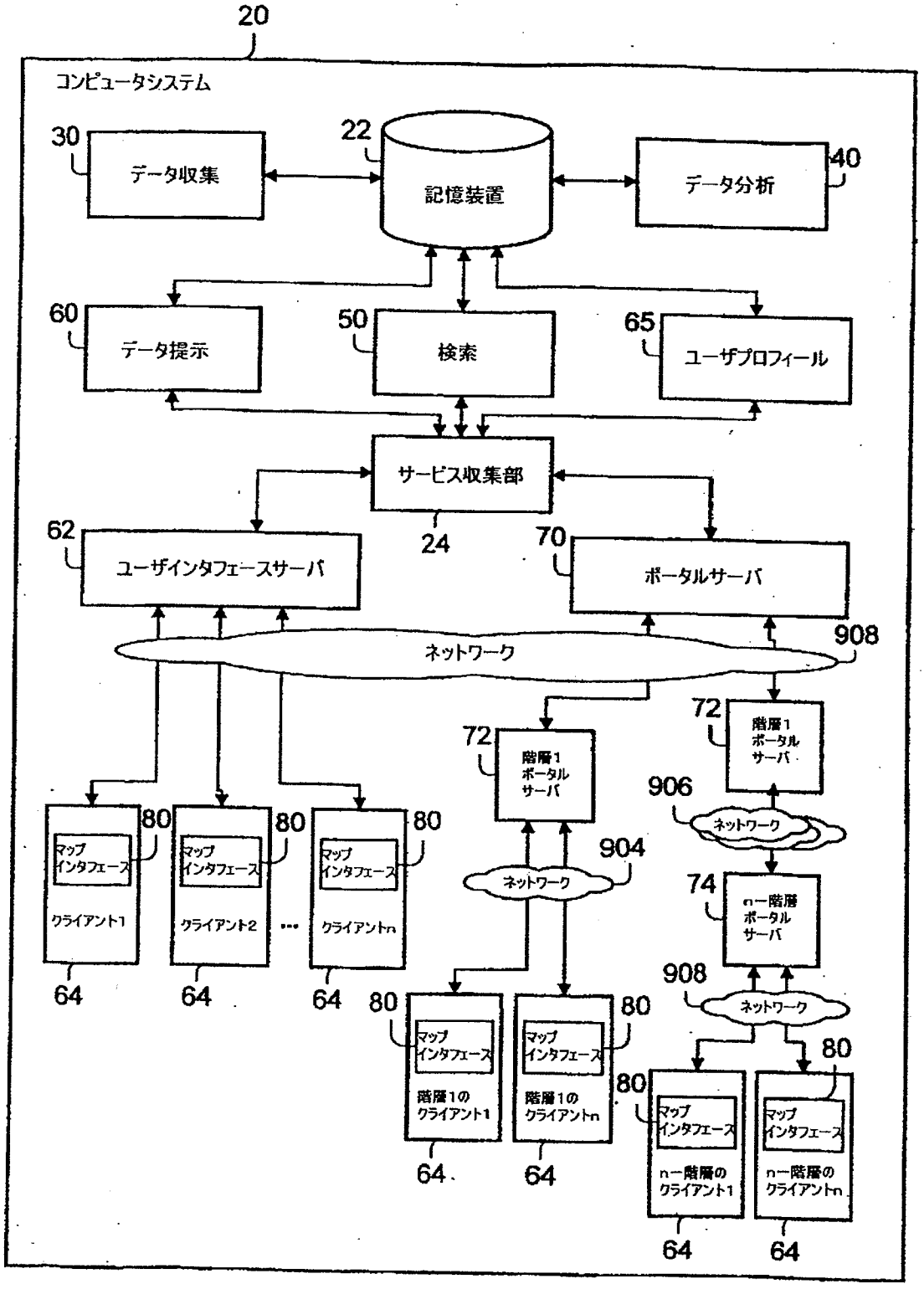
[Translation done.]

* NOTICES *

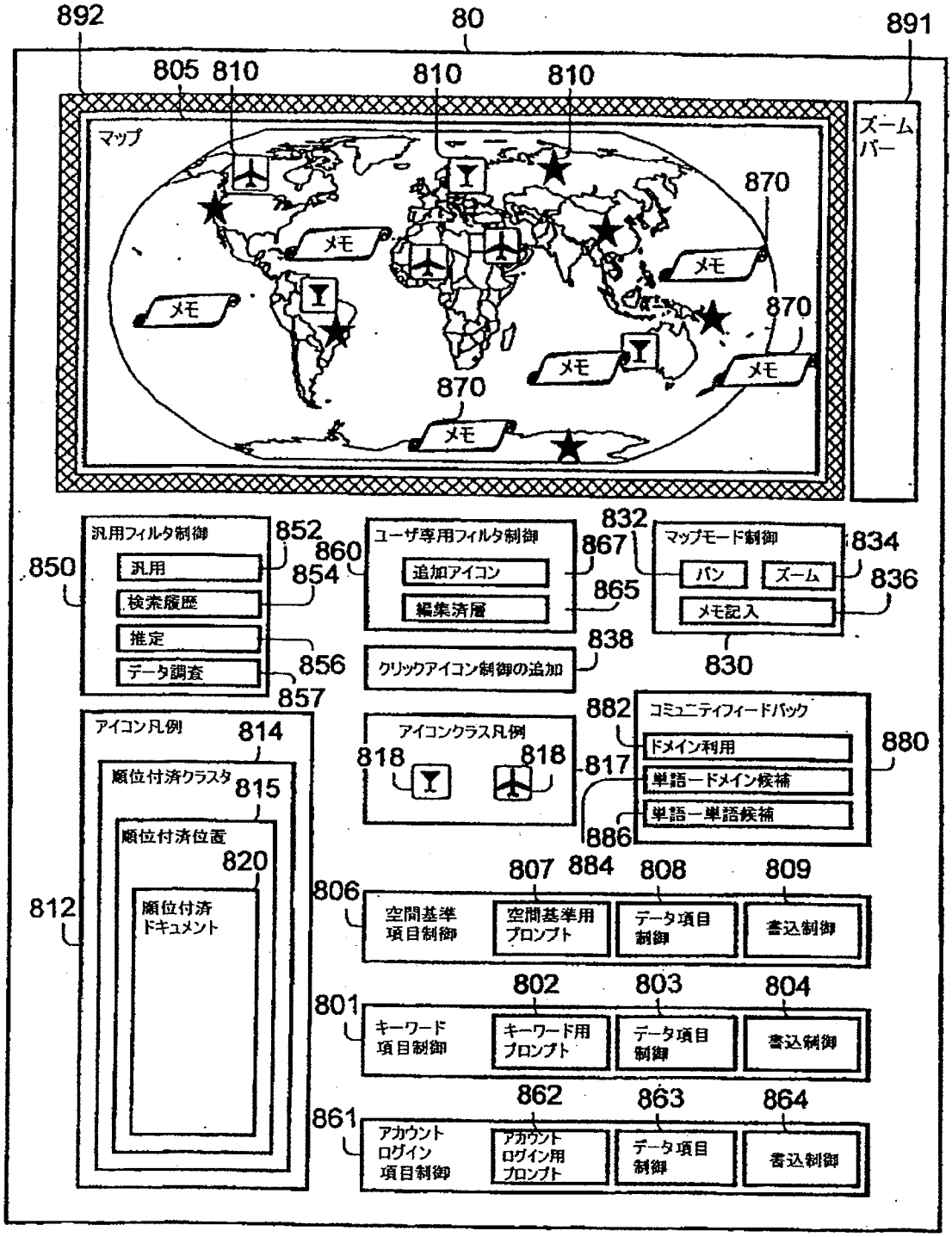
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

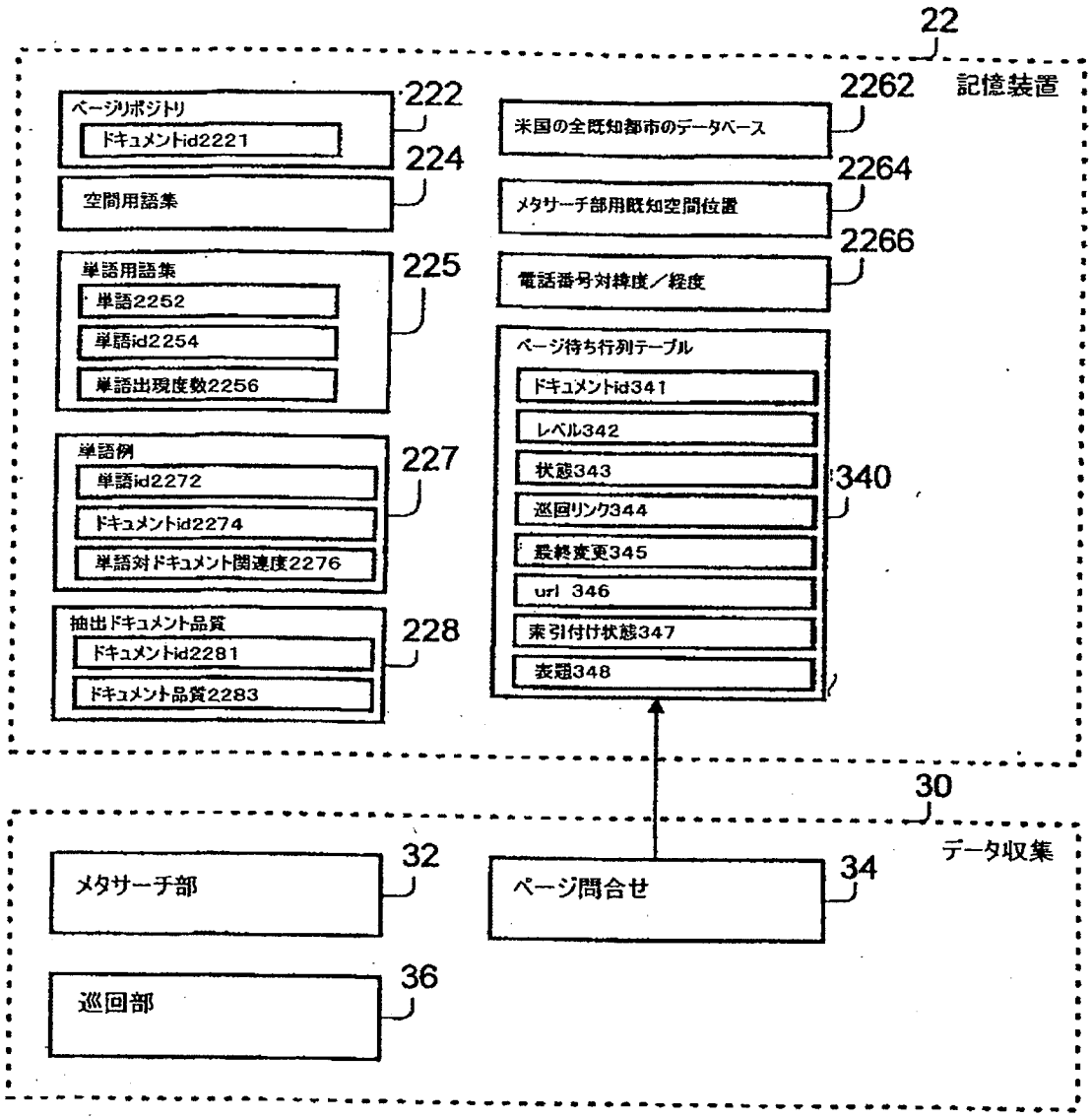
[Drawing 1]


20
コンピュータシステム
30
データ収集
22
記憶装置
40
データ分析
60
データ提示
50
検索
65
ユーザプロフィール
サービス収集部
24
62
ユーザインタフェースサーバ
70
ポータルサーバ
908
ネットワーク
72
階層1
ポータル
サーバ
906
ネットワーク
904
ネットワーク
74
n一階層
ポータル
サーバ
908
ネットワーク
80
マップ
インタフェース
クライアント1
クライアント2
クライアントn
64
階層1の
クライアント1
階層1の
クライアントn
n一階層の
クライアント1
n一階層の
クライアントn

[Drawing 2]


892
80
891
805
810
810
810
マップ
ズームバー
870
メモ
メモ
メモ
870
メモ
メモ
870
メモ
汎用フィルタ制御
852
850
汎用
860
854
検索履歴
推定
856
データ調査
857
ユーザ専用フィルタ制御
832
867
追加アイコン
865
編集済層
838
クリックアイコン制御の追加
マップモード制御
834
パン
ズーム
836
メモ記入
830
アイコン凡例
814
順位付済クラスタ
815
順位付済位置
820
順位付済ドキュメント
812
アイコンクラス凡例
818
818
882
コミュニティフィードバック
ドメイン利用
880
817
単語ードメイン候補
単語ー単語候補
886
807
884
808
809
806
空間基準項目制御
空間基準用プロンプト
データ項目制御
書込制御
802
803
804
801
キーワード項目制御
キーワード用プロンプト
データ項目制御
書込制御
862
863
864
861
アカウントログイン項目制御
アカウントログイン用プロンプト
データ項目制御
書込制御

[Drawing 3]


22
記憶装置
ページリポジトリ
ドキュメントid2221
222
空間用語集
224
単語用語集
単語2252
単語id2254
単語出現度数2256
225
単語例
単語id2272
ドキュメントid2274
単語対ドキュメント関連度2276
227
抽出ドキュメント品質
ドキュメントid2281
ドキュメント品質2283
228
米国の全既知都市のデータベース
2262
メタサーチ部用既知空間位置
2264
電話番号対緯度／経度
2266
ページ待ち行列テーブル
ドキュメントid341
レベル342
状態343
巡回リンク344
最終変更345
url 346
索引付け状態347
表題348
340
30
データ収集
メタサーチ部
32
ページ問合せ
34
巡回部
36

[Drawing 4]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
Based on USC35Section119 (e) and (1), this application, U.S. temporary patent application the 60/183 of the point for which it applied on February 22, 2000, No. 971 and a title "information retrieval engine of a Metacarta:map base, and catalog", And the benefit of the 60/201 for which it applied on May 3, 2000, and No. 839 and a title "the method for making information equivalent to physical objects and a position, system, and method of database extension" is asserted.
The citation reference of these both is carried out into this Description.

[0002]
Technical field Especially this invention relates to visualization of a space database, a document data base, a search engine, and data about a computer system.
[0003]
Background Arrangement of a document and access have a tool of available many with various interfaces which are useful for a user's information retrieval. There is a thing with which a user enables it to search the document which is in agreement with special standards, such as a document containing the specified keyword, in these tools. The operating direction etc. which are displayed on a map have some things which provide the information about a geographical region or a space domain in these tools.
The Internet of these tools, etc. may be available on an open network available on a secret computer system. The user can collect information using these tools.
[0004]
Outline of an invention In the computer system which shows a user a map interface, especially this invention is carried out as [be / the check of a list / possible] as a result of the inquiry

which made possible the inquiry through the map interface by a user, and has been arranged as an icon on a map. A map and an icon react to user action further, and change of the map range, change of inquiry conditions, or the more detailed check of the subset of a result is included in this.

[0005]

The object of an inquiry is a document. Text-based computer filing, a text-based file and the file which includes spacial information selectively, and the computer entity that can be accessed via a document-like interface are contained in the example of a document. A document may also contain other documents and, in addition to a document-like interface, may have other interfaces. Each document has an address. In the case of a World-Wide-Web document, this address is usually URL.

[0006]

A document exists on computer systems arranged at the whole computer network, such as a secret network and the Internet. The hyperlink of the document may be carried out, namely, it may also include the reference (hyperlink) to the address of other documents. The copy of a document can be memorized to a page repository.

[0007]

A space perception part process investigates a document about the existence of spacial information contents. If a space perception part judges that a certain document has spacial information contents, the document will be added to a space document group.

[0008]

A document ranking process assigns a space degree-of-association score to each document of a space document group. A space degree-of-association score is a measure of the degree relevant to the spatial position where the document was described by the spacial information contents. When a document has two or more instances of spacial information contents, a document attaches a score to each instance.

[0009]

(19)日本国特許庁（ＪＰ）　(12) **公表特許公報**（Ａ）　(11)特許出願公表番号
特表2003－524259
（P2003－524259A）
(43)公表日　平成15年８月12日(2003.8.12)

(51)Int.Cl.[7]	識別記号	ＦＩ	テーマコード（参考）
Ｇ０６Ｆ　17/30	３６０	Ｇ０６Ｆ　17/30	３６０Ｚ　２Ｃ０３２
	２１０		２１０Ａ　５Ｂ０７５
	３５０		３５０Ｃ
	４１９		４１９Ａ
Ｇ０９Ｂ　29/00		Ｇ０９Ｂ　29/00	Ａ

審査請求　未請求　予備審査請求　有　（全 75 頁）　最終頁に続く

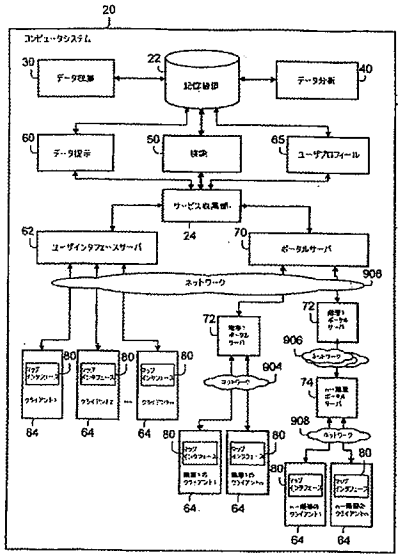
(21)出願番号　特願2001－562372(P2001－562372)
(86)(22)出願日　平成13年２月22日(2001.2.22)
(85)翻訳文提出日　平成14年８月22日(2002.8.22)
(86)国際出願番号　ＰＣＴ／ＵＳ０１／４０１７３
(87)国際公開番号　ＷＯ０１／０６３４７９
(87)国際公開日　平成13年８月30日(2001.8.30)
(31)優先権主張番号　６０／１８３，９７１
(32)優先日　平成12年２月22日(2000.2.22)
(33)優先権主張国　米国（ＵＳ）
(31)優先権主張番号　６０／２０１，８３９
(32)優先日　平成12年５月３日(2000.5.3)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　メタカルタ　インコーポレイテッド
ＭＥＴＡＣＡＲＴＡ，ＩＮＣ．
アメリカ合衆国　02139　マサチューセッツ州　ケンブリッジ　プロスペクト　ストリート　126　スイート　５
(72)発明者　フランク、ジョン　アール．
アメリカ合衆国　02139　マサチューセッツ州　ケンブリッジ　ピー．オー．　ボックス　397207
(74)代理人　弁理士　恩田　博宣　（外１名）

最終頁に続く

(54)【発明の名称】　情報の空間符号化及び表示

(57)【要約】
コンピュータシステム（２０）は、記憶装置システム（２２）を含み、記憶装置２２には、ドキュメント形式の情報とそのドキュメントに関する空間情報とが含まれる。また、コンピュータシステム（２０）は、データ収集（３０）、データ分析（４０）、検索（５０）、データ提示（６０）、及びポータルサービス（７０）用のサブシステムを含む。更に、コンピュータシステム（２０）は、地図インタフェース（８０）を含む。地図インタフェース（８０）によって、ユーザは、記憶装置（２２）への問合せが可能であり、また問合せ結果の一覧を地図に配置して閲覧できる。



【特許請求の範囲】

【請求項１】　コンピュータ可読媒体上に記憶されたインタフェースプログラムであって、

表示装置を有するコンピュータシステムに以下の機能、すなわち、

自由テキスト項目問合せと、ドメインを識別するドメイン識別子とを含む検索基準をユーザから受信する機能と、

ユーザからの前記検索基準の受信に応じて、対応するレコードを各々が識別する複数のレコード識別子を検索する機能であって、前記対応するレコードは、（１）ドメイン識別子によって識別されたドメイン内の特定の位置でそれを位置付ける、それと関連した位置識別子を有し、また、（２）自由テキスト項目問合せに応答する情報を含むものである、前記複数のレコード識別子を検索する機能、

表示装置上に前記ドメインの一覧を表示する機能と、

前記複数のレコード識別子によって識別されたレコードの表示として、複数のアイコンを表示装置上に表示する機能と、

を実行させ、ここに、前記複数のレコード識別子各々に対して、複数のアイコンの内の対応する１つが、表示装置上に表示されている前記ドメインの表示内に表示され、前記複数のレコード識別子の各々に対応するアイコンは、対応するレコード用の位置識別子に対応するドメイン内にある座標のドメインの表示内に位置付けられるものであることを特徴とするインタフェースプログラム。

【請求項２】　請求項１に記載のインタフェースプログラムにおいて、

ドメインは、地理的な領域であり、前記表示は、地理的な領域の多次元地図であることを特徴とするインタフェースプログラム。

【請求項３】　請求項１に記載のインタフェースプログラムにおいて、

前記表示は、地理的な領域の２次元地図であることを特徴とするインタフェースプログラム。

【請求項４】　請求項２に記載のインタフェースプログラムにおいて、

入力を受信する機能には、更に、指定されたカテゴリのユーザによる指定を受信する機能が含まれ、また、複数の検索されたレコード識別子に対応するレコードには各々、指定されたカテゴリ内にある情報が含まれることを特徴とするイン

タフェースプログラム。

【請求項５】　請求項４に記載のインタフェースプログラムにおいて、

あるカテゴリのユーザによる前記指定を受信する機能には、予め定義されたカテゴリのリストをユーザに提示する機能と、指定されたカテゴリとして、そのリストからユーザによる選択を受信する機能と、が含まれることを特徴とするインタフェースプログラム。

【請求項６】　請求項３に記載のインタフェースプログラムにおいて、コンピュータに更に以下の機能、すなわち、

複数のレコード識別子各々に対応するアイコンを表示した後、ユーザから更なる検索基準を受信する機能であって、前記更なる検索基準は、ドメイン識別子入力タイプ、自由テキスト項目問合せ入力タイプ、及びカテゴリタイプから成る検索基準タイプの群から選択される前記機能と、

ユーザからの前記更なる検索基準の受信に応じて、前記複数のレコード識別子の部分集合を検索する機能であって、前記複数のレコード識別子の前記部分集合は、前記更なる検索基準内にある前記複数のレコード識別子間の全レコード識別子を識別する前記機能と、

改版された地理的な領域の２次元地図を前記更なる検索基準に応答する表示装置上に表示する機能と、

複数のレコード識別子における前記部分集合の各レコード識別子に対して、前記表示された地図内に対応するアイコンを表示する機能であって、前記複数のレコード識別子における前記部分集合のレコード識別子各々に対応するアイコンは、対応するレコード用の位置識別子に対応する座標の表示された地図内に位置付けられる前記機能と、

を実行させることを特徴とするインタフェースプログラム。

【請求項７】　請求項６に記載のインタフェースプログラムにおいて、コンピュータに更に以下の機能、すなわち、

前記更なる検索基準と組み合せて、最初に述べた検索基準をフィルタとして記憶する機能であって、前記記憶されたフィルタは、インタフェースを介して、更なる検索の指定にユーザが用いるために検索可能である前記機能を実行させるこ

とを特徴とするインタフェースプログラム。

【請求項８】　請求項７に記載のインタフェースプログラムにおいて、

前記更なる検索基準と組み合せた最初に述べた検索基準は、順序付けした一連の入力であり、また、記憶されたフィルタは、一連の入力順を維持する順序付けした一連の入力であることを特徴とするインタフェースプログラム。

【請求項９】　請求項３に記載のインタフェースプログラムにおいて、コンピュータに更に以下の機能、すなわち、

表示装置を介して、ユーザに地図を提示する機能と、

表示された地図と対話することによって、検索基準の一部としてユーザが前記ドメイン識別子を入力できるようにする機能と、

を実行させることを特徴とするインタフェースプログラム。

【請求項１０】　請求項３に記載のインタフェースプログラムにおいて、

前記複数のアイコンには、第１アイコンクラスのアイコンと第２アイコンクラスのアイコンとが含まれ、また、第１アイコンクラスのアイコンは、第１視覚特性を有し、また、第２アイコンクラスのアイコンは、第１アイコンクラスに対応する視覚特性とは異なる第２視覚特性を有していることを特徴とするインタフェースプログラム。

【請求項１１】　請求項１０に記載のインタフェースプログラムにおいて、

前記複数のレコード識別子によって識別された少なくとも幾つかのレコードは、第１タイプであり、前記複数のレコード識別子によって識別された少なくとも幾つかの他のレコードは、第２タイプであり、また、第１タイプのレコードは、第１アイコンクラスのアイコンを用いて表示され、第２タイプのレコードは、第２アイコンクラスのアイコンを用いて表示されることを特徴とするインタフェースプログラム。

【請求項１２】　請求項３に記載のインタフェースプログラムにおいて、

複数のアイコンの内、少なくとも１つのアイコンは、前記複数のレコード識別子によって識別される多数のレコードを表し、前記多数のレコードは各々、中央位置周辺付近内において、そのレコードを位置付ける位置識別子を有することを特徴とするインタフェースプログラム。

【請求項１３】　請求項３に記載のインタフェースプログラムにおいて、コンピュータに更に以下の機能、すなわち、

ユーザからの縮尺要求の変更を受信する機能と、

前記縮尺要求の変更受信に応じて、前記最初に述べた複数のアイコンにおけるアイコン数より数が少ない第２複数のアイコンを形成するために、少なくとも幾つかの前記複数のアイコンを互いに統合する機能と、

前記縮尺要求の変更受信に応じて、前記複数のレコード識別子によって識別されたレコードの位置を識別するために、小さい縮尺を用いて、また、第２複数のアイコンを用いて、前記ドメインを再表示する機能と、

を実行させることを特徴とするインタフェースプログラム。

【請求項１４】　請求項３に記載のインタフェースプログラムにおいて、コンピュータに更に以下の機能、すなわち、

地図内に関連する位置を有する電子メモの指定をユーザから受信する機能と、

関連する位置に対応する位置の地図上に貼付メモアイコンを表示する機能と、

を実行させることを特徴とするインタフェースプログラム。

【請求項１５】　請求項１４に記載のインタフェースプログラムにおいて、

電子メモには、それ自身の外部アクセス可能なアドレスであって、そのアドレスを介してそのコンテンツに電子的にアクセスを可能にするアドレスを有するウェッブページが含まれることを特徴とするインタフェースプログラム。

【請求項１６】　コンピュータ可読媒体上に記憶されたデータベースシステムにおいて、コンピュータシステムに以下の機能、すなわち、

（１）テキスト、（２）ドメインを識別するドメイン識別子、及び（３）フィルタを識別するフィルタ識別子の内、少なくとも１つが含まれる検索基準を受信する機能と、

複数のレコード識別子であって、その各々が、検索基準のテキスト、ドメイン識別子、又はフィルタ識別子をそれと関連付けた対応するレコードを識別する前記レコード識別子を検索する機能と、

を実行させ、ここに、前記検索する機能は、空間キーワードドキュメント索引で実行されることを特徴とするデータベースシステム。

【請求項１７】　複数のレコード識別子を検索する方法であって、

前記複数のレコード識別子の各々は、それと関連する、検索基準によって指定されるテキスト、ドメイン識別子、及び層識別子の内の少なくとも１つを有し、かかる検索は、空間キーワードドキュメント索引で実行されることを特徴とする方法。

【請求項１８】　請求項１７に記載の方法において、

空間キーワード・ドキュメント索引には、ドキュメントと、前記空間索引ツリーと同じ構造を有する複数のツリーと、を参照するためには拡張されるが、特定の用語集項目やフィルタに対しては修正される空間索引ツリーが含まれることを特徴とする方法。

【請求項１９】　請求項１７に記載の方法において、

前記複数のレコード識別子は、空間キーワード・ドキュメント索引ツリーにおいて検索され、また、前記ツリーの分岐構造が解析されて地理的な現象が識別されることを特徴とする方法。

【請求項２０】　請求項１９に記載の方法において、

地理的な現象は、所定のわずかな枝が、所定の数より多い親ノードを共有する空間キーワード・ドキュメント索引ツリーの分岐構造であることを特徴とする方法。

【請求項２１】　コンピュータ可読媒体上に記憶されたプログラムであって、

コンピュータシステムに以下の機能、すなわち、

ドキュメントアドレスによって参照されるドキュメントを読み込む機能と、

更に多くのドキュメントアドレスを読み込むために、それらのドキュメントを構文解析する機能と、

可能な空間識別子が無いか、それらのドキュメントを構文解析する機能と、

を実行させることを特徴とするプログラム。

【請求項２２】　請求項２１に記載のプログラムにおいて、

コンピュータシステムに更に以下の機能、すなわち、

ドメイン中における位置を決定するために、可能な空間識別子を解析する機能

を実行させることを特徴とするプログラム。

【請求項２３】　請求項２１に記載のプログラムにおいて、

一部のドキュメントアドレスが、メタ検索部プロセスによって収集され、メタ検索部プロセスは、空間ドメインを参照するテキストを用いて、他のコンピュータシステムに問合せることを特徴とするプログラム。

【請求項２４】　請求項２１に記載のプログラムにおいて、

コンピュータシステムに更に以下の機能、すなわち、

各ドキュメントに見いだされた各可能な空間識別子に対する関連度スコアを算出する機能を実行させることを特徴とするプログラム。

【請求項２５】　請求項２１に記載のプログラムにおいて、

関連度スコアには、

（１）ドキュメントにおける可能な空間識別子の位置と、

（２）ドキュメントにおける他の可能な空間識別子の数と、

（３）可能な空間識別子が、文中又は孤立しているか否かと、

（４）可能な空間識別子中の文字のフォーマット化された強調と、の内の１つ以上が含まれることを特徴とするプログラム。

【請求項２６】　請求項２１に記載のプログラムにおいて、

コンピュータシステムに更に以下の機能、すなわち、

読み込み前に、関連度スコア毎にドキュメントアドレスを記憶する機能を実行させることを特徴とするプログラム。

【請求項２７】　コンピュータ可読媒体上に記憶されたプログラムであって、

コンピュータシステムに以下の機能、すなわち、

ドキュメントアドレスによって参照されるドキュメントを読み込む機能と、

更に多くのドキュメントアドレスを読み込むために、それらのドキュメントを構文解析する機能と、

可能な空間識別子が無いか、それらのドキュメントを構文解析する機能と、

ドメイン中における位置を決定するために、可能な空間識別子を解析する機能と、

を実行させることを特徴とするプログラム。

【請求項２８】　空間的に符号化された情報を表示するための方法において、

自動化されたコンピュータプロセスによって、データベースにドキュメントを収集する段階と、

空間情報を含むと判断し得るドキュメントの部分集合を選択する段階と、

少なくとも１つの空間識別子を部分集合中の各ドキュメントと対応させる段階と、

ドキュメントに索引付けする段階であって、前記索引付けは、空間識別子に関する索引とキーワードに関する索引とが含まれる前記段階と、

コンピュータインタフェースによって、ユーザは、空間情報が含まれる問合せを投稿し得る該コンピュータインタフェースを提供する段階と、

ドキュメントが含まれる結果集合で問合せに応答する段階と、

コンピュータインタフェースを介して、ユーザに結果集合を表示する段階と、が含まれることを特徴とする方法。

【請求項２９】　請求項２８に記載の方法において、

結果集合は、１つ以上の要素を含む場合、空間的な近接性によって構成される複数の群を含み、その各々の群は、結果集合の少なくとも１つのドキュメントを含むことを特徴とする方法。

【請求項３０】　請求項２９に記載の方法において、

複数の群が、基準に対する関連度を表す群に関する所定の機能に従って整列されることを特徴とする方法。

【請求項３１】　請求項２９に記載の方法において、

各群のコンテンツが、基準に対する関連度を表す群に関する所定の機能に従って整列されることを特徴とする方法。

【請求項３２】　請求項２８に記載の方法において、

基準には、キーワードが含まれることを特徴とする方法。

【請求項３３】　空間情報を含むハイパーリンクされたドキュメントを空間ドキュメントデータベースに格納するための方法において、

収集可能なドキュメントの可能な発信源が含まれる宛先データベースを提供する段階と、

ドキュメントが収集された既知の発信源の履歴データベースを提供する段階と、

ドキュメントのハイパーリンクに追従して、ハイパーリンクによって指定された収集可能なドキュメントの可能な発信源にアクセスし得る巡回部コンピュータプロセスを提供する段階と、

巡回部のブートストラッピングを行う段階と、

宛先データベース上で巡回部を反復する段階であって、

収集可能なドキュメントの可能な発信源を宛先データベースから履歴データベースに移動するステップと、

収集可能なドキュメントが無いか、可能な発信源を調べるステップと、

このような収集可能なあらゆるドキュメントを空間ドキュメントデータベースに記憶するステップと、

収集可能なドキュメントにおいてハイパーリンクによって参照される収集可能なドキュメントの全ての可能な発信源を宛先データベースに追加するステップと、を含む前記反復する段階と、

を含むことを特徴とする方法。

【請求項３４】　請求項３３に記載の方法において、

ブートストラッピングは、

既知の対象である複数の位置を提供する段階と、

宛先データベースに複数のメタ発信源を提供する段階であって、各メタ発信源は、収集可能なドキュメントの可能な発信源の発信源であり、また、各メタ発信源は、収集可能なドキュメントの可能な発信源を含む結果集合で、コンピュータプロセスによる問合せに応答する前記段階と、

準備部プロセスを繰り返し動作させることによって、宛先データベースを準備する段階であって、

既知の対象である少なくとも１つの位置に基づき問合せを形成するステップと、

問合せをメタ発信源に宛てるステップと、

問合せ用の結果集合を受信するステップと、

結果集合の各要素を宛先データベースに挿入するステップと、を含む前記準備する段階と、

を含むことを特徴とする方法。

【請求項３５】　既知の空間識別子を含むドキュメントに対する問合せへの適合度にスコア付けするための方法であって、

問合せにおいて指定された場所へのドキュメントの適合度に対してドキュメントにスコア付けする段階と、

問合せにおいて指定された単語へのドキュメントの適合度に対してドキュメントにスコア付けする段階と、

ドキュメントの品質に対してドキュメントにスコア付けする段階と、

スコアを組み合せて単一のスコアにする段階と、

を含むことを特徴とする方法。

【請求項３６】　複数のドキュメントに索引付けして、キーワードと空間情報とを含む問合せを可能にするための方法において、

所定の度合の主空間ツリーをコンピュータデータ構造として初期化する段階であって、

（１）主空間ツリーの各葉ノードが、ドキュメントを表し、

（２）主空間ツリーの各非葉ノードが、空間の範囲を表し、

（３）主空間ツリーの根ノードが、有効な問合せが参照し得る任意の空間を包含する空間の範囲を表し、そして、

（４）親ノードを基準にして、非葉子ノードの集合が、親によって表される空間の範囲上の区画を定義する、ように初期化する前記段階と、

子ノードをツリーに再帰的に追加する段階と、

を含むことを特徴とする方法。

【発明の詳細な説明】

【０００１】

ＵＳＣ３５§１１９（ｅ）（１）に基づき、本出願は、２０００年2月２２日に出願された先の米国仮特許出願第６０／１８３、９７１号・表題“Ｍｅｔａｃａｒｔａ：地図ベースの情報検索エンジンとカタログ”、及び２０００年5月３日に出願された第６０／２０１、８３９号・表題“情報を物理的オブジェクトと位置に対応させるための方法とシステム及びデータベース拡張の方法”の恩恵を主張するものであり、この両方を本明細書中に引用参照する。

【０００２】

技術分野

本発明は、コンピュータシステムに関し、特に、空間データベース、ドキュメントデータベース、検索エンジン、及びデータの視覚化に関する。

【０００３】

背景

ユーザの情報検索に役立つ様々なインタフェースによってドキュメントの整理やアクセスに利用可能な多くのツールがある。これらのツールには、指定されたキーワードを含むドキュメント等、特別な基準に一致するドキュメントをユーザが検索できるようにするものがある。これらのツールには、地図上に表示される運転方向等、地理的領域又は空間ドメインに関する情報を提供するものが幾つかある。

これらのツールは、非公開のコンピュータシステム上で利用可能であり、またインターネット等、公開ネットワーク上で利用可能な場合もある。ユーザは、これらのツールを用いて、情報を収集し得る。

【０００４】

発明の概要

地図インタフェースをユーザに提示するコンピュータシステムにおいて、本発明は、特に、ユーザによる地図インタフェースを介した問合せを可能にし、また地図上にアイコンとして配置した問合せの結果一覧のチェックが可能なようにする。地図とアイコンは、更にユーザアクションに反応し、これには、地図範囲の

変更、問合せ条件の変更、又は結果の部分集合のより綿密なチェックが含まれる。

【０００５】

問い合せの対象は、ドキュメントである。ドキュメントの例には、テキストベースのコンピュータファイル、並びに、部分的にテキストベースのファイル、空間情報を含むファイル、及びドキュメントライクなインタフェースを介してアクセスし得るコンピュータエンティティが含まれる。ドキュメントは、他のドキュメントを含んでもよく、またドキュメントライクなインタフェースに加えて他のインタフェースを有してもよい。各ドキュメントは、アドレスを有する。ワールドワイドウェブ・ドキュメントの場合、通常、このアドレスはＵＲＬである。

【０００６】

ドキュメントは、非公開ネットワークやインターネット等、コンピュータネットワーク全体に配置されたコンピュータシステム上に存在する。ドキュメントは、ハイパーリンクされてもよく、すなわち、他のドキュメントのアドレスへの参照（ハイパーリンク）を含んでもよい。ドキュメントのコピーは、ページリポジトリに記憶し得る。

【０００７】

空間認識部プロセスは、空間情報コンテンツの有無に関してドキュメントを調べる。あるドキュメントに空間情報コンテンツがあると空間認識部が判断すると、そのドキュメントは、空間ドキュメント集合に追加される。

【０００８】

ドキュメント順位付けプロセスは、空間ドキュメント集合の各ドキュメントに空間関連度スコアを割当てる。空間関連度スコアとは、ドキュメントがその空間情報コンテンツに記述された空間位置に関連する度合の尺度である。ドキュメントが、空間情報コンテンツのインスタンスを２つ以上有する場合、ドキュメントは、各インスタンスに対してスコアを付ける。

【０００９】

空間・キーワード・ドキュメント索引付け部は、空間ドキュメント集合の各ドキュメントを調べ、また、それを空間・キーワード・ドキュメント索引データ構

造で表す。空間・キーワード・ドキュメント索引付け部は、キーワードと少なくとも１つの空間情報コンテンツのインスタンスとの両方によって、ドキュメントに索引付けする。通常、空間・キーワード・ドキュメント索引によって、コンピュータシステムは、空間基準をキーワード基準と組み合せた問合せに対して、非常に高速に反応し得る。

【００１０】

巡回部は、既知のドキュメントに含まれるハイパーリンクを調べて、既知のドキュメントの集合を拡張する。ハイパーリンクが、これまで未知のドキュメントを参照する場合、巡回部は、未知のドキュメントを既知のドキュメントの集合に追加し、それらを調べて、そして新しいハイパーリンクが続くようにする。

【００１１】

巡回部は、部分的に空間関連度スコアに基づき、巡回部が従うハイパーリンクの優先順位付けを行うことができる。

コンピュータシステムは、既知のドキュメントの集合を初期化するためのメタサーチ部プロセスを含む。この初期化ステップは、ブートストラッピングとして知られ、またこの技術では公知である。メタサーチ部は、インターネット上の検索エンジンウェブサイト等、他のコンピュータシステムやドキュメント資源に関する情報を記憶するように、既知の所定の検索エンジンに問合せる。メタサーチ部の管理人は、それに、既知の空間位置の集合を提供する。メタサーチ部は、これらの空間位置に基づき、問い合せを形成し、その問い合せを検索エンジンに宛てる。各問合せの後、結果は、既知のドキュメントの集合と比較され、新しければ、追加される。

【００１２】

しかしながら、検索エンジンが、単一の問合せに返し得る結果の最大数を制限するのは一般的である。メタサーチ部は、累進的に更に空間的に限定する後続の問合せを発行して、結果の制限に対応し得る。累進的に更に空間的に限定する一連の例として、“Ｎｅｗ　Ｙｏｒｋ　ｓｔａｔｅ”、“Ｎｅｗ　Ｙｏｒｋ、ＮＹ”、“Ｔｉｍｅｓ　Ｓｑｕａｒｅ、Ｎｅｗ　Ｙｏｒｋ、ＮＹ”等がある。その問合せ範囲を累進的に狭めることによって、メタサーチ部は、結果の数が制限内に

収まるまで、結果の数を減らす。累進的に空間的に限定すると、特定の空間位置により緊密に一致する情報並びに任意の検索エンジンから入手可能な包括的な結果のサンプルが生成される。同時に、前の問合せは一般的であるため、どんな結果も逃さないように、なるべく広い範囲に網が投げられる。その結果、メタサーチ部が見いだすドキュメントは、巡回部が開始するための多種多様であるが空間的に高度に洗練されたサンプルを形成する。

【００１３】

一般的に、１つの側面において、本発明は、表示装置を有するコンピュータシステムに一組の機能を実行させるためのコンピュータ可読媒体上に記憶されたインタフェースプログラムである。これらの機能は、自由テキスト項目問合せと、ドメインを識別するドメイン識別子とを含む検索基準をユーザから受信する機能である。更に、ユーザからの検索基準の受信に応じて、対応するレコードを各々識別する複数のレコード識別子において、（１）ドメイン識別子によって識別されたドメイン内の特定の位置でそれを位置付ける位置識別子をそれに対応させる前記レコードであって、また、（２）自由テキスト項目問合せに応答する情報が含まれる前記レコードを、識別する前記レコード識別子を検索する機能である。更に、表示装置上にドメインの一覧を表示する機能と、複数のレコード識別子によって識別されたレコードの表示として、複数のアイコンを表示装置上に表示する機能である。複数のレコード識別子各々に対して、複数のアイコンの内対応する１つが、表示装置上に表示されているドメインの表示内に表示される。複数のレコード識別子各々に対応するアイコンは、対応するレコード用の位置識別子に対応するドメイン内にある座標のドメインの表示内に位置付けられる。

【００１４】

好適な実施形態には、以下の特徴が１つ以上含まれる。ドメインは、地理的な領域であり、表示は、地理的な領域の多次元地図である。更に、具体的には、表示は、地理的な領域の２次元地図である。入力を受信する段階には、更に、指定されたカテゴリのユーザによる指定を受信する機能が含まれ、複数の検索されたレコード識別子に対応するレコードには各々、指定されたカテゴリ内にある情報が含まれる。あるカテゴリのユーザによる指定を受信する段階には、予め定義さ

れたカテゴリのリストをユーザに提示する機能と、指定されたカテゴリとして、そのリストからユーザによる選択を受信する機能と、が含まれる。また、インタフェースプログラムは、コンピュータに更に以下の機能を実行させるためのインタフェースプログラムであって、複数のレコード識別子各々に対応するアイコンを表示した後、ユーザから更なる検索基準を受信する機能を実行させる。この更なる検索基準は、ドメイン識別子入力タイプ、自由テキスト項目問合せ入力タイプ、及びカテゴリタイプから成る検索基準タイプの群から選択される。また、ユーザからの更なる検索基準の受信に応じて、コンピュータに次の機能を実行させる。すなわち、（１）複数のレコード識別子の部分集合を検索する機能であって、複数のレコード識別子の部分集合は、更なる検索基準内にある複数のレコード識別子間の全レコード識別子を識別する前記機能と、（２）改版された地理的な領域の２次元地図を更なる検索基準に応答する表示装置上に表示する機能と、（３）複数のレコード識別子における部分集合の各レコード識別子に対して、表示された地図内に対応するアイコンを表示する機能であって、複数のレコード識別子における部分集合のレコード識別子各々に対応するアイコンは、対応するレコード用の位置識別子に対応する座標の表示された地図内に位置付けられる前記機能と、（４）更なる検索基準と組み合せて、最初に述べた検索基準をフィルタとして記憶する機能であって、記憶されたフィルタは、インタフェースを介して、更なる検索の指定にユーザが用いるために検索可能である前記機能と、である。更なる検索基準と組み合せた最初に述べた検索基準は、順序付けした一連の入力であり、また、記憶されたフィルタは、一連の入力順を維持する順序付けした一連の入力である。

【００１５】

また、好適な実施形態には、以下の特徴が１つ以上含まれる。インタフェースプログラムは、コンピュータに更に以下の機能を実行させる。すなわち、表示装置を介して、ユーザに地図を提示する機能と、表示された地図と対話することによって、検索基準の一部としてユーザがドメイン識別子を入力できるようにする機能と、を実行させる。複数のアイコンには、第１アイコンクラスのアイコンと第２アイコンクラスのアイコンとが含まれ、また、第１アイコンクラスのアイコ

ンは、第１視覚特性を有し、また、第２アイコンクラスのアイコンは、第１アイコンクラスに対応する視覚特性とは異なる第２視覚特性を有する。複数のレコード識別子によって識別された少なくとも幾つかのレコードは、第１タイプであり、複数のレコード識別子によって識別された少なくとも幾つかの他のレコードは、第２タイプであり、また、第１タイプのレコードは、第１アイコンクラスのアイコンを用いて表示され、第２タイプのレコードは、第２アイコンクラスのアイコンを用いて表示される。複数のアイコンの内、少なくとも１つのアイコンは、複数のレコード識別子によって識別される多数のレコードを表し、多数のレコードは各々、中央位置周辺付近内において、そのレコードを位置付ける位置識別子を有する。

【００１６】

また、好適な実施形態において、インタフェースプログラムは、コンピュータに更に以下の機能を実行させる。すなわち、ユーザからの縮尺要求の変更を受信する機能と、縮尺要求の変更受信に応じて、最初に述べた複数のアイコンにおけるアイコン数より数が少ない第２複数のアイコンを形成するために、少なくとも幾つかの複数のアイコンを互いに統合する機能と、を実行させる。更に、縮尺要求の変更受信に応じて、複数のレコード識別子によって識別されたレコードの位置を識別するために、小さい縮尺を用いて、また、第２複数のアイコンを用いて、ドメインを再表示する機能を実行させる。また、更に、コンピュータに更に以下の機能を実行させる。すなわち、地図内に関連する位置を有する電子メモの指定をユーザから受信する機能と、関連する位置に対応する位置の地図上に貼付メモアイコンを表示する機能と、を実行させる。電子メモには、それ自身の外部アクセス可能なアドレスであって、そのアドレスを介してそのコンテンツに電子的にアクセスを可能にするアドレスを有するウェッブページが含まれる。

【００１７】

一般的に、他の側面において、本発明は、上述した機能を実行する方法である。

一般的に、更に他の側面において、本発明は、コンピュータシステムに以下の機能を実行させるためのコンピュータ可読媒体上に記憶されたデータベースシス

テムである。すなわち、（１）テキスト、（２）ドメインを識別するドメイン識別子、及び（３）フィルタを識別するフィルタ識別子の内、少なくとも１つが含まれる検索基準を受信する機能と、複数のレコード識別子であって、その各々が、検索基準のテキスト、ドメイン識別子、又はフィルタ識別子をそれと関連付けた対応するレコードを識別する前記レコード識別子を検索する機能と、を実行させる。検索機能は、空間キーワードドキュメント索引で実行される。

【００１８】

一般的に、更に他の側面において、本発明は、複数のレコード識別子を検索する方法であって、複数のレコード識別子各々は、検索基準によって指定されるテキスト、ドメイン識別子、及び層識別子の内、少なくとも１つをそれと関連付けた対応するレコードを識別し、空間キーワードドキュメント索引での検索機能が実行される。

【００１９】

好適な実施形態には、以下の特徴が１つ以上含まれる。空間キーワードドキュメント索引には、ドキュメントと、空間索引ツリーと同じ構造を有する複数のツリーと、を参照するためには拡張されるが、特定の用語集項目やフィルタに対しては修正される空間索引ツリーが含まれる。複数のレコード識別子は、空間キーワードドキュメント索引ツリーにおいて検索され、また、ツリーの分岐構造が解析されて地理的な現象が識別される。地理的な現象は、所定のわずかな枝が、所定の数より多い親ノードを共有する空間キーワードドキュメント索引ツリーの分岐構造である。

【００２０】

一般的に、更に他の側面において、本発明は、コンピュータシステムに以下の機能を実行させるためのコンピュータ可読媒体上に記憶されたプログラムである。すなわち、ドキュメントアドレスによって参照されるドキュメントを読み込む機能と、更に多くのドキュメントアドレスを読み込むために、それらのドキュメントを構文解析する機能と、可能な空間識別子が無いか、それらのドキュメントを構文解析する機能と、を実行させる。

【００２１】

好適な実施形態には、以下の特徴が１つ以上含まれる。このプログラムは、更に、コンピュータシステムに以下の機能を実行させるためのプログラムである。すなわち、ドメイン中における位置を決定するために、可能な空間識別子を解析する機能を実行させる。一部のドキュメントアドレスが、メタ検索部プロセスによって収集され、メタ検索部プロセスは、空間ドメインを参照するテキストを用いて、他のコンピュータシステムに問合せる。このプログラムは、更に、コンピュータシステムに以下の機能を実行させる。すなわち、各ドキュメントに見いだされた各可能な空間識別子に対する関連度スコアを算出する機能を実行させる。関連度スコアには、（１）ドキュメントにおける可能な空間識別子の位置と、（２）ドキュメントにおける他の可能な空間識別子の数と、（３）可能な空間識別子が、文中又は孤立しているか否かと、（４）可能な空間識別子中の文字のフォーマット化された強調と、の内の１つ以上が含まれる。このプログラムは、更に、コンピュータシステムに以下の機能を実行させる。すなわち、読み込み前に、関連度スコア毎にドキュメントアドレスを記憶する機能を実行させる。

【００２２】

一般的に、更に他の側面において、本発明は、コンピュータシステムに以下の機能を実行させるためのコンピュータ可読媒体上に記憶されたプログラムである。すなわち、ドキュメントアドレスによって参照されるドキュメントを読み込む機能と、更に多くのドキュメントアドレスを読み込むために、それらのドキュメントを構文解析する機能と、可能な空間識別子が無いか、それらのドキュメントを構文解析する機能と、ドメイン中における位置を決定するために、可能な空間識別子を解析する機能と、を実行させる。

【００２３】

本発明の１つ以上の実施形態に関する詳細は、添付の図面及び以下の説明中に記載する。本発明の他の特徴、目的、及び利点は、この説明及び図面から、更に、請求項から明らかになるであろう。

【００２４】

詳細な説明

図１において、通常、コンピュータシステム２０は、記憶装置２２システムを

含み、記憶装置２２には、ドキュメント形式の情報とそのドキュメントに関する空間情報とが含まれる。また、コンピュータシステム２０は、データ収集３０、データ分析４０、検索５０、データ提示６０、及びポータルサービス７０用のサブシステムを含む。更に、コンピュータシステム２０は、様々なクライアントによってユーザに提供される地図インタフェース８０を含む。地図インタフェース８０によって、ユーザは、記憶装置２２への問合せが可能であり、また問合せ結果の一覧を地図に配置して閲覧できる。

【００２５】

ドキュメント

ユーザの問合せ対象は、ドキュメントである。ドキュメントの例には、テキストベースのコンピュータファイルに加え、部分的にテキストベースのファイル、非テキストファイル、空間情報を含むファイル、及びドキュメントライクなインタフェースを介してアクセスし得るコンピュータエンティティが含まれる。ドキュメントは、他のドキュメントを含むことができ、又、ドキュメントライクなインタフェースに加えて他のインタフェースを含み得る。各ドキュメントは、アドレスを有する。通常、ワールドワイドウェブ・ドキュメントの場合、このアドレスはＵＲＬである。ＵＲＬの場合と同様、アドレスの一部は、ドキュメントを供給するコンピュータサーバプロセスに渡される命令又はパラメータを含み得る。

【００２６】

ドキュメントは、非公開ネットワークやインターネット等のコンピュータネットワーク全体に配置されたコンピュータシステム上に存在する。ドキュメントは、ハイパーリンク化してもよく、すなわち、他のドキュメントのアドレスを含み得る。ドキュメントのコピーは、ページリポジトリ２２２（図３）に記憶してもよい。

【００２７】

汎用ユーザインタフェース

図２において、地図インタフェース８０が、ユーザインタフェースを有するコンピュータ処理装置上のユーザに提示される。ユーザインタフェースは、グラフィカル（ＧＵＩ）、音声ベース、又はテキスト専用ユーザインタフェースでよい

。各ＧＵＩ機能は、可能な限り、音声ベース又はテキスト専用ユーザインタフェースにおいて再現される。

【００２８】

この技術分野では、通常、ＧＵＩは、マウス、接触感知領域、又は方向ボタンの組合せ等、ユーザのポインティングデバイスの操作に反応するポインタ記号を含む。ポインタ記号は、ＧＵＩコンテンツに一致させる。また、ＧＵＩは、ユーザによるクリックイベントにも反応する。通常、クリックは、ポインティングデバイス上やその付近でのユーザのボタン操作に対応するが、コンピュータ処理装置やそのオペレーティングシステムによって、他の方法でも実施し得る。クライアントプロセスは、コンピュータ処理装置のオペレーティングシステムからクリックイベントとポインタ記号の位置とを受信する。

【００２９】

地図インタフェース８０は、地図８０５を含む。地図８０５は、部分的な場合が多いが、少なくとも１つの空間ドメインの一覧である。空間ドメインとは、空間認識部４８には既知である位置尺度を有する何らかの空間である。１つの実施形態において、地球表面は、緯度と経度との２次元位置尺度下の空間ドメイン（以降、標準地理ドメイン）である。他の実施形態において、“ＧＰＳドメイン”は、３次元尺度のＧＰＳ（全地球測位衛星）データの下における地球表面周囲の体積によって定義される。

【００３０】

空間上の尺度は、空間点の位置を識別する必要はない。ドキュメントは、空間点の位置付近にあると識別し得る。例えば、ドキュメントは、“ペンシルベニアのＩ－８０から出口１９付近“と識別し得る。ドキュメントは、プラム島州立公園のような拡張領域を指し得る。

【００３１】

地図８０５は、ドメインの表示に縮尺を用いる。縮尺は、ドメインのどの部分集合が地図８０５に表示されるかを示す。通常、任意のドメインに適した範囲の縮尺がある。小さい縮尺を選択すると、ユーザは、単位当たりの詳細と引き換えに、ドメイン全体の内の狭い部分を調べ得る。

【００３２】

ドメインには、１つの空間的な連続体に沿って存在し得るドメインが幾つかある。例えば、１つの実施形態において、地図８０５は、初期的には、標準地理ドメイン下の地球表面全体の一覧を表示する。次に、地図８０５を変更して、１つの大陸のみ表示し得る。地図８０５は、この時もドメインの一部を表示するが、縮尺が変更されている。

【００３３】

しかしながら、縮尺は、地図８０５が接写したコンサートホールのみを表示する点まで変わり得る。この時点では、地図インタフェース８０は、ドメインをコンサートホールのドメインに変え得るが、ここで、位置は、例えば、区画、列、及び座席番号によって表し得る。従って、ドメインは、連続の縮尺に沿って交差又は重なり合い得る。

【００３４】

ドメインが地理的な意味を有する場合、地図８０５は、通りや水路等、標準的な地図の機能を含み得る。地形用のデータは、米国国勢調査局、米国国土地理院、及びＬａｆａｙｅｔｔｅ・Ｓｔｒｅｅｔ・１１、Ｌｅｂａｎｏｎ、NHのＧＤＴ社、又はＨｉｇｇｉｎｓ・Ｒｏａｄ・１０４００Ｗ、Ｒｏｓｅｍｏｎｔ、ＩＬのＮａｖＴｅｃｈ社等の企業から入手可能である。地図には、縮尺が個々の通りを表示できる充分な程度になるまで現れない通りの名称等、同じドメイン内において、ある縮尺にのみ対応する空間的な陸標地形を含み得る。

【００３５】

地図８０５は、物理的に存在するドメインを表示する必要はなく、地図８０５は、計画住宅開発の仮想レイアウト等、それ自体が一覧であれば何でも表し得る。また更に抽象的には、地図８０５は、空間寸法が実際の空間寸法に対応しない空間レイアウトにおいてエンティティを表示し得る。例えば、ドメインは、平面上に配置された系図でもよく、ここで、平面の１軸は、時間の直線的推移を表す。

【００３６】

ドメイン位置とは、地図８０５が表す空間の位置である。ドメイン位置は、表

示位置と区別すると役に立つが、この表示位置は、地図８０５が表示する要素の配置を示す。ドメイン位置間の距離の大きさは、ドメインに依存するのに対して、表示位置間の距離は、クライアント６４のホストであるコンピュータ処理装置の画素数で測定される。

ドメインフレームとは、ある状態の地図８０５が表示するドメイン全体の部分集合である（全ドメインを含む場合もある）。

【００３７】

地図表示図の変更

ユーザは、地図８０５が表示する図を幾通りかの方法で調整し得る。

ユーザは、ズームバー８９１をクリックして、地図８０５の縮尺を変更できる。ズームバー８９１は、地図インタフェース８０が地図８０５に表示可能な複数の縮尺を視覚的に表す。任意の状態においてズームバー８９１が表示する縮尺は、地図インタフェース８０が地図８０５に表示可能な全縮尺の部分集合であってもよい。この部分集合は、縮尺の変更を含み、状態の変化に適応し得る。例えば、地理的な意味合いにおいて、最初の状態の地図インタフェース８０は、地図８０５に地球全体を表示し得る。この最初の状態において、ズームバー８９１は、例えば、地球的レベルから通りのレベルまでの縮尺を表示し得る。ユーザが、表示領域の範囲をコンサートホールに対応する通りの住所に狭める場合、ズームバー８９１は、ホール内の座席区画から個々の座席までの縮尺を表示する。

【００３８】

ユーザが地図枠８９２をクリックすると、地図８０５の中心は移動できる。地図枠８９２は、地図８０５を取り囲んでいる。

地図モード制御８３０との対話によって、ユーザは、地図インタフェース８０が、地図８０５上のクリックにどのように応ずべきか指定できる。地図モード制御８３０は、パン８３２、ズーム８３４、及びメモ記入８３６のための制御を含む。パン８３２制御及びズーム８３４制御は各々、“オン”状態と“オフ”状態とを含む状態を有する。パン８３２制御が“オン”状態である場合、地図８０５をクリックすると、地図インタフェース８０は命令を受けて、クリックで示した位置周辺の地図８０５を中心に再配置する。同様に、ズーム８３４制御が“オン

”状態である場合、地図８０５をクリックすると、地図インタフェース８０は命令を受けて、クリックで示した位置周辺の地図８０５をズームする。メモ記入８３６制御は、以下の電子メモの項で説明する。

【００３９】

地図インタフェース８０は、空間基準項目制御８０６を含む。空間基準項目制御８０６は、データ項目制御８０８、書込制御８０９、及び空間基準用プロンプト８０７を含む。空間基準用プロンプト８０７は、データ項目制御８０８の目的をユーザに指示する。空間基準用プロンプト８０７は、静的な指示を含んでもよく、あるいはデータ項目制御８０８上のポインタ記号の動き等、ユーザ対話に動的に反応し得る。空間基準用プロンプト８０７は、音声を含み得る。ユーザは、書込制御８０９を呼出して、データ項目制御８０８のデータが完全であることをクライアントプロセス６４に通知する。

【００４０】

空間基準の例には、緯度、経度、又は高度等の地理的測定値、郵便住所情報、あるいはコンサートホールの例に戻ると、列及び座席番号が含まれる。また空間基準は、間接的に空間的な基準、すなわち、空間属性は表さないが、空間属性さえあればどのようなエンティティも指定する基準を含む。間接的な空間基準の例には、小包の追跡番号がある。追跡番号は、それ自体に空間的な意味合いは無いが、任意の時点で、小包の最新の位置を地図８０５上に表示して知り得る。

【００４１】

地図インタフェース８０は、キーワード項目制御８０１を含む。キーワード項目制御８０１は、データ項目制御８０３、書込制御８０４、及びキーワード用プロンプト８０２を含む。キーワード用プロンプト８０２は、データ項目制御８０３の目的をユーザに指示する。空間基準用プロンプト８０７と同様に、キーワード用プロンプト８０２は、静的な指示や音声を含み得、また、ユーザ対話動作に動的に反応し得る。キーワード項目制御８０１内の書込制御８０４の役割は、空間基準項目制御８０６内の書込制御８０９の役割と同じである。

【００４２】

一部又は全ての空間基準項目制御８０６及びキーワード項目制御８０１は、同

じインタフェース構成要素を用い得ることに留意されたい。例えば、ユーザが、“ＭＡ州ケンブリッジ付近の靴”とテキストを入力した場合、システムは、これを空間基準及びキーワード基準の両方であると見なし得る。

【００４３】

キーワードの例には、ユーザが関心を持つ何らかの単語又は単に文字列が含まれる。コンピュータシステム２０は、データ項目制御８０３のデータを記憶装置２２にあるドキュメントのコンテンツと比較する。ドキュメントと一致し得るキーワードに対する所定の制約は無い。

【００４４】

ユーザは、全てのテキスト入力ツールにおいてユーザが望むいかなるテキストも入力し得る。コンピュータシステム２０は、項目を構文解析して、可能なドメイン変更命令やキーワード問合せを得る。キーワード問合せは、如何なる形態でもよい。コンピュータシステム２０は、予め定義されたカテゴリに項目を限定しない。その代わり、コンピュータシステム２０は、問合せテキストとコーパスの全ドキュメントにあるテキストとの照合を試みる。

【００４５】

問合せテキストの照合方法の１つは、空白で分割した個々の文字列にそのテキストを分割することであるが、ここで、この技術分野において、空白は、タブ、スペース、改行復帰、及び正規表現文字“＼ｓ”で通常参照される他の文字として通常定義される。次に、これら個々の文字列は各々、ドキュメントにおいて検索し得る。

【００４６】

同様に、ドキュメントのテキストコンテンツも、空白で分割した個々の文字列に分割し得る。従って、ユーザが入力したテキストがドキュメントコーパスのいずれかの文字列と一致する場合、コンピュータシステム２０は、結果を取得し得る。

この“自由テキスト項目問合せ”によって、所定のカテゴリによる検索に比べて、更により多様な検索が可能になる。

【００４７】

アイコン

地図インタフェース８０は、地図８０５に重ね合わせたアイコン８１０を１つ以上含み得る。アイコン８１０は、クライアント６４がユーザに最初に提供される場合、地図インタフェース８０に存在する必要はない。しかしながら、ユーザが問合せを送信した後、地図インタフェース８０は、アイコン８１０を用いて、検索５０プロセスが決定する程度に問合せ基準を満たす記憶装置２２のドキュメントを表示し得る。

【００４８】

アイコン８１０の表示位置は、そのドキュメントと、対応するドメイン位置との間の相関関係を表す。特に、任意のアイコン８１０がドメイン位置を有する場合、また各ドキュメントがアイコン８１０に対応している場合、データ分析２０用のサブシステムは、ドキュメントがドメイン位置に関係すると判断済みでなければならない。データ分析２０用のサブシステムは、ドキュメントに対するその位置のユーザ入力から、このような関係を判断する。ドキュメントは、２つ以上のドメイン位置に関係でき、従って、２つ以上のアイコン８１０で表し得ることに留意されたい。

【００４９】

個々のアイコン８１０は、１つのアイコンクラスに属する。同じアイコンクラスのアイコン８１０は、形状、色、大きさ、指標付け方式（例えば、ローマ数字対文字）、又はアニメーション動作を含み得る視覚特性を共有する。アイコン面８１８は、アイコンクラスの要件を満たす地図インタフェース８０のインタフェース要素である。１つの実施形態において、クライアントプロセス６４は、画素サイズが約０．２８ｍｍのモニタを備えるコンピュータ上で動作するが、この画素サイズは、本発明を行った時点において、ほぼデスクトップコンピュータの業界標準である。この画素サイズの場合、代表的なアイコンの直径は、１５乃至２０画素である。

【００５０】

アイコンクラスの要件を満たす方法は、複数あり、従って、アイコンクラスは、複数のアイコン面８１８を持ち得ることに留意されたい。例については、下記

のアイコン下位クラスを参照されたい。

【００５１】

アイコンクラスがもたらす視覚的な類似性を用いて、アイコン８１０が表すドキュメント間の項目別の類似性を表し得る。例えば、レストランのメニューに関連するドキュメントは、フォークやナイフの形状を共有するアイコン８１０によって表示し得る。フォークやナイフの形状は、アイコンクラスの特性である。

【００５２】

異なるアイコン８１０の色、形状、色合い、及びアニメーション動作によって、異なるドキュメントの機能をアイコン８１０で表し得る。

１つのクラスのアイコンは、同じ幾何学的な形状を共有し得るが、異なる色すなわち異なる濃淡の同じ色を有し得る。異なる濃淡によって、幾つの異なる特性のドキュメントをアイコンで表し得る。異なる特性のドキュメントには、ドキュメントが作成されてからの経過時間、ドキュメントがシステムに導入されてからの経過時間、ドキュメントの関連度、及びドキュメントサイズが含まれる。

【００５３】

アイコンクラスの他の特徴は、アイコン下位クラスである。２つのアイコンクラスは、第３クラスに関連する特性を共有するが、少なくとも１つの他の特性が一貫して意味があるように変わる場合、第３クラスの下位クラスであり得る。例えば、レストラン用のアイコンクラスは、ある新聞のレストラン批評で評価された質に対する下位クラスを有し得る。レストランの質に対するアイコン下位クラスのアイコン８１０は全てフォークやナイフの形状を共通に持つが、アイコン８１０は、良い批評の場合は緑、悪い批評の場合は赤、又は批評が分かれる場合は黄と色付けし得る。アイコン８１０は、更に円グラフを用いて分割し、各種の批評の割合を示し得る。従って、広範な視覚的類似性を用いると、１つのレベルで広範な項目別の類似性を意味することができ、一方、視覚的副次変化を用いて、第２レベルでの項目別の副次変化を意味することができる。アイコン凡例８１２を用いると、このような取決めをユーザに通知し得る。

【００５４】

アイコンクラスＢが、アイコンクラスＡの下位クラスである場合、アイコンク

ラスＡは、アイコンクラスＢの親クラスである。

多くのドメイン位置は、その位置を参照する複数のドキュメントを有する。これをユーザに示すために、地図インタフェースのその位置で用いられるアイコン８１０は、他のアイコン８１０とは、大きさ、色、又は形状が異なる。例えば、アイコン８１０は、幾つかのアイコン８１０がほぼ重なり合うように、積み重ねて見えるように作成し得る。他の例では、アイコン８１０は、異なるアイコン８１０を部分的に接合しているように見せてもよい。

【００５５】

好適な実施形態において、色が異なるアイコン８１０で異なる層のドキュメントを表し、形状を変えて様々な数のドキュメントを表し、濃淡を変えて、その背後にあるドキュメントに対する様々な関連度の数を表す。与えられたドメイン位置を参照する１組のドキュメントの関連度は、個々のドキュメントの関連度を平均又は合計することによって計算し得る。

【００５６】

アイコン８１０は、ドメインの１つの位置又は幾つかの隣接する位置を表し得る。位置の数は、地図８０５の縮尺を含む幾つかの要因に依存する。複数のアイコン８１０が、コンピュータシステム２０が決める許容範囲内に表示位置を有する場合、地図インタフェース８０は、アイコン８１０を統合して、視覚的明瞭性を高める。ユーザが、地図をズームしてその縮尺を変更する場合、地図インタフェース８０は、アイコン８１０を統合するか否か再計算する。アイコンの統合がそれを越えて行われる許容差は、変更し得る。統合決定の主な要因は、アイコン８１０が重なり合っているか否かである。多くのアイコン８１０の場合、重なり合いの良い判断方法は、表示位置が、アイコン面８１８の平均直径の２倍よりも近接しているか否かである。統合決定の他の要因には、アイコン面８１８の視覚的特性、地図８０５の視覚的特性、ドメインの特性、ドキュメントの特性、及び表示部に現存するアイコンの数及び種類が含まれる。

【００５７】

統合アイコン８１０は、複数の空間ドメインを表し得る。例えば、ワシントンＤＣを含む標準地理ドメイン及びコンサートホールの座席尺度におけるフォード

劇場用の他のドメインについて考える。ある表示縮尺において、リンカーン記念館を表すドキュメントは、リンカーンが撃たれたフォード劇場の特定の席について述べたドキュメントと同じアイコンに統合される。本例において、リンカーン記念館ドキュメントは、標準地理ドメインに関連する。フォード劇場ドキュメントは、フォード劇場に固有のドメインに関連するが、本例では、フォード劇場ドメイン全体が、ユーザが要求するドメインサイズに比べて、かなり小さい領域上にマッピングし得るため、標準的な地理に表示し得る。

【００５８】

またアイコン８１０は、アイコン８１０が統合されるか否かに関係なく、そのドキュメント間における複数の項目別カテゴリを表し得る。この場合、多数の項目を反映するためにアイコン面８１８を変更し得る。

【００５９】

アイコン凡例８１２は、地図インタフェース８０の他の要素である。アイコン凡例８１２は、アイコン８１０をそれが表すドキュメントに関連付ける。アイコン凡例８１２は、ドキュメントリストから構成される。このリストは、様々にグループ分け又は順序付けし得る。

【００６０】

アイコン８１０は、検索５０プロセスが編集した順序に従って、アイコン凡例８１２にリスト化される。

非統合アイコン８１０は、単一の表示位置を表す。アイコン凡例８１２にリスト化されたそのドキュメントの順序は、検索５０プロセスが編集した関連度順位に基づく。関連度順位は、ユーザの問合せ基準に対して各ドキュメントを評価する。

【００６１】

統合アイコン８１０は、複数のドメイン位置を表し得る。統合アイコン８１０は、複数のアイコンクラスを表し得る。アイコンクラスが異なれば、項目別カテゴリも必然的に異なる。アイコン凡例８１２は、これらの項目別カテゴリに基づきドキュメントリストを区分し得る。例えば、それらのカテゴリ毎のグループ化、リストの各項目へのフィールド追加によるカテゴリの指定、又は視覚的強調の

追加によって、区分し得る。視覚的強調は、字体の変更、色の変更、又はカテゴリに関連するアイコンタイプの存在を含み得る。隣接グループ間の背景色の変化と組み合せたカテゴリ毎のグループ化等、幾つかの効果を組み合せ得る。

【００６２】

フィルタ

地図インタフェース８０は、フィルタを管理するための２グループの制御、すなわち、汎用フィルタ表示８５０及びユーザ専用フィルタ表示８６０を含む。

図３において、フィルタは、ページリポジトリ２２２におけるドキュメントのコーパスの部分集合を選択する。フィルタは、再帰的に定義され、フィルタは、要素のリストであり、ここで、各要素は、キーワード文字列、１組の空間基準、人が編集したリストのドキュメント、ドメインフレーム、又は他のフィルタのいずれであってもよい。要素は、ユーザがドキュメントの集合を選択し得る順序に定義し得る。フィルタの順序は、ブール演算子ＡＮＤと組み合せて、如何なる順序のフィルタに対しても同じの交差ドキュメント集合を生成し得る。２組のフィルタは、ブール演算子ＯＲと組み合せてもよい。地図８０５の１組のドキュメントを閲覧する場合、ユーザは、地図表示を変更して、このドキュメント組の部分集合を表示し得るが、この部分集合は、地図表示を変更した後ユーザがフィルタ処理を行った場合とは異なる場合がある。従って、ユーザの問合せ毎にフィルタが定義されるが、これは、フィルタが、キーワード、空間基準、ドメインフレームへの変更、又はこれら幾つかの組合せのいずれかを含むためである。地図インタフェース８０の初期状態によって、ユーザがそれとまだ対話していなくても、フィルタが定義されるが、これは、地図８０５が、それに対応する少なくとも１つのドメインフレームを有するためである。同様に、空状態ではない地図８０５によってフィルタが定義されるため、地図８０５をズーム又はパンすると、前のフィルタと新しいドメインフレームに基づき、新しいフィルタが常に定義される。アイコン８１０の各グループは、それ自体の固有フィルタ、すなわち、地図８０５の現在の状態によって定義されるフィルタを定義するが、その結果得られるドキュメントは、そのグループの少なくとも１つのアイコン８１０に対応したドキュメントに限定される。このように、アイコン８１０をクリックすると、フィ

ルタを定義できるが、これは、単一のアイコン８１０は、単に１つのグループのアイコンであることによる。

【００６３】

汎用フィルタ表示８５０は、ユーザ用に作成されたフィルタを含む。ユーザ専用フィルタ表示８６０は、ユーザが作成したフィルタを含む。２組の制御８５０、８６０は、切り離してもよく、あるいは地図インタフェース８０の制御を共有し得る。

【００６４】

汎用フィルタ表示８５０は、汎用８５２フィルタ、検索履歴８５４フィルタ、及び推定８５６フィルタを含む。汎用８５２フィルタは、コンピュータシステム２０が予め定義したフィルタである。これには、ユーザ集団が一般的に関心を持っているとして編集人が選んだフィルタと、ユーザ集団の利用パターンにおいて繰返し頻度が高いためにアルゴリズム的に選択されるフィルタとが含まれる。検索履歴８５４フィルタは、現ユーザが、それを記憶するようにシステムに明確に命令せずに、現在又は前のセッションに適用した可能性のあるフィルタである。検索履歴８５４フィルタへの簡単なアクセスを提供すると、システムでは、ユーザは、ユーザが以前作成してユーザ専用フィルタ表示８６０へ追加しなかったフィルタを再適用できる。

【００６５】

推定８５６フィルタは、現ユーザの利用パターンに基づき、アルゴリズムで選択されたフィルタである。

データ調査８５７フィルタとは、ページリポジトリ２２２のドキュメントのコンテンツやハイパーリンクを分析して、１つの特性を共有する１組のドキュメントを作成する手続によってアルゴリズムで作成されるフィルタである。この特性は、例えば、“料理法に関連すると思われる全ドキュメント”のように、発見的に決定し得る。このようなフィルタを構築するアルゴリズムには、ベイズ学習法、統計的分析、及び単語や句の存在論の使用法が含まれる。

【００６６】

ユーザ専用フィルタ表示８６０は、地図インタフェース８０の状態によっては

示されないことがある。例えば、コンピュータシステム２０が、正しいユーザプロフィールを判断できず、現ユーザに適用できない場合、又はプロフィールに対応するセキュリティ基準を満たせない場合、ユーザ専用フィルタ表示８６０は、非表示又は無効にできる。

【００６７】

表示され有効になった場合、ユーザ専用フィルタ表示８６０は、ユーザプロフィールに対応するフィルタを含む。ユーザは、これらのフィルタを追加、変更、又は削除でき、またユーザ定義のグループに割当て得る。

【００６８】

ユーザがユーザ専用フィルタ表示８６０に追加できるフィルタには、汎用フィルタ表示８５０のフィルタ、地図８０５の現在の状態で定義されるフィルタ、ユーザがポインタ記号を用いて指定可能なあるグループのアイコン８１０で定義されるフィルタ、少なくとも２つの既存のフィルタから組み合せたフィルタ、及びユーザが新しい名前での保存を選択する変更済フィルタが含まれる。

【００６９】

ユーザが、ユーザ専用フィルタ表示８６０のフィルタに適用し得る変更には、フィルタの名前変更と、そのリストでの要素の追加、削除、又は並び替えと、フィルタに対応するアイコンクラスの変更又はそのフィルタ用の新しいアイコンクラスの定義とが含まれる。ユーザが編集し得るアイコンクラスの特性には、その名前と、そのアイコン面８１８と、その親アイコンクラスと、ドキュメントのテキストによる要約と、アイコンクラス凡例８１７に表示される何らかの特性とが含まれる。

【００７０】

電子メモ

メモドキュメントは、ドメイン位置に対応するドキュメントである。またこれは、ユーザプロフィールにも対応付けてもよく、あるいは匿名でも存在し得る。電子貼付メモ８７０は、このメモドキュメントに対応付けたドメイン位置に対応する表示位置において、地図８０５上に表示したメモドキュメントの一覧である。メモドキュメントは、記憶装置２２のドキュメントが含み得るあらゆる形態の

情報を含み得る。例えば、メモドキュメントは、テキスト、図形、音声、映像、ハイパーリンク、又はそれらの組合せを含み得る。メモドキュメントは、それ自体のＵＲＬを持つことができ、またウェブページとして機能し得る。

【００７１】

メモ記入８３６制御は、地図８０５の次のクリックによって、新しいメモドキュメントが生成されるように、地図インタフェース８０の状態を変更する。メモドキュメントは、クリックされた表示位置に対応するドメイン位置に対応付けられ、また電子貼付メモ８７０は、前記表示位置に表示され、その表示位置に表示されるドメイン位置に対応付けられる。

【００７２】

１つの実施形態において、地図インタフェース８０が然るべき状態に置かれた場合、ユーザは、クライアントプロセス外部から地図８０５上にドキュメントコンテンツを移動して、メモドキュメント作成を開始し得る。コンテンツは、コンピュータ処理環境及びメディア種類に適した他の方法の中でも、ドラッグ・アンド・ドロップ又はコピー・アンド・ペーストによって移動し得る。例えば、ドキュメントコンテンツは、コンピュータシステム２０が記録を開始するメディアストリームであってもよい。コンテンツは、新しいメモドキュメントの一部になり、またメモドキュメントには、ＵＲＬ等、少なくとも１つの外部アクセス可能なアドレスが与えられる。地図インタフェース８０が然るべき状態にある場合、ユーザは、例えば、一回の素早い動作で、ウェブページを作成できる。本実施形態において、ユーザがコンテンツをドラッグ・アンド・ドロップ又はコピー・アンド・ペーストできる機構は、オペレーティングシステムによって提供される。用語“ドラッグ・アンド・ドロップ”及び“コピー・アンド・ペースト”は、この技術分野では良く知られている。

【００７３】

これらメモドキュメントの他の機能には、説明が必要なものが幾つかある。ユーザは、ドキュメントが公開されない場合、カレンダの日付及び／又は時刻を指定できるが、さもなければ全て期限切れになる。メモドキュメントは、期限切れになると、記憶装置から削除してもよく、あるいはインタフェースに現れないよ

うにできる。これによって、ユーザは、地理的位置において時間に左右される情報を記入できる。短命のメモドキュメントは、地図インタフェース上にアニメーションアイコンを作成するのに用い得る。このようなアイコンは、移動オブジェクト又はドメインを通るユーザの大体の経路に追従する。

【００７４】

ユーザは、他のユーザに対してメモドキュメントの信頼性を保証するためにメモドキュメントにデジタル署名を行うことができる。ＰＧＰ等の公開鍵暗号法が、この技術分野では標準的であり、これを真似て用い得る。同じ種類の公開鍵暗号法を用いて、又はユーザ識別を認証する非公開パスワードでのログインをユーザに要求して、ドキュメントの閲覧者を限定し得る。メモドキュメントの作成者は、特定のメモドキュメントの閲覧が許可された登録ユーザのリストを決定できる。他の選択肢として、作成者は、メモドキュメントを開くのに必要な暗号鍵を配布し得る。これによって、ユーザは、メモドキュメントを登録リストに掲載し得る。

【００７５】

ユーザは、非公開コンピュータシステム上においてユーザ自身のメモドキュメントのホストとなり得る。このような非公開コンピュータシステムは、コンピュータシステム２０の一部又は全てのコピーが許可されたシステムであってもよい。このような非公開で維持されるメモドキュメントは、セキュリティ対策によって保護し得る。このようなメモドキュメントの作成者は、他の人々又は企業が所有し得るコンピュータシステム２０の他のインスタンスにおいて、新たなメモドキュメントを作成できる。これらの新たなメモドキュメントは、作成者の非公開コンピュータシステム上の１つ又は多くのメモドキュメントにポインタを提供し得る。これらの新たなメモドキュメントは、元のメモドキュメントの要約を含み得る。コンピュータシステム２０の１つのインスタンスのユーザは、ある他のインスタンスのコンピュータシステム２０へアクセスし得る。このアクセス権は、各インスタンスの所有者によって決定される。これによって、コンピュータシステム２０の多くのインスタンスが、地図上に位置付けられたメモドキュメントのホスティング及び配信に参加できる。

【００７６】

如何なるメディアタイプでもメモドキュメントに容易に組み込めるため、コンピュータシステム２０のインスタンスの所有者が、所有者の制御下で他のコンピュータシステムからのデータでメモドキュメントを作成することは容易である。例えば、商店主は、商店主のインスタンスのコンピュータシステム２０においてメモドキュメントに商店主の在庫データベースをコピーできる。このように店舗データベースを地理的に位置付けしたメモドキュメントに変換すると、店舗の実際のエリアに関心を持つ他のユーザに在庫情報を提供することが簡単になる。

【００７７】

ユーザは、地図インタフェースへフォルダのドキュメントをドラッグ・アンド・ドロップする等、一回の動作で、メモドキュメントの集合をアップロード又は作成できる。ドキュメントは、位置情報を含む場合、地図インタフェースに自動的に記入し得る。そうでない場合、ユーザは、各ドキュメントに対する位置の選択を催促されることがある。

【００７８】

このようなメモドキュメントの集合は、ユーザ専用フィルタ表示８６０のフィルタにグループ化される。このようなグループ化したメモドキュメントの例には、休暇で撮影した写真の集合、街周辺で録音した音声記録の集合、様々なセンサから集めた１組のデータ、新聞記事の一連の出来事、又は小道案内用の１組の説明が含まれる。１つの集合は、地図８０５上の様々なアイコンを接続する色付きの線を有してもよく、これによって、ドメインのユーザが追従し得る経路が示される。

【００７９】

このような集合は、サービス又は装置によってユーザ用に作成し得る。例えば、ユーザのカメラには、ＧＰＳ又は各写真に位置スタンプを刻印する他の空間位置特定装置を含む。更に、写真のアップロードは、非常に簡単であり、スタンプを用いると、各写真を地図８０５上に配置できる。ユーザのためにこのようなサービスを行い得る。例えば、病院は、ユーザの医療記録にユーザが処置を受けた位置の注釈を付け、それらをユーザや他の医療提供者用の非公開メモドキュメン

ト集合として記入し得る。

【００８０】

ユーザは、討論掲示板、注文入力ツール、電話接続サービス、又は他のソフトウェアバックアップ式ツール等の動的ソフトウェアを含むメモドキュメントを記入できる。販売機の位置に記入されたメモドキュメントは、ユーザが、クレジットカード又は他の支払方法を用いて販売機から商品を購入できる販売機に接続された注文入力ツールを有し得る。これによって、ユーザは、現金を支払わずに、あるいはクレジットカードを携行さえせずに実際の商品を購入できる。

【００８１】

店舗に記入されたメモドキュメントに、テキスト又は他のメディア入力ツールがある討論掲示板を含むと、一般の人々がその位置での討論に参加し得る。このようなメッセージ掲示板は、携帯電話からのテキストメッセージ送信を受信したり、討論掲示板を閲覧するユーザにそれらを一斉送信したりできる。

【００８２】

メモドキュメントには、クリックするとユーザの電話機がサービスに電話をかけるツールを含み得る。このようなメモドキュメントは、電話予約が必要なレストランや劇場に記入し得る。

【００８３】

コミュニティフィードバック

地図インタフェース８０は、コミュニティフィードバック８８０制御を用いて、他のユーザの挙動から集められたユーザ情報を示し得る。コミュニティフィードバック８８０制御の特徴は、ドメイン利用フィードバック８８２、単語ードメイン候補８８４、及び単語ー単語候補８８６を含む。

【００８４】

ユーザが空間ドメインを閲覧する場合、ドメイン利用フィードバック８８２によって、ユーザは、最近そのドメイン又はそのドメインの一部を何人が閲覧したか知り得る。例えば、“２３人が最近１８分間でこの領域を閲覧した”ことを知り得る。

【００８５】

ユーザが、空間ドメインを閲覧する場合、単語－ドメイン候補８８４は、このドメインに関連するキーワードをユーザに知らせ得る。これらの単語は、この領域を参照するドキュメントを分析して、そのドメインで最も頻発する単語を見つけることによって、集めることができる。また、これらの単語は、この領域を閲覧する際、他のユーザが入力したキーワードを記録することによっても集めることができる。最も多く検索された単語をユーザに提供し得る。

【００８６】

ユーザが、キーワード問合せを入力する際、単語－単語候補８８６は、たった今入力したキーワード（群）に関連する新たなキーワードをユーザに通知し得る。これらの新たなキーワード候補は、他のユーザが入力した一連の問合せを記録することによって構築し得る類語辞典から得られる。多くのユーザが、同じキーワードを一緒に又は単一セッションで入力する場合、これらのキーワードは、関連性があると考え得る。例えば、多くのユーザが、“ｃｈｏｃｏｌａｔｅ”を検索し、次に、“ｃｈｏｃｏｌａｔｉｅｒ”を検索する場合、コンピュータシステム２０は、“ｃｈｏｃｏｌａｔｅ”と入力する次のユーザに“ｃｈｏｃｏｌａｔｉｅｒ”に対するキーワード問合せを試すように提案し得る。この候補は、ユーザが望むものを見つけるのに役立つ。

【００８７】

データ収集

コンピュータシステム２０は、新しいドキュメントを集めるためのデータ収集３０プロセスを含む。図３において、データ収集３０プロセスは、巡回部３６プロセス、ページ待ち行列３４、及びメタサーチ部３２プロセスを含む。

【００８８】

巡回部及びページ待ち行列

巡回部３６は、ネットワークを介してドキュメントを読込み、それをページリポジトリ２２２に保存し、それをハイパーリンクが無いかスキャンする。これらのハイパーリンクを繰返し追従すると、ネットワーク化された体系のドキュメントの多くを見つけることができ、またページリポジトリ２２２に保存し得る。このように、巡回部３６は、コンピュータシステム２０にドキュメントを集める。

１つの実施形態において、これらのドキュメントは、インターネット上で利用可能なワールドワイドウェブ・ページである。この場合、ページのダウンロードは、ハイパーテキスト転送プロトコル（ｈｔｔｐ）、ファイル転送プロトコル（ｆｔｐ）、ゴーファ、ニューズ、ＷＡＩＳ、及びその他のプロトコルを含む様々なインターネットプロトコルのいずれかを用いて行い得る。

【００８９】

ページ待ち行列３４は、ドキュメントアドレスを記憶する。巡回部３６、新規開拓部４８、及びメタサーチ部３２は、ドキュメントアドレスを追加する。ページ待ち行列３４は、データベーステーブル、すなわち、ページ待ち行列テーブル３４０から構成される。

【００９０】

巡回部３６は、ドキュメントアドレスを得て、ページ待ち行列３４から巡回する。巡回部３６が、これまで未知のドキュメントを読込む場合、このドキュメントを新規開拓部４８プロセスに渡す。新規開拓部４８は、新しいドキュメントへのハイパーリンクが無いかドキュメントのコンテンツを解析する。新規開拓部４８は、このようなハイパーリンクが参照するあらゆるアドレスを、ページ待ち行列３４に追加する。

【００９１】

巡回部３６は、空間的に関連する確率はリンクと相関関係にある、すなわち、空間的に関連するページにリンクされるページは、平均の空間的な関連より高い確率を有するという事実を利用する。巡回されたＵＲＬには各々、空間関連度が割当てられる。空間関連度を考慮すると、巡回部３６が、時間及び他の資源を効率的に用いることに役立つ。

【００９２】

巡回部はまず、所定の閾値よりも高い空間関連度を有するページからリンクされたページを巡回する。ページがダウンロードされ、その空間関連度が計算された後、その空間関連度レベル３４２フィールドは、再較正され、見出した実際の関連度を反映し得る。

【００９３】

メタサーチ部

メタサーチ部３２は、既知のドキュメントの集合を初期化する。この初期化ステップは、“シーディング”又は“ブートストラッピング”と呼ばれる。コンピュータシステムは、各ドメイン用にシードしなければならない場合がある。例えば、個別のブートストラッピング動作は、米国や仏国の郵便住所に用い得る。

【００９４】

メタサーチ部は、インターネット上のウェブサイト検索エンジン等、ドメインに対応した情報を記憶することで知られる検索エンジンに問合せを行う。メタサーチ部の管理人は、それに、ドメインに対応する既知の空間位置の集合を提供する。メタサーチ部は、これらの空間位置に基づき問合せを形成し、その問合せを検索エンジンに宛てる。その結果は、既知のドキュメントの集合と比較され、新しければ追加される。

【００９５】

巡回は、ネットワーク上の発見可能なドキュメントが全て見つかると完了する。実際には、このことは、集合が極端に静的でない限り、大きなドキュメント集合上ではほとんど起こらない。従って、完全な巡回は、ほとんど起こらないことから、巡回速度が、重要な設計上の関心事である。巡回速度は、新しいページが、以前ダウンロードされたページ上のリンクを介して発見される速度によって制限される。この巡回を速める良い方法は、ドキュメント集合の少なくとも一部を既に巡回した既存の検索エンジンに問合せることであり、これは、ウェブであり得る。これらの検索エンジンから得られた結果は、データ収集３０プロセスをブートストラップするために用いられる。

【００９６】

１つの実施形態において、メタサーチ部３２は、米国の地理の情報をブートストラップする。このブートストラップのプロセスは、６つのステップから構成される。異なるプロセスを要求する他のドメインもあり得る。

【００９７】

このステップは、既存の検索エンジンから最も有用な空間的ＵＲＬを集めることを意図したレベルのシステムである。通常、検索エンジンは、単一の問合せに

対して返す結果の数を限定することから、検索は、集めたい結果の全てを返さない場合がある。例えば、このことは、地理の問合せで、“ＭＡ州ボストン”のような都市名で起こる。このような場合、その都市にある通りの全ての名称等、問合せに他の単語を指定すると良い。

【００９８】

主な検索エンジンには、ＡｌｔａＶｉｓｔａ、Ｆａｓｔ、Ｌｙｃｏｓ、ＭｅｔａＣｒａｗｌｅｒ、ＤｏｇＰｉｌｅ、ＮｏｒｔｈｅｒｎＬｉｇｈｔがある。各エンジンは、更に多くのページが問合せを満たすと分かっていても、その問合せに対しては、ある最大数の結果を返す。メタサーチ問合せがこの数を超過する場合、メタサーチ部３２は、単語を問合せに加えて、更にＵＲＬを絞り込む。

【００９９】

ステップ１において、メタサーチ部３２は、例えば、“ｂｏｓｔｏｎ”、“ｃａｍｂｒｉｄｇｅ”、“ｎｅｗｙｏｒｋ”、“ｍａｄｉｓｏｎ”、“ｓａｎ ａｎｔｏｎｉｏ”等の都市名のみで、検索エンジンに問合せる。

【０１００】

ステップ２において、ある都市名が、そのエンジンに対して最大数の結果に至る場合、メタサーチ部３２は、例えば、“ｂｏｓｔｏｎ ｍａ”、“ｂｏｓｔｏｎ ｍａｓｓ”、“ｂｏｓｔｏｎ ｍａｓｓａｃｈｕｓｅｔｔｓ”、“ｃａｍｂｒｉｄｇｅ ｍａ”・・・等、“ｎｅｗｙｏｒｋ ｎｙ” ・・・等、“ｍａｄｉｓｏｎ ｎｊ” ・・・等、“ｍａｄｉｓｏｎ ｎｙ” ・・・等、都市と州で検索エンジンに再度問合せる。

【０１０１】

ステップ３において、メタサーチ部３２は、更に情報を含む第２テーブルに切り替わる。第２テーブルは、米国の各都市の通りを全て含む。いずれかの都市－州の対が特定のエンジン上でオーバーフローする場合、メタサーチ部３２は、例えば、“ｈｉｇｈｌａｎｄ ｓｏｍｅｒｖｉｌｌｅ”、“ｈａｎｃｏｃｋ ｓｏｍｅｒｖｉｌｌｅ”、“ｅｌｍ ｓｏｍｅｒｖｉｌｌｅ”等、各通りに対して問合せる。

【０１０２】

ステップ４において、メタサーチ部３２は、例えば、“ｈｉｇｈｌａｎｄ ｓｏｍｅｒｖｉｌｌｅ ｍａ”、“ｈａｎｃｏｃｋ ｓｏｍｅｒｖｉｌｌｅ ｍａ”、“ｅｌｍ ｓｏｍｅｒｖｉｌｌｅ ｍａ”等、通り名称に州名を追加する。

【０１０３】

ステップ５において、メタサーチ部３２は、例えば、“ｈｉｇｈｌａｎｄ ａｖｅ ｓｏｍｅｒｖｉｌｌｅ”、“ｈｉｇｈｌａｎｄ ａｖｅｎｕｅ ｓｏｍｅｒｖｉｌｌｅ”、“ｈａｎｃｏｃｋ ｓｔ ｓｏｍｅｒｖｉｌｌｅ”、“ｅｌｍ ｓｔ ｓｏｍｅｒｖｉｌｌｅ”等、通りの種類を追加する。

【０１０４】

ステップ６において、メタサーチ部３２は、例えば、“ｈｉｇｈｌａｎｄ ａｖｅ ｓｏｍｅｒｖｉｌｌｅ ｍａ”、“ｈｉｇｈｌａｎｄ ａｖｅｎｕｅ ｓｏｍｅｒｖｉｌｌｅ ｍａ”、“ｈｉｇｈｌａｎｄ ａｖｅｎｕｅ ｓｏｍｅｒｖｉｌｌｅ ｍａｓｓａｃｈｕｓｅｔｔｓ”等、通りの種類と州名とを追加する。このレベルに達する場所はほとんど無い。

【０１０５】

ページ問合せテーブル３４０は、空間関連度レベル３４２を含むが、これは、巡回部３６を空間的に関連するドキュメントに拘束するのに役立つ。メタサーチ部３２がドキュメントを集める場合、そのドキュメントには、レベル“ゼロ”が与えられる。

【０１０６】

データ分析

図４において、コンピュータシステム２０は、ドキュメントからの抽出情報及びメタ情報用のデータ分析４０プロセスを含む。データ分析４０は、空間認識部４２プロセス、空間符号化部４３プロセス、キーワード構文解析部４４プロセス、索引付け部４６プロセス、空間ドキュメント順位付け４５プロセス、及び新規開拓部４８プロセスを含む。新規開拓部４８プロセスの役割は、データ収集３０の項で説明する。データ分析の項においては、標準的な緯度／経度によって識別されるが、郵便体系住所、地域、及び電話番号によっても識別される米国用の標準地理ドメインの例を繰返し引用する。

【０１０７】

空間認識部

新しいドキュメントが、ページリポジトリ２２２に保存される際、空間認識部４２は、各ドキュメントを開き、そのコンテンツをスキャンする。それは、空間識別子の一部に類似するパターンを検索する。例えば、米国用の標準地理ドメインにおいて、パターンは、米国郵便体系の通りの住所、地域、及び電話番号を含む。

【０１０８】

ステップ４２２において、空間認識部４２は、非体系化テキストでの候補空間データを捜す。見込のある空間識別子に対する候補空間データは、ＰＳＩと呼ばれる。

【０１０９】

ステップ４２４において、空間認識部４２は、候補空間データのテキストを構文解析して、その構造を判断して、ＰＳＩを形成する。ここで、住所を米国郵便体系で用いられる標準的な一組のフィールドに分断する。同様なフォーマットが他の郵便体系に存在するが、これは、他のドメインとして表される。ＰＳＩの構成部分が識別される。１つのドキュメントに全て存在することはあり得ず、地域及び電話番号の場合、都市、州、及び恐らくＺＩＰやＺＩＰ＋４だけが用いられる。構成部分は、以下のものを含む。

【０１１０】

住居番号

通りの接頭部（例えば、東、南）

通り名称

通りの接尾部（例えば、東、南）

通りの種類（例えば、通り、有料高速道路、広場）

都市

州

ＺＩＰ

４桁拡張ＺＩＰ

【０１１１】

ＰＳＩは、空間用語集２２４に記憶され更に分析される。これらの見込空間識別子（ＰＳＩ）用のテーブルは、この場合、標準地理ドメインに対してマッピングされるが、緯度及び経度用のフィールドを含む。ドメインにかかわらず、テーブルには、空間符号化信頼度、この場所に配置されたドキュメント数、空間符号化の状態、及びこの場所に配置されたドキュメントの関連度の合計を含み得る。

【０１１２】

関連度スコアラ４２６は、関連度スコアをドキュメントに割当てる。

関連度スコアラ４２６は、多重空間基準区画部４２６２プロセスを含む。多くのドキュメントが多重空間基準を有する。全ての空間識別子が、ドキュメント全体に関連する場合がある。例えば、店舗チェーンの支店位置をリスト化したウェブページがある。しかしながら、今度は、各空間識別子が、ページの然るべき部分集合にのみ関連している場合があり得る。この例として、多くのレストランの短評を提供するページがある。このようなページは、複部構成のドキュメントである。

【０１１３】

複部構成のドキュメントには、キーワードによってドキュメント集合を検索する場合、問題が生じる。ドキュメントが一括してキーワードによって索引付けされるならば、ドキュメントのある部分の単語は、実際にはその単語がそのドキュメントの異なる部分の住所に関連し得ない場合も、その部分に関連しているかのように索引付けされる。

【０１１４】

複部構成のドキュメントを検出する場合、多重空間基準区画部４２６２は、複部構成クラスタ測定４２６２５プロセスを呼出す。複部構成クラスタ測定４２６２５プロセスは、まず、ある数の住所（通常は５）よりも少ない又はある数の単語（約２００）より短い全てのドキュメントを拒否する。複部構成クラスタ測定４２６２５プロセスは、ページの各ＰＳＩの小数位置を含むアレイを計算する。例えば、１０００ワードのドキュメントにおける２００番目の単語から始まる住所は、小数位置０．２である。次に、Ｇｉｎｉ係数等のクラスタ統計を適用して

、住所がページ上でどの程度集中しているか表すクラスタスコアを生成する。ドキュメントのクラスタスコアが低い（住所が、均一に分散していることを示す）と、複部構成のドキュメントになりやすい。最大クラスタスコアの閾値は、経験的に決定され、ドメインによって変わり得る。

【０１１５】

多重空間基準区画部４２６２は、次のように、境界としてＰＳＩを用いて、１つのＰＳＩを各々含むセグメントにドキュメントを区分する。ｎ番目のセグメントは、ＰＳＩｎを含み、ＰＳＩｎ－１の末尾に続く単語から始まり、ＰＳＩｎ＋１の前の単語で終わる。ｎ＝１の場合、セグメントは、最初の単語から始まる。ページ上で最後のＰＳＩの場合、セグメントは、ページの最後で終わる。

【０１１６】

更に、各セグメントは、それに加えられたドキュメントの表題部分を有する。タグ認識部４４２は、ドキュメントの表題部分を決定する１つの方法を提供する。

【０１１７】

セグメントは、ページリポジトリ２２２に記憶され個別に索引付けされる。セグメントが検索結果として見つかった場合、全てのドキュメントを返すことができるように、セグメント化されていないページは保持されるが、この場合、ドキュメントは、そのセグメントをユーザに提示する前に、そのセグメントにスクロールし得るように、アンカがセギメントの開始部に配置される。

【０１１８】

空間符号化部

ＰＳＩを更に分析するために、空間符号化部４３プロセスは、ドメイン位置をドキュメントコンテンツの様々な識別子と対応付ける幾つかのプロセスを実行する。標準地理ドメインにおいて、緯度／経度点又は境界ポリゴンを識別子と対応付けし得るが、このプロセスは、居住者地域別分類として知られている。ＰＳＩと一致し得る緯度／経度がない場合、空間符号化部４３は、それを不正認識とマーキングする。そうでない場合、空間符号化部４３は、ＰＳＩを既知の空間識別子、すなわち、ＫＳＩに変える。これによって、上述の空間用語集２２４への入

力が完了する。

【０１１９】

米国の標準地理ドメイン用空間符号化部４３は、住所符号化部４３２、地域符号化部４３４、及び電話番号符号化部４３６を含む。

再び米国の標準地理ドメインにおいて、住所は最良の組合せと見なされる。従って、あるページがそこに住所を有する場合、“ｃａｍｂｒｉｄｇｅ、ＭＡ”等の単純な場所の名前や電話番号は、そのページを空間的に符号化するためには用いない。ページは、複数のＫＳＩを有し得るが、このことによって、その空間関連度（空間ドキュメント順位付け４５を参照）が低くなり、そのために、主に少数の大きく注目されるＫＳＩのみを有するページを探す。注目されるＫＳＩとは、空間符号化部４３が、“緯度／経度空間”（緯度及び経度によって識別される空間）の狭い領域を高い信頼度で関連付けることを意味する。従って、例えば、電話番号は、少なくとも数平方マイルはある電話交換局規模の領域と対応するが、郵便住所は、通常、仮説的な屋根の中心点で表される“屋根”サイズの領域と対応する。ドキュメントの電話番号と住所の両方が、ページの位置で一致する場合、ドキュメントの順位付けを改善し得る（空間ドキュメント順位付け４５）。

【０１２０】

住所符号化部４３２：米国及び他国の郵便住所は、通常、建物サイズの狭い地理的領域と対応付けできる。標準的な居住者地域別分類手続は、これを点で近似する。例えば、以下のようなＰＳＩ、

７７　ｍａｓｓａｃｈｕｓｅｔｔｓ　ａｖｅ｜ｃａｍｂｒｉｄｇｅ｜ｍａ｜０２１３９

が与えられた場合、対応する緯度／経度は、何らかの標準的な住所の居住者地域別分類によるプロダクトにそのテキスト文字列を供給することで見いだせる。例には、Ｅｔａｋ´ｓ　Ｅａｇｌｅｃｏｄｅｒ、Ｓａｇｅｎｔ´ｓ　ＧｅｏＳｔａｎ、及びＥＳＲＩ´ｓ　ＡｒｃＩＮＦＯ居住者地域別分類プラグインが含まれる。Ｅｔａｋ´ｓ　Ｅａｇｌｅｃｏｄｅｒの出力は、以下のようになる。

【０１２１】

＜コマンドラインインタフェース＞ｊｒｆ＠ｒａａｇ：～　ｍｃ／ｌｉｂ／ｅ

ｔａｋ／ｒｉｅ‐ｂ

＜ＰＳＩの入力テキスト＞７７　ｍａｓｓａｃｈｕｓｅｔｔｓ　ａｖｅ｜ｃａｍｂｒｉｄｇｅ｜ｍａ｜０２１３９

＜居住者地域別分類部の出力＞７７　ＭＡＳＳＡＣＨＵＳＥＴＴＳ　ＡＶＥ、ＣＡＭＢＲＩＤＧＥ、ＭＡ、０２１３９、４２．３５８９６８、０７１．０９３９９７

出力の第３ラインは、この住所と対応する緯度／経度情報を含む。従って、このＰＳＩは、ＫＳＩに変換し得る。

【０１２２】

地域符号化部４３４：“ｂｏｓｔｏｎ、ＭＡ”や“ワシントン記念碑”等の場所の名前は、その場所の中心の緯度・経度と共に、米国国勢調査によってリスト化される。これによって、これらの居住者地域別分類が簡単になる。地域符号化部４３４は、住所符号化部４３２と同様、都市や州名であり得る候補文字列を検索する。しかしながら、地域符号化部４３４は、米国における既知の都市全てのデータベース２２６２にある都市名を調査し、そこで見つからない場合、その都市名を拒否するという点が異なる。

【０１２３】

電話番号符号化部４３６：電話番号符号化部４３６は、電話番号対場所テーブル２２６６にあるエリアコード及び交換局を調査することで、電話番号を地理的位置に変換する。電話対場所テーブル２２６６は、エリアコードと交換局の対を都市名と州名の対にマッピングする。更に、この対は、地域名として扱われるが、例外は、その関連度スコアが、（発見的に決定される）小さい定数分だけ減じられ、このようにして得られた都市は、名指しされる都市に比べて幾分価値が低いという事実を反映する点である。単一の電話会社の中央局が、特に都市周辺位置の複数の都市をカバーしてもよくが、電話番号は、隣接都市に実際は配置される可能性がある。

【０１２４】

空間意味推定

空間符号化部４３は、空間意味推定４３８プロセス、すなわち、ＳＭＩ４３８

を含むが、これは、特別な種類の空間符号化を実行し得る。ＳＭＩ４３８は、意味論的解釈ではなく、空間・キーワード・ドキュメント索引５０５の然るべき部分の統計的な特性に基づき、用語（単語や句）に対する空間関連度を推定し得る。

【０１２５】

単語及び句には、地理的位置に対応するものがあるが、既存の居住者地域別分類サービスがその全てを記録するわけではない。これらの地理的関係を見いだすために、ＳＭＩ４３８は、候補の単語や句とＫＳＩとの相関関係を統計的に分析する。ＳＭＩ４３８では、ある句が、同じ場所の住所を有するドキュメントに頻発する場合、その句はまた、恐らくその場所に関係するという前提が用いられる。例えば、“ｔｈｅ　ｂｉｇ　ａｐｐｌｅ”は、単語“ｎｅｗｙｏｒｋ、ＮＹ”及びニューヨーク市の住所を有する多くのページに存在する。ＳＭＩ４３８も、“ｔｈｅ　ｂｉｇ　ａｐｐｌｅ”は、ニューヨーク市に関するものであると推定し得る。

【０１２６】

ＳＭＩ４３８は、以下の如く、空間関連度を推定する。空間―キーワードドキュメント索引５０５は、各々索引付けされた用語、すなわち、単語用語集２２５の各用語用のツリーを含む。任意の文字列の各単語について、ＳＭＩ４３８は、その単語に対応するツリーを調査する。この調査には、不均衡測定部４３９を呼出してツリー構造の不均衡の度合を測定する段階が含まれるが、このツリーは、空間ドキュメント索引５０３の修正版であることから、修正した結果、著しく不均衡な場合がある。不均衡測定部４３９については、後述する。一般的に言って、また更に詳細に説明するように、文字列の充分な用語が、同様に不均衡なツリーを有する場合、ＳＭＩ４３８は、その文字列を、前記ツリーの不均衡な部分によって説明される空間領域と対応付ける。

【０１２７】

前述の例に戻ると、句“ｔｈｅ　ｂｉｇ　ａｐｐｌｅ”中の単語は各々、多くのドキュメントに見られる。境界ボックスを指定することなく、空間・キーワード・ドキュメント索引５０５上でその句の検索を実行すると、ニューヨーク市付

近のドキュメント数に大きな“ピーク”が見つかる。このことは、修正後のツリーにおける不均衡の度合によって立証される。これら３つの単語の共通部分から得られるツリーは、ニューヨーク市をカバーする緯度・経度領域に、多くの枝を有する。このことによって、互いに隣り合うこれら３つの単語を有するページは、この緯度／経度領域を指している可能性があることがわかる。

【０１２８】

このような単語や句を“地理的事象”と呼ぶ。

ツリーアドレスは、以下の如く定義される。空間・キーワード・ドキュメント索引５０５の場合、索引ツリーのあらゆるノード又は葉は、そのノードに到達するために通過しなければならない子ノードの並びを示す一組の値によって識別し得る。例えば、２進ツリーでは、ツリーアドレス０１１０は、根ノードから出発して、第１子ノード、第２子ノード、第２子ノード、第１子ノードと進むことによって見つけられるノードを指定する。１６通りのツリーでは、“０ｘ４ｆ８”の如く１６進法で書かれたツリーアドレスは、根ノードから出発して、第５子ノード、第１６子ノード、第９子ノードと進むことによって見つけられるノードを指定する。

【０１２９】

空間・キーワード・ドキュメント索引５０５を用いることなく、特定の句の“ピーク性”を測定する場合、不均衡測定部４３９は、まず、平均的な単語の“標準ピーク性”を計算し、更に、候補をそれと比較する。１つの実施形態において、不均衡測定部４３９は、ランダムサンプリングした単語を取り出すことによって、標準ピーク性を計算し、またそれらの単語の各々に対して、その単語を含むドキュメントによって参照される点の２次元の分散を計算する。特に単語に関連するドキュメントは、その分散の算出において特別な重みを与え得る。例えば、その位置で複数のドキュメントを表すかのように、関連性の高いドキュメントを直線的に調整し得る。分散のこのランダムな集合の場合、不均衡測定部４３９は、平均分散を計算する。平均分散は、地理的に関連する句や単語を検出するための基準線として用い得る。基準線を大きく下回る分散を有する単語や句は全て地理的事象である。

【０１３０】

空間・キーワード・ドキュメント索引５０５を用いると、ＳＭＩ４３８は著しく簡略化される。空間・キーワード・ドキュメント索引５０５のツリーは、コンピュータシステム２０に既知のドキュメント全てに及ぶことから、ＳＭＩ４３８は、修正後のツリーにおける葉のツリーアドレス集合を考慮すると、単に地理的事象を検出すればよい。例えば、候補の単語や句の場合、ＳＭＩ４３８は、空間・キーワード・ドキュメント索引５０５に、この単語や句に対する修正後のツリーを獲得するように問合せて、このアドレスリスト上の以下の動作を実行する。

【０１３１】

そのツリーから、ＳＭＩ４３８は、各葉のツリーアドレスのリストを作成する。全アドレスの第１桁から出発して、ＳＭＩ４３８は、このレベルの（すなわち、この桁に対する）最も一般的な枝番号を見つける。枝はツリーの分岐点であり、候補位置の方向を示すことから、この桁により索引付けされた枝は、“候補分岐点”と呼ばれる。ＳＭＩ４３８は、そのレベルにおける候補分岐点に続くアドレスの小数部を計算する。

【０１３２】

次のレベルにおいて、ＳＭＩ４３８は、最終レベルの候補分岐点を取った全てのアドレスを考慮し、再度最も一般的な分岐点方向を見つけ、それを次の分岐点方向として用いる。ＳＭＩ４３８は、その候補分岐点に更に続くアドレスの小数部を再び計算する。

【０１３３】

ＳＭＩ４３８は、この候補分岐点に更に続くアドレスの割合が、コンピュータシステム２０の操作者が調整可能な所定の閾値を下回るまで、これを繰返す。個々の閾値は、各ドメインに対して調整し得る。閾値を調整すると、考慮対象の照合品質が調整される。閾値は経験的に設定される。

【０１３４】

例えば、簡単の簡略化のために、２進ツリーであって、そのノードがドメイン空間を矩形に分割する２進ツリーについて、また幾つかのレベルに対して共に分岐する以下の４つのアドレスについて考える。

【０１３５】

１０１１１０１０１０１１１１１

１０１１１０１０１１１０１０１０

１０１１１０１０１１１０１１１１

１０１１１０１０１１１０１１０１

レベル１：分岐１＝１００％

レベル２：分岐０＝１００％

レベル３：分岐１＝１００％

レベル４：分岐１＝１００％

レベル５：分岐１＝１００％

レベル６：分岐０＝７５％

レベル７：分岐１＝７５％

レベル８：分岐０＝７５％

レベル９：分岐１＝７５％

レベル１０：分岐１＝７５％

レベル１１：分岐１＝７５％

レベル１２：分岐０＝７５％

レベル１３：分岐１＝７５％

レベル１４：分岐１＝５０％

レベル１５：分岐０＝２５％（５０％閾値を下回る）。

【０１３６】

これらのツリーアドレスは、単語が、アドレス１０１１１による空間索引ツリー５０２の矩形によって定義される領域に１００％関連し、また矩形１０１１１０１０１１１に７５％関連することを示す。

【０１３７】

特定の単語がほとんど無い、すなわち、ページリポジトリ２２２全体で数回しか現れないが、その出現度数が、同じ場所の地理的識別子と高度に相関している場合、その単語は、点位置に関連付け得る。例えば、単語“ＥＶＯＯ”は、米国、ＭＡ州、ｓｏｍｅｒｖｉｌｌｅにあるレストランの名前である。単語“ＥＶＯ

Ｏ”は、コーパス全体には数回しか現れない。このほとんどの場合、このレストランのアドレスを有するページに現れる。その他の場合、そのレストランを批評するページに現れる。“ＥＶＯＯ”がレストランのアドレスと強い相関がある場合、単語“ＥＶＯＯ”を同じ緯度／経度点で居住者地域別分類し得る。これによって、他のページをその同じ点で居住者地域別分類し得る。緯度／経度点は、単語リンク“ＥＶＯＯ”を介してページ間で送信される。

【０１３８】

通常、空間意味推定４３８プロセスは、点として注目される位置と句を対応付けできないことに留意されたい。より一般的な結果は、境界ポリゴンである。これらの句を居住者地域別分類する主な目的は、ドキュメントの順位付けを改善することであり、空間ドキュメント順位付け４５の項で議論する。

【０１３９】

キーワード構文解析部

非地理的な検索用語（キーワード）は、以下の如く識別される。ドキュメントがページリポジトリ２２２に保存される際、キーワード構文解析部４４プロセスは、各ドキュメントを開き、そのキーワードをスキャンする。これらのキーワードは、単語例２２７と呼ばれるデータベーステーブルに記憶されるが、これには、単語ＩＤ２２７２、ドキュメントＩＤ２２７４、及び単語・ドキュメント関連度フロート２２７６等のフィールドを含む。単語例２２７テーブルは、任意のキーワードを、それを含む１組のドキュメントと対応させる。

【０１４０】

単語ＩＤは、その単語の文字列を置き換える数字である。これによって、記憶装置の要求事項が低減され、また“ｔｈｅ　ｂｉｇ　ａｐｐｌｅ”等の句を単一データベース項目として扱い得る。単語用語集２２５は、全単語及びそれらの対応する単語ＩＤの辞書として機能するデータベーステーブルである。単語用語集２２５テーブルは、単語２２６２１、単語ｉｄ２２６２３、及び単語出現度数２２６２５のフィールドを含む。

【０１４１】

キーワード構文解析部４４は、ＳＧＭＬ又は関連する規格であるＨＴＭＬ及び

ＸＭＬ等、タグ付きテキストを含むドキュメントを構文解析するためのタグ認識部４４２を含む。様々なドキュメント規格用のタグ認識部が、コンピュータ処理技術では良く知られており、コンピュータ処理システムの１つの特徴であり得る。

【０１４２】

この技術分野における標準的な方法を用いて、句検索のためにドキュメントに索引付けしてもよく、これによって、ユーザは、ドキュメント中で近くに又は直ぐ近くにある単語の集合に対する問合せを発行し得る。

【０１４３】

空間ドキュメント順位付け

潜在的に極めて大量の情報の場合、ドキュメント順位付けは、非常に重要である。ユーザの問合せに関連する結果は、関連の無い結果で圧倒されてはならず、さもないとシステムが役に立たなくなる。

【０１４４】

空間ドキュメント順位付け４５プロセスは、ドキュメント対場所関連度４５２、ドキュメント対単語関連度４５４、及び抽出品質４５６の評価を含むドキュメントの順位を生成する。評価は、各ドキュメントと問合せとの関連度を示す浮動小数点数に組み合わされる。

【０１４５】

ドキュメント対場所関連度４５２スコアは、ドメイン位置へのドキュメントの関連度を示すが、ここで、ドメイン位置は、ドキュメント内のＰＳＩ又はＫＳＩによって示される。以下は、１つのＳＩ（ＰＳＩ又はＫＳＩであり得る空間識別子）の１つのドキュメントに対する関連度を考慮する方法である。同じドキュメントにおいて、幾つかの異なるＳＩに対して、これを算出することが可能である。これらのＳＩは、全て同じ地理的領域を参照する場合、組み合せ得る。例えば、ドキュメントは、居住者地域別分類できる住所や電話番号を有し得る。住所が、電話番号のエリアの内側に組み込まれた点に対するものである場合、その住所へのドキュメントの地理的関連度は改善し得る。関連度の増加は、複数のＳＩがページ上で集約し得る様々な環境に対して選択された手製の重みに影響される。

この改善は、以下の方法で計算される関連度にとって二次的なものである。

【０１４６】

ドキュメント対場所関連度

ドキュメント対場所関連度４５２スコアは、ページ中の位置４５２１、末尾からの距離４５２３、他のＳＩの数４５２５、文中４５２７、及び強調４５２９のスコアを含む（付録Ａを参照）。ページ中の位置４５２１スコアは、発見機能であり、ＳＩを数多く観測することで較正される。この機能は、もっと関連度がありそうなＳＩがドキュメントに以前現れたという前提で、スコアを割当てる。距離は、文字又はバイトで測り得る。（スクロールの必要が無くページが最初に読込まれた時のスクリーン上で）“重なり部上方”に現れるＳＩが、最も関連性があると見なされる。

【０１４７】

ＳＩがドキュメントのフッタに現れる場合、末尾からの距離４５２３スコアは、ドキュメント対場所関連度４５２スコアをわずかに増やすが、これは、位置発見的にそれに割当てる低いスコアを部分的に打ち消す。

【０１４８】

他のＳＩの数４５２５スコアは、他のＳＩが幾つ同じドキュメントにあるかに基づき、ＳＩの関連度を薄める発見機能である。住所が多数あるドキュメントは、リストでありがちであり、ここでは、個々の住所は全てドキュメントと関連する確率が低い。

【０１４９】

文中４５２７スコアは、文中で記述されるのとは対照的に、独立しているＳＩをわずかに増やす。

強調４５２９スコアは、太字の大活字であることやページの表題中にあることを含み、ＳＩテキストの強調の度合を反映する。このスコアは、１．０が標準（強調解除も強調も無い）状態であると仮定された１０進数の形態を取り、これより小さい数値は、強調の欠如（テキストが小さい等）を示し、またこれより大きい数値は、目立つ（ｐｒｏｍｉｎｅｎｃｅ）ことを示す。

【０１５０】

ドキュメント対単語関連度

ドキュメント対単語関連度４５４スコアは、個々の単語の、それを含む個々のドキュメントに対する関連度を示す。ドキュメントに対する単語の関連度を測定する手段は、この技術では良く知られている。例えば、Ｓ．Ｅ．ロバートソン（Ｒｏｂｅｒｔｓｏｎ）及びＫ．スパーク・ジョーンズ（Ｓｐａｒｃｋ・Ｊｏｎｅｓ）による“テキスト検索の簡単且つ確実な手法”　（ケンブリッジ大学、コンピュータ研究技術報告書、１９９７年５月）を参照されたい。

【０１５１】

また、句の検索は、ドキュメント関連度にも影響を及ぼし得る。通常、この種の関連度は、個々の句に対するユーザの問合せ時点において、その場で計算される。この技術分野においては、この種の関連度を計算するための標準的な方法がある。

【０１５２】

抽出品質

抽出品質４５６スコアは、任意の単語又は場所とは独立なドキュメント値を表す。これを測定する方法には幾つかあるが、ドキュメントにリンクするページ数、検索結果として提供される時ドキュメントをクリックする回数、及び同じ単語と場所を参照する他のドキュメントの数（すなわち、他の多くのドキュメントと同様なドキュメントである場合、その抽出値は、それが含む個々の単語とは無関係に低いと考えられる）が含まれる。

【０１５３】

抽出品質４５６スコアは、ネットワーク接続性４５６２及び手入力更新４５６４用の構成要素を含む。ネットワーク接続性４５６２は、ページがウェブのランダム巡回によって選択される確率から計算される。更に、この確率は、スコアにマッピングされる。任意のいずれかのドキュメントが見つかる確率は、集合の大きさに反比例するため、選択された個々のマッピングは、ページリポジトリ２２２のドキュメント集合の大きさに依存する。

【０１５４】

手入力更新４５６４スコアは、編集者の入力を取り入れるようになっている。

編集者は、個々のドキュメントの抽出品質４５６を調整する規則を作成し得る。例えば、編集者は、ドキュメント品質の尺度を大きくするだけで、他のドキュメントよりも良いものとして、特定のサイト内の全ドキュメントを重み付けできる。編集者は、Ｚａｇａｔ．ｃｏｍ等、それ自体、編集者が念入りに作成したものであるサイトを用いてこれを行い得る。

【０１５５】

抽出品質４５６スコアは、抽出ドキュメント品質２２８テーブルに記憶されるが、これには、ドキュメントｉｄ２２８１及びドキュメント品質２２８３のフィールドが含まれる。ドキュメントｉｄ２２８１フィールドは、ページリポジトリ２２２のドキュメントｉｄ２２２１を参照する外部鍵である。

【０１５６】

索引付け部

索引付け部４６は、ドキュメントを分析して、検索５０プロセスを加速するデータ構造を準備する。索引付け部４６は、空間索引付け部４６２、空間・キーワード索引付け部４６５、及びツリー次数変換部４６６を含む。

【０１５７】

空間索引付け部

図７において、空間索引付け部４６２は、ドメイン空間に対する空間索引５０２及び空間ドキュメント索引５０３を生成する。空間索引５０２は、２進ツリーである。空間ドキュメント索引５０３は、空間索引５０２に基づくツリーであるが、次数は２（全２進ツリーの次数）より大きくてもよい。

【０１５８】

ステップ４６２１において、空間索引付け部４６２は、ページリポジトリ２２２のドキュメントによって参照される全ドメイン位置の集合を集め、更に、ステップ４６２２の空間索引５０２ツリー用の根ノードを生成する。空間索引付け部４６２は、根ノード及びその集合をステップ４６２４に渡すが、ここで、再帰的空間索引付けサブルーチン（すなわち、ＲＳＩＳ）４６２０の開始位置にマーキングする。

【０１５９】

ステップ４６２４において、ＲＳＩＳ４６２０は、ノードと集合を受信する。ステップ４６２５において、ＲＳＩＳ４６２０は、その集合を調べて、その集合が２つ以上の要素を含むか否か判断する。含まない場合、ステップ４６２９５において、ＲＳＩＳ４６２０は、現在のノードを１要素のドメイン位置と対応付け、ステップ４６２９に進み、それを呼出したルーチンに制御を返す。含む場合、ＲＳＩＳ４６２０は、ステップ４６２６に進み、ここで、ＲＳＩＳ４６２０は、その集合を空間分割部Ｄに沿って集合ＬとＲに空間的に分割するが、この分割は、できるだけＬとＲの数が等しくなるように行なわれる。ドメイン空間が、平面である場合、空間分割部Ｄは、平面内の直線である。ドメイン空間が、３次元である場合、空間分割部Ｄは、３次元空間を通る平面である。一般的に、ドメイン空間が、Ｘ次元である場合、空間分割体は、次元数Ｘマイナス１の多様体である。またステップ４６２６において、ＲＳＩＳ４６２０は、ノードＮにおける空間分割部Ｄ用の基準も記憶する。従って、各ノードは、位置の主集合を２つの下位集合に分割する基準を含む。

【０１６０】

またステップ４６２６において、ＲＳＩＳ４６２０も、ステップ４６２４に渡されたノード上に、左ノードと右ノードを生成する。これによって、索引として機能する２進ツリーに分岐点が生成される。ツリー全体が空間索引５０２になる。

【０１６１】

ＲＳＩＳ４６２０は、各下位集合上で、それ自体を呼出すことで再帰的になる。特に、ステップ４６２７において、ＲＳＩＳ４６２０は、下位集合Ｌ及び現在の左ノードをステップ４６２４に渡し、一方、ステップ４６２８において、ＲＳＩＳ４６２０は、下位集合Ｒ及び現在の右ノードをステップ４６２４に渡す。ＲＳＩＳ４６２０は、各集合が、子無しノードに対応する単一要素の集合に分割されるまで繰返される。他の全てのノードは、分割基準と、それらから降順の２つのノードを有する。

【０１６２】

空間索引付け部４６２が、空間索引５０２ツリーを構築し、これが、ドキュメ

ントのコーパスにおいて参照される点に索引付けした後、空間索引付け部４６２は、空間索引５０２ツリーのコピーを拡張して、同じ空間点を参照する複数のドキュメントをカバーすることによって、空間ドキュメント索引５０３を構築する。空間索引付け部４６２は、ツリー次数変換部４６６を呼出して、次数ｋのツリーに表れる空間索引５０２のバージョンを作成する。

【０１６３】

空間索引５０２を拡張すると新しい枝が生成され、これらの枝は、もはや空間分割を反映しないが、代わりに、その点を参照するドキュメントの区分を反映する。特に、（空間索引５０２から引き継がれたノードが継続して行うように）ドメイン内での空間分割を定義する基準を含むノードの代わりに、拡張後に追加されたノードは、ドキュメントのドキュメントＩＤ２２２１数の空間内において分岐用の基準を含む。データベーステーブルの（ドキュメントＩＤ２２２１等の）鍵の値に基づく区分化は、この技術では標準的である。このように区分すると、鍵としてドキュメントＩＤ２２２１数を用いてドキュメント上にｋ通りのツリーが作成される。

【０１６４】

次数Ｋ

次数ｋの索引ツリーの重要な最適化には、ｋの選択が含まれる。ｋ通りに分岐する構造は、ツリーを構築又は記憶する前に選択しなければならない。ｋは、ドキュメント数と、可能性として基となるコンピュータ処理プラットフォームに依存して、小さい場合２、大きい場合数千もしくは数万であり得る。次数ｋのツリーは、Ｌレベルでｋ Ｌのドキュメントを索引付けできる。

【０１６５】

ｋの値が大きいと、幾つかのドキュメントのみに現れるキーワードの処理が、高速にそして記憶効率が更に良くなる。ページリポジトリ２２２の稀な単語の数が多い場合、ｋの値が大きいと、小さい場合と比べて、更に記憶効率が高まる。しかしながら、ｋの値が小さいと、更に検索効率が高まり得るが、これは、（問合せに応じて）横断プロセスは、その制約基準を満たさないツリーの枝を無視し得るためである。

【０１６６】

ｋの選択は、ページリポジトリ２２２において、ドキュメントの各集合が索引付けされるように実行し得る経験的なプロセスである。これは、単一プロセッサ命令で処理されるバイト数及びディスクドライブが読込むブロック数等、ハードウェアの制限の影響される。ｋの選択で最も重要な要因は、単語・頻度分布である。例えば、ウェブページ用のキーワード用語集は、１つ又は２つのドキュメントのみに現れる膨大な数の単語を示すが、より一般的な単語は多くのドキュメントに現れる。これらの一般的な単語は、“裾野が厚い”分布を生成する。特定集合ドキュメントの分布の正確な形状が、最適なｋを決定する。値がｋである場合、特定の用語集及びドキュメント集合用の単語ツリーの記憶に用いるバイト数をカウントする計算が簡単である。

【０１６７】

ツリー次数変換部

ツリー次数変換部４６６は、２進ツリー及び整数ｋを含むパラメータを受取る機能であり、２進ツリーの構造とデータを組み込んだ次数ｋのツリーをその出力として返す。この変換方法は、コンピュータ処理技術では公知である。

【０１６８】

空間・キーワード索引付け部

空間・キーワード索引付け部４６５は、ドキュメントに対する問合せに応答する空間・キーワード・ドキュメント索引５０５を構築する。問合せは、キーワード基準、空間基準、あるいはその両方を有し得る。

【０１６９】

空間・キーワード索引付け部４６５は、ページリポジトリ２２２のドキュメントが参照するドメイン位置を全て集める。

空間・キーワード索引付け部４６５は、空間索引付け部４６２が生成した空間ドキュメント索引５０３を用いる。空間ドキュメント索引５０３は、このドキュメントリスト上のｋ通りのツリーである。空間・キーワード索引付け部４６５は、空間ドキュメント索引５０３をコピーして、各キーワード用のキーワードツリー５０６を作成する。各キーワードツリー５０６では、空間・キーワード索引付

け部４６５は、その特定のキーワードを含まないドキュメントを全て修正して取り除く。ドキュメントの修正後、キーワードツリー５０６のノードから決まる下位ツリーが、ドキュメントを含まない場合、空間・キーワード索引付け部４６５は、そのノードを（従って、その下位ツリーを）除去する。

【０１７０】

空間・キーワード索引付け部４６５は、ページリポジトリ２２２のドキュメントのコーパスにキーワードを関係付ける最小キーワードツリー５０６を各キーワードに対して作成する。更に、空間・キーワード索引付け部４６５は、１つの分岐構造が、空間ドキュメント索引５０３ツリーと同様に全てのキーワードツリーに共通であるように保証する。

【０１７１】

検索

図５において、検索５０プロセスは、関連度毎に順位付けされた１組のドキュメントで問合せに応答する。

語彙ツリー５０８は、空間ドキュメント索引５０３ツリーの何らかのコピーであるが、修正されている場合もある。従って、各キーワードツリー５０６は、空間ドキュメント索引５０３ツリー自体にある語彙ツリー５０８である。また、フィルタは、１組のドキュメントを決定し、いずれかの組のドキュメントが、空間ドキュメント索引５０３ツリーの修正を決定することから、フィルタは、いずれも語彙ツリー５０８として表し得る。従って、語彙ツリー５０８を構築して、ドキュメントの複合集合を任意に索引付けし得る。

【０１７２】

検索５０プロセスは、空間ドキュメント索引５０３及び空間・キーワード・ドキュメント索引５０５を用いて、任意の組のドメイン位置又は領域を参照するドキュメント、及び単語用語集２２５に存在する任意の組のキーワードに対応するドキュメントを見つける。また検索５０プロセスは、語彙ツリー５０８を用いて、フィルタを表し得るようなドキュメントを見つけることができる。従って、検索５０プロセスは、空間ドメイン基準、キーワード基準、フィルタ、又はこれらのいずれかを組合せに従って、ドキュメントを探す問合せに応じ得る。更に、検

索５０プロセスは、ドキュメント順位付け部５６プロセスを呼出して、問合せ用語に対して関連度毎にドキュメントの結果集合を順位付けし得る。

【０１７３】

検索５０プロセスは、図６の手続を介して、問合せに応える。問合せには、閉じた形状を指定する境界領域（通常、２次元のポリゴン）、単語、句、及び層の内少なくとも１つが含まれる。境界領域は、地図インタフェース８０からのドメインフレームであり得る。

【０１７４】

問合せの各要素に対して、検索５０プロセスは、以下の如く決定される適当なツリーのコピーを読込む。境界領域が指定される場合、ステップ７０３において、空間ドキュメント索引５０３が読込まれる。キーワードが指定される場合、ステップ７０２において、各キーワードに対して、空間・キーワード・ドキュメント索引５０５ツリーが読込まれる。句が指定され、またその句が、単語用語集２２５の単一の項目ではない場合、ステップ７０２において、各単語の空間・キーワード・ドキュメント索引５０５が読込まれる。句が指定され、また単語用語集２２５の単一の項目である場合、ステップ７０２では、その句の空間・キーワード・ドキュメント索引５０５のみを読込む必要がある。層が指定される場合、その名前で、ステップ７０２で読込まれる適切な語彙ツリー５０８が識別される。

【０１７５】

検索５０プロセスは、これら各ツリーの葉の数をカウントする。ステップ７０３において、検索５０プロセスは、境界領域の面積にコーパスの点の平均密度を乗じて、問合せ境界領域で囲まれる空間ドキュメント索引５０３の葉の数を推定する。ステップ７０４において、これらの数を用いて、最小ツリーを最初にして、ツリーをリストに並べる。

【０１７６】

ステップ７０５において、この最小ツリーは、結果ツリーとしてラベル換えされ、また修正され、最終的な結果ツリーを生成する。部分的に修正された結果ツリー中に存在する各ノードに対して、検索５０プロセスは、そのノードが含まれるか否か全てのツリーをチェックする。ステップ７０８及び７１２において、検

索５０プロセスは、リスト順にツリーをチェックする。そのノードが欠落したツリーがある場合、検索５０プロセスは、チェックを中止し、ステップ７０９において、結果ツリーのノードの下にある下位ツリーを削除する（付録Ｂを参照）。ステップ７１０及び７１１において、ツリーが詳細にチェックされる。検索５０プロセスは、葉ノードだけ残るまで、結果ツリーの全ノードをチェックし続ける。これらの葉ノードは、ドキュメントの結果集合を表す。ステップ７１３では、結果ツリーが返される。

【０１７７】

空間・キーワード・ドキュメント索引５０３ツリーの葉は、単語関連度、及び各ドキュメントにおける単語の位置と文脈上の強調のリストを有する。空間ドキュメント索引５０３は、各ドキュメント用の空間関連度を有する。各層の語彙ツリー５０８は、いくつかのドキュメント用の抽出ドキュメント品質４５６を持つことがある。これらの関連度は、結果集合の各ドキュメントに対して組み合せられる。組合せ手続は、平均化、合計、又は重み付け平均であり得る。

【０１７８】

第２プロセスは、ドキュメント内における複数の問合せ単語の強調と近似を考慮することによって、ドキュメント関連度に対する調整量を算出し得る。この標準的な手続は、問合せ単語が密接して見えるドキュメントに高い関連度を与えるだけである。

ドキュメントの最終結果リストは、ソートしてユーザに返し得る。ソート手続は、高い関連度を有するドキュメントの一部だけを抽出してよい。

【０１７９】

ドキュメント順位付け部

ドキュメント順位付け部５６は、結果集合の各ドキュメント用の様々な関連度スコアを組み合せて、この組み合せた関連度によって、ドキュメントをソートする。組合せ関数は、平均化又は重み付け加算又は用いた様々な関連度スコアに合せて調整した他の組合せ関数であってもよい。ドキュメント順位付け部５６は、幾つかのデータベースシステムからソート済結果集合のストリームを取り出してもよく、また、それらをマージソートして、新しい結果集合を生成し得る。

【０１８０】

アイコン順位付け部

アイコン順位付け部５７は、ドキュメント順位付け部５６から結果のソート済リストを受信する。これらのドキュメントを要求したユーザにこのリストを提示する場合、アイコン順位付け部５７は、アイコンの項で説明した方法に従って、重なり合うアイコンを統合する。この統合アイコンのリストは、各アイコン８１０の次にあるサブリストを有するユーザに提供される。これらのサブリストは、アイコン８１０に統合されたドキュメントを識別する。

【０１８１】

アイコン順位付け部５７は、以下の如く、アイコン８１０にドキュメントをグループ化する。アイコン順位付け部５７は、ソート済結果リストから第１ドキュメントを取り出し、それをアイコンリストの第１アイコン８１０にする。その結果リストに仮表示位置を有する後続の各ドキュメントに対して、アイコン順位付け部５７は、その仮表示位置に位置するアイコン８１０が、アイコンリストに既にあるいずれかのアイコン８１０と衝突するか否か判断する。衝突が起こる場合、アイコン順位付け部５７は、衝突するドキュメントを既存のアイコンと対応させる。衝突が起きない場合、アイコン順位付け部５７は、アイコン８１０をアイコンリストに追加し、現在のドキュメントを前記アイコン８１０と対応させる。この手続は、アイコンの数が、ユーザが決定する最大数又はコンピュータシステム２０のカスタマイズ可能な動作パラメータである所定の数字のいずれか小さい方に到達した場合いつでも終了し得る。

【０１８２】

ドキュメントが、特定のアイコンクラスと項目別に関連している場合、アイコン順位付け部５７は、ドキュメントを表すアイコン８１０に前記アイコンクラスからのアイコン面８１８を割当てる。複数のアイコンクラスが、単一アイコン８１０で表されるドキュメントと関連している場合、アイコン順位付け部５７は、前記アイコンクラスの１つを選択して、前記アイコン８１０に割当ててもよく、又は前記複数のアイコンクラスを反映するよう構築された新しいアイコンクラスを割当ててもよい。

【０１８３】

ユーザプロフィール

ユーザプロフィール６５プロセスは、ユーザアカウントに特有の情報を管理する。この情報は、ユーザが、過去にコンピュータシステム２０と如何に対話したかという内容を含み得る。記録し得る他の要素には、対話開始時にユーザに表示する既定位置、以前収集された層の集合、以前記入されたメモドキュメントの集合、以前の検索、及び以前のクリックパターン又は挙動が含まれる。この情報の一部又は全てを、ユーザが直接閲覧し編集可能なようにできる。

【０１８４】

またユーザプロフィール６５プロセスによって、ユーザは、ユーザ名及び場合によりパスワードでコンピュータシステム２０にログインできる。ユーザ名は、この技術では一般的なように、ユーザをユーザアカウントで識別する。地図インタフェース８０は、アカウントログイン項目制御８６１を含んでもよく、これは、アカウントログイン用プロンプト８６２、データ項目制御８６３、及び書込制御８６４を含む。

【０１８５】

データ提示

データ提示６０プロセスは、各ユーザセッション用の地図インタフェース８０の状態を管理する。ユーザが、例えば、問合せの発行、制御の選択、及びインタフェースツールの一般的な利用によって、地図インタフェース８０の状態を変更する場合、データ提示６０システムは、これらの変更やその順序の記録をとる。この記録履歴を用いると、以前の結果集合内での問合せが可能になる。例えば、ユーザは、“ｓｈｏｅｓ”を参照するドキュメントに対して“ｃａｍｂｒｉｄｇｅ、ｍａ”での問合せができ、また後続の対話において、ユーザは、単語“ｓｔｏｒｅ”を含むドキュメントだけを更に要求することによって、この組のドキュメントをフィルタ処理し得る。この結果、“ｓｈｏｅｓ”及び“ｓｔｏｒｅ”を含み、“ｃａｍｂｒｉｄｇｅ、ｍａ”を参照するドキュメントのリストが得られる。更に、ユーザは、縮小を行って、地図にまだ表示されるこれらのドキュメントで領域を拡大して閲覧する。この拡大したドメインにおいてキーワード問合せ

に合致し得る新しいドキュメントを閲覧する場合、ユーザは、問合せを再発行し得る。

【０１８６】

同様に、ユーザは、１組のドキュメントを異なる問合せにより選択された他の組のドキュメントと組み合せてもよい。

後続のフィルタ動作又は結果集合の組合せは何回でも実行し得るが、性能上の理由から、コンピュータシステム２０の記憶資源、又はオプションとして、コンピュータシステム２０に構築されたパラメータによってのみ制限される。コンピュータシステム２０が、後続の各対話において、正しい組のドキュメントをユーザに提示し得るように、データ提示６０システムは、任意のユーザによるフィルタ動作を追跡する。

【０１８７】

サービス収集部

サービス収集部２４には、ユーザインタフェースサーバ６２及びポータルサーバ７０が、データ提示６０、検索５０、及びユーザプロフィール６５用のプロセスと通信するプロクシが含まれる。

【０１８８】

ポータルサーバ

コンピュータシステム２０は、ポータルサーバ７０プロセスを含む。ポータルサーバ７０は、遠隔手続呼出し及び他のネットワークプロトコルを介して、コンピュータシステム２０の少なくとも幾つかのサービスを提供する。これによって、公開ポータルシステムを介して又は直接個人にコンピュータシステム２０のサービス、データ、及びツールを配信し得る。公開ポータルシステムを提供する企業の例には、ＣＡ州、サンタクララ、Ｃｅｎｔｒａｌ・Ｅｘｐｒｅｓｓｗａｙ３４２０のＹａｈｏｏ！社、及びＫＹ州、ロンドン、私書箱８０７７のＳｐｒｉｎｔＰＣＳが含まれる。

【０１８９】

他の選択可能な実施形態

本発明の多くの実施形態について説明した。しかしながら、様々な変更を本発

明の精神と範囲から逸脱することなく行い得ることが理解されるであろう。従って、他の実施形態も以下の請求項の範囲内にある。

【０１９０】

付録Ｂ

ツリーＴのノードアドレスＮの存在を以下のように調べる。

```
ｉｆ（Ｔがキーワード／階層ツリーである）｛
ノードＱ＝Ｔの根ノードへのポインタ；
ノードアドレスＮにおいて、ｆｏｒｅａｃｈ＄ｓｔｅｐ｛
  ｎｅｘｔノードＱ＝前ノードＱから子番号＄ｓｔｅｐへのポインタ；
  ｉｆ（ノードＱが有効な子である）｛
    ｆｏｒｅａｃｈループを継続；
  ｝ｅｌｓｅ｛
    戻り値＝“ｆａｌｓｅ”で終了；
  ｝
｝
ｒｅｔｕｒｎ　“ｔｒｕｅ”；＃非存在子をヒットせずにループ終了
｝
ｉｆ（Ｔが空間ツリーである）｛
以下、ポリゴンＰは、ユーザが与える境界領域；
ノードＱ＝Ｔの根ノードへのポインタ；
  ノードアドレスＮにおいて、ｆｏｒｅａｃｈ＄ｓｔｅｐ｛
  ｎｅｘｔノードＱ＝前ノードＱから子番号＄ｓｔｅｐへのポインタ；
  ｉｆ　（ノードＱ下方の領域が、ポリゴンＰと重なり合う）｛
  ｝ｅｌｓｅ｛
    戻り値＝“ｆａｌｓｅ”で終了；
  ｝
｝
ｒｅｔｕｒｎ　“ｔｒｕｅ”；＃問合せの境界領域外の分割部をヒットせずに
                    ループ終了
```

}

【０１９１】

付録Ａ

場所にドキュメントの関連度を割当てるための擬似コード抜粋。

＃発見的に決定されたパラメータ：

＃＄ｅｍｐｈａｓｉｓ＿ｂｏｎｕｓ＿ｍｏｄｉｆｉｅｒは、強調ビットの重要性を決定する。

＃＄ｓｅｎｔｅｎｃｅ＿ｐｅｎａｌｔｙ＿ｍｏｄｉｆｉｅｒは、文中ビットの重要性を決定する

＃＄ｓｐ＿ｆｕｌｌ＿ｐｏｉｎｔ：この後文ペナルティが完全に該当する位置

＃＄ｓｐ＿ｔｒａｎｓｉｔｉｏｎ＿ｐｏｉｎｔ：この後、文ペナルティが該当し始める位置；この位置でのゼロから＄ｓｐ＿ｆｕｌｌ＿ｐｏｉｎｔでの＄ｓｅｎｔｅｎｃｅ＿ｐｅｎａｌｔｙ＿ｍｏｄｉｆｉｅｒになる

＃＄ｅｎｄ＿ｂｏｎｕｓ＿ｓｉｚｅ：ドキュメント末尾ボーナスが該当するドキュメントの末尾からの最大文字数

＃＄ｅｎｄ＿ｂｏｎｕｓ＿ｍａｘ：ドキュメント末尾ボーナスが該当する最大関連度値

＃＄ｅｎｄ＿ｂｏｎｕｓ＿ｍｕｌｔｉｐｌｉｅｒは、ドキュメント末尾ボーナスの重みを決定する

＃位置発見関数で開始する。これは、位置ゼロに対して１に規格化される非増加関数である。“重なり部”の平均位置となるある位置ｐ＿ｆ、すなわち、ユーザに最初に表示された時、通常のドキュメントの可視領域終端が現れる場所まで徐々に減少する。位置がｐ＿ｆよりも大きい場合、更に速く減少するが、位置が広範な場合、横ばい状態になる。正確な形状は、通常のドキュメントの数多いインスタンスのＰＳＩにスコアを手入力で割当てることで、また、これらのスコアに関数を当てはめることで、発見的に決定される。

＄ｒｅｌｅｖａｎｃｅ＝＆ｐｏｓｉｔｉｏｎ＿ｆｕｎｃｔｉｏｎ（ｐｏｓ）；

＃太字体である、大型字体である、表題中に含まれる等に対するボーナス。＄ｅｍｐｈａｓｉｓは、ＰＳＩがどの程度強調されるかに基づき割当てたＰＳＩの

発見関数である。

【０１９２】

$emphasis_bonus=emphasis_bonus_modifier*emphasis;

#文中にあることに対するペナルティ。例：“ＭＡ０１７４８、ホプキントン、メイン・ストリート５２のホプキントン・ドラッグストアを通して、当社製品が幾つか入手可能なことをお知らせ致します”。最初の$sp_transition_point文字のＰＳＩには、ペナルティは割当てられず、$sp_full_point文字以降完全なペナルティになる。

【０１９３】

```
if ($pos>$sp_full_point) {
$sentence_penalty=sentence_penalty_modifier*in_sentence;
} else {
if ($pos>sentence_penalty_transition_point) {
  $sentence_penalty=$in_sentence*$sentence_penalty_modifier*(($pos-$sp_transition_point)/$sp_full_point-$sp_transition_point);
} else {
  $sentence_penalty=0.0;
}
}
```

【０１９４】

$relevance+=$emphasis_bonus-$sentence_penalty;

#長いドキュメントの場合ドキュメントの末尾にあることに対するボーナス。関連度がどの程度既に低いかに比例し、このため、既に高いスコアのＰＳＩは、

末尾にあることに対するボーナスを受取らない。このことは、ＰＳＩ数の関数の前であるため、この関数によって、これは小さくなる（また大きなリストにある最後のＰＳＩのスコアは、高くなり過ぎることはない）。

【０１９５】

```
if ($size-$pos<$end_bonus_size&&$relevance<$end_bonus_max) {
$relevance+=($end_bonus_max-$relevance)*$end_bonus_multiplier;
}
```

【０１９６】

```
#ここで、他のＰＳＩが幾つページ上に現れるかに基づき、上記スコアを小さくする。num_psi_function（$num）は、ＰＳＩが他のＰＳＩと共に起こる際、ＰＳＩがどの程度価値が無いか判断する関数である。この関数は非増加であり、また$num＝１に対して１であり、$numが小さい時は、速く減少し、$numが大きなると、ゆっくり減少する。この関数は、位置関数に対して、上述の如く発見的に決定される。
$relevance*=&num_psi_function($num);
```

【図面の簡単な説明】

【図１】　本発明の実施形態に基づくコンピュータシステムの全体構成を示す概略図。

【図２】　本発明の実施形態に基づく地図インタフェースに対する制御配置を示す概略図。

【図３】　記憶エンティティとデータ収集プロセスにおけるエンティティとの説明図。

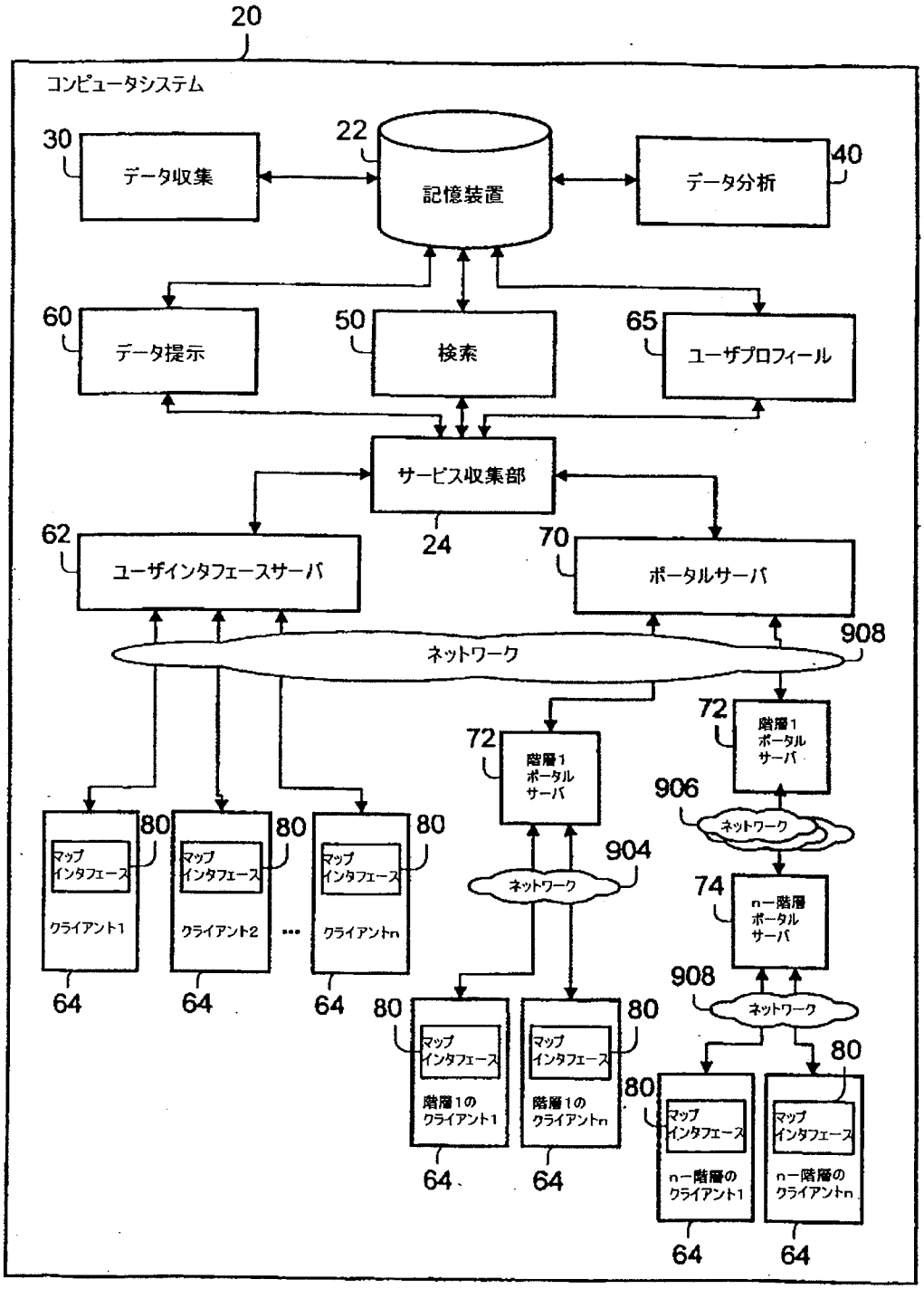
【図４】　データ分析プロセスにおけるエンティティの説明図。

【図５】　検索プロセスにおけるエンティティの説明図。

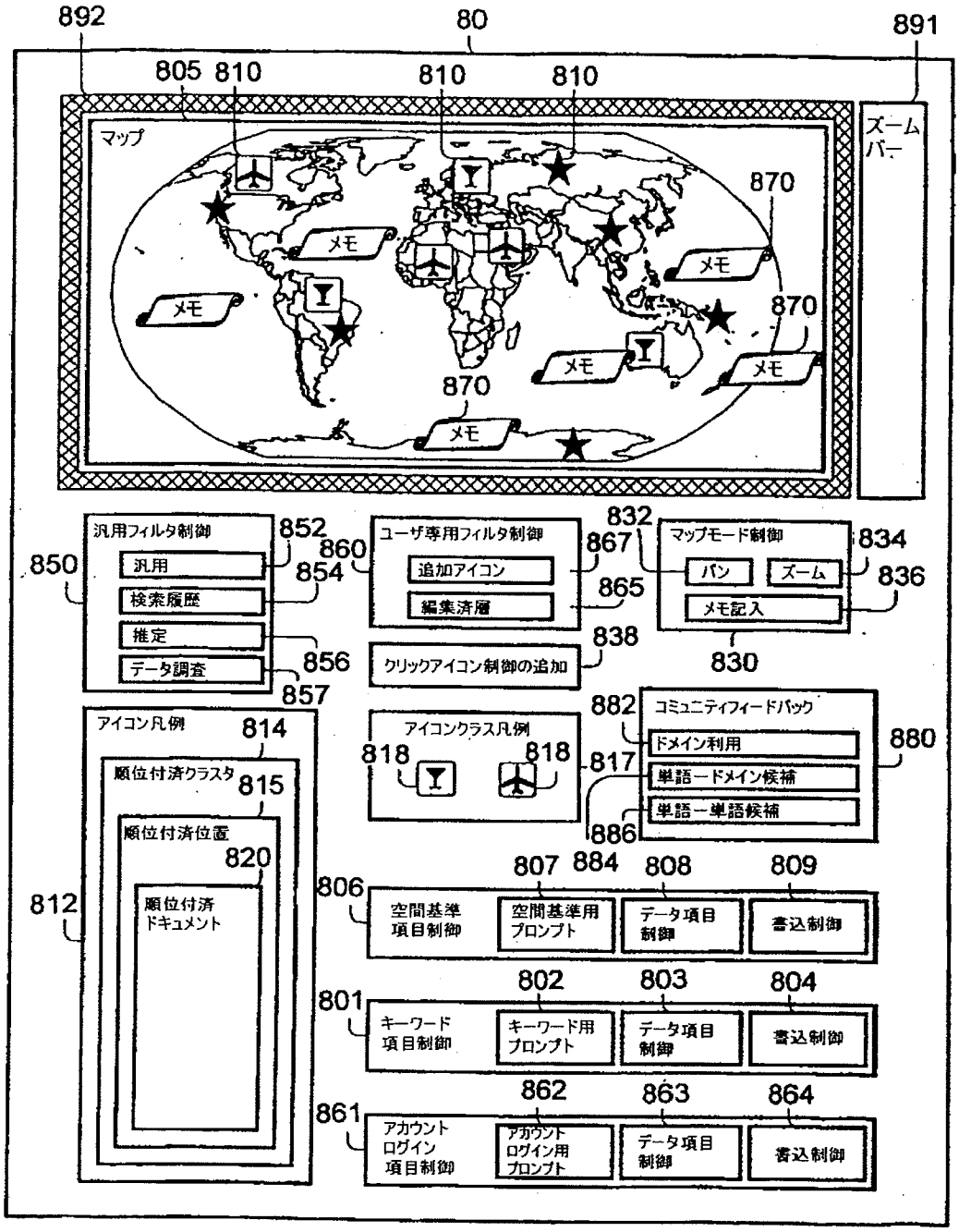
【図６】　空間キーワード索引付け部の構築プロセスにおけるステップを示す説明図。

【図７】　空間索引付け部のプロセスにおけるステップを示す説明図。

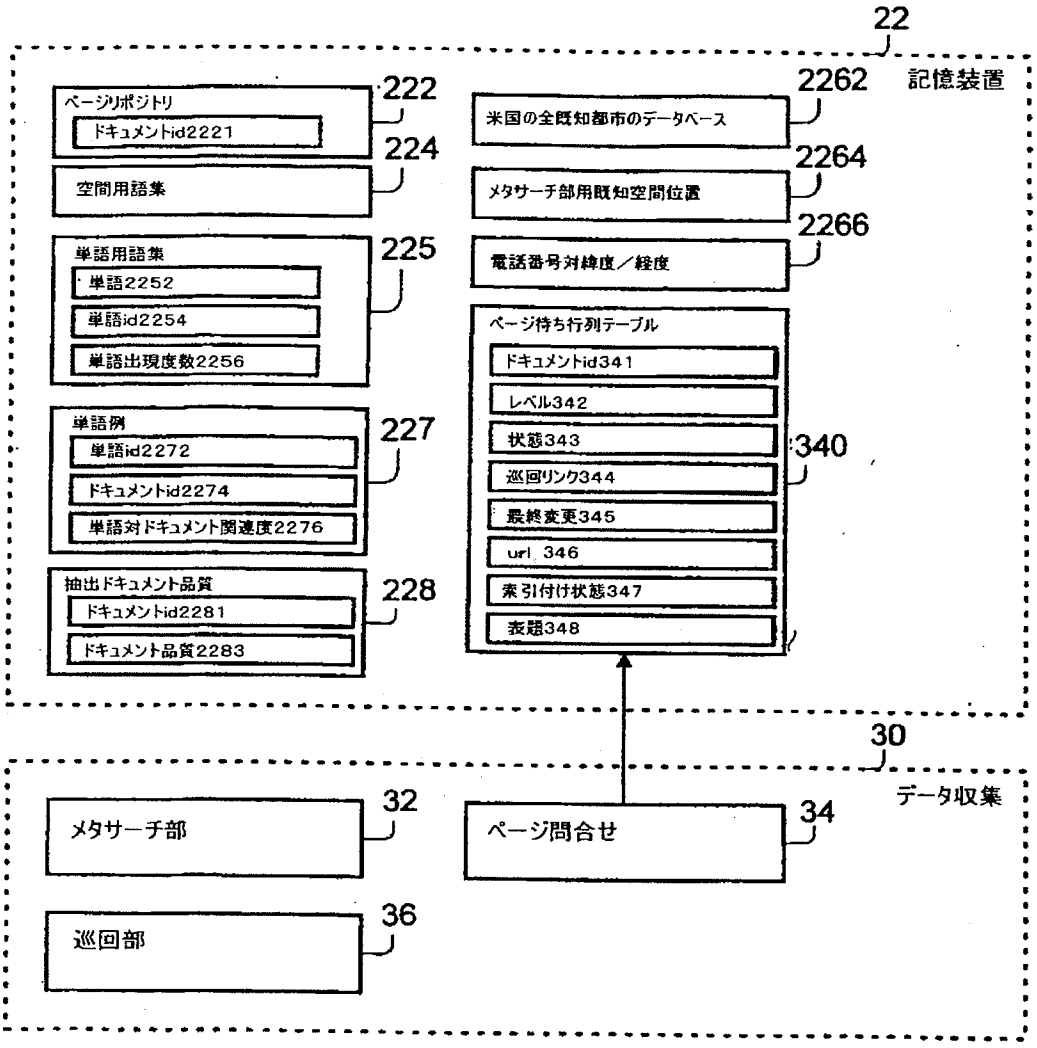
【図１】



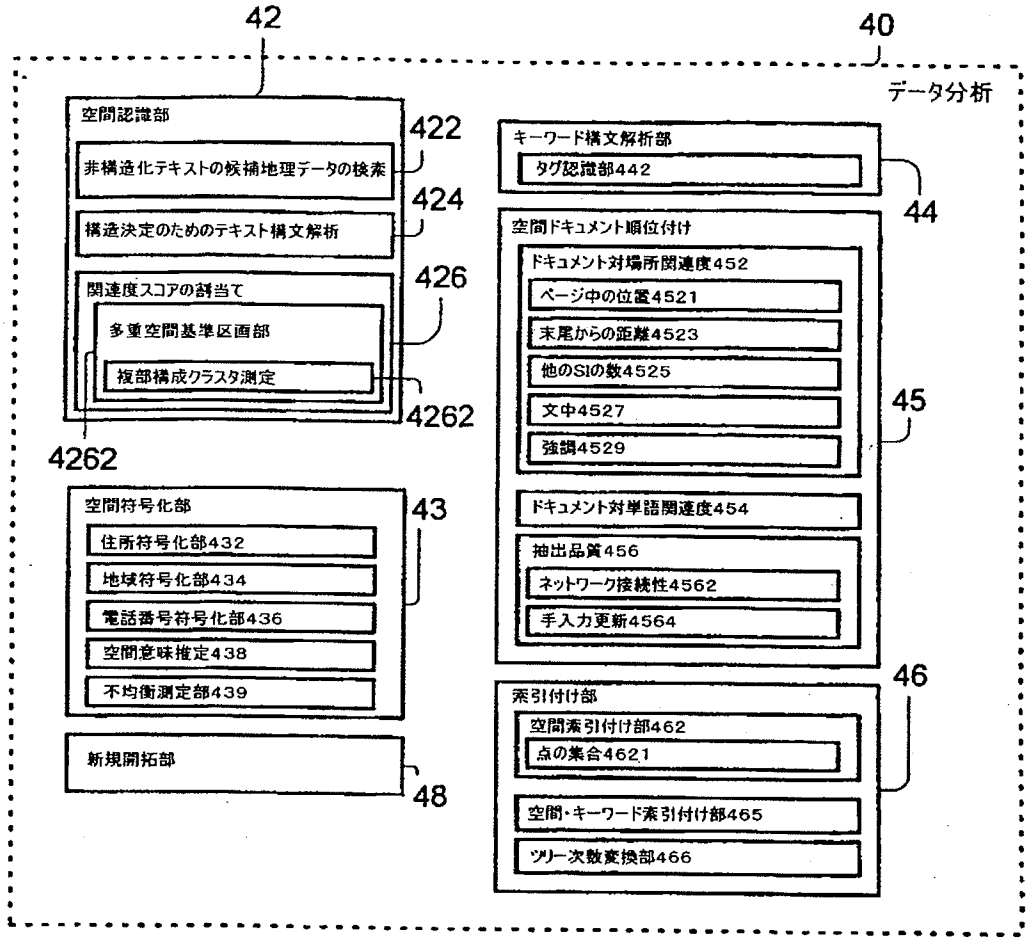
【図２】


892
80
891
805
810
810
810
マップ
ズームバー
870
メモ
メモ
メモ
870
メモ
メモ
870
メモ
汎用フィルタ制御
852
850
汎用
検索履歴
推定
データ調査
854
856
857
860
ユーザ専用フィルタ制御
追加アイコン
編集済層
832
867
865
マップモード制御
834
パン
ズーム
メモ記入
836
830
838
クリックアイコン制御の追加
アイコン凡例
814
順位付済クラスタ
815
順位付済位置
820
順位付済ドキュメント
812
アイコンクラス凡例
818
818
882
コミュニティフィードバック
ドメイン利用
単語ードメイン候補
単語ー単語候補
880
817
886
807
884
808
809
806
空間基準項目制御
空間基準用プロンプト
データ項目制御
書込制御
802
803
804
801
キーワード項目制御
キーワード用プロンプト
データ項目制御
書込制御
862
863
864
861
アカウントログイン項目制御
アカウントログイン用プロンプト
データ項目制御
書込制御

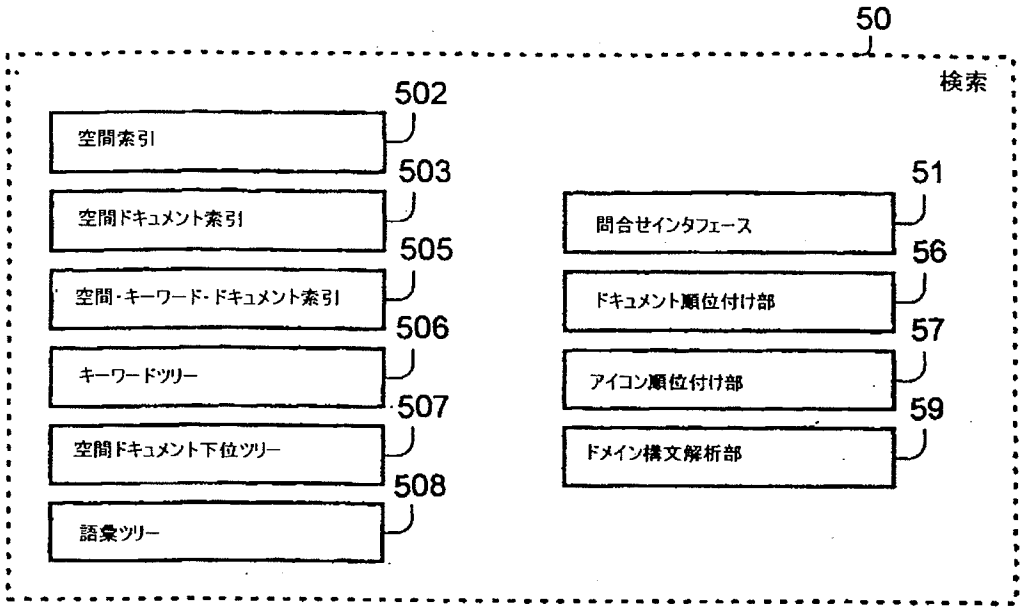
【図３】


22
記憶装置
222
ページリポジトリ
ドキュメントid2221
224
空間用語集
225
単語用語集
単語2252
単語id2254
単語出現度数2256
227
単語例
単語id2272
ドキュメントid2274
単語対ドキュメント関連度2276
228
抽出ドキュメント品質
ドキュメントid2281
ドキュメント品質2283
2262
米国の全既知都市のデータベース
2264
メタサーチ部用既知空間位置
2266
電話番号対緯度／経度
ページ待ち行列テーブル
ドキュメントid341
レベル342
状態343
巡回リンク344
最終変更345
url 346
索引付け状態347
表題348
340
30
データ収集
32
メタサーチ部
34
ページ問合せ
36
巡回部

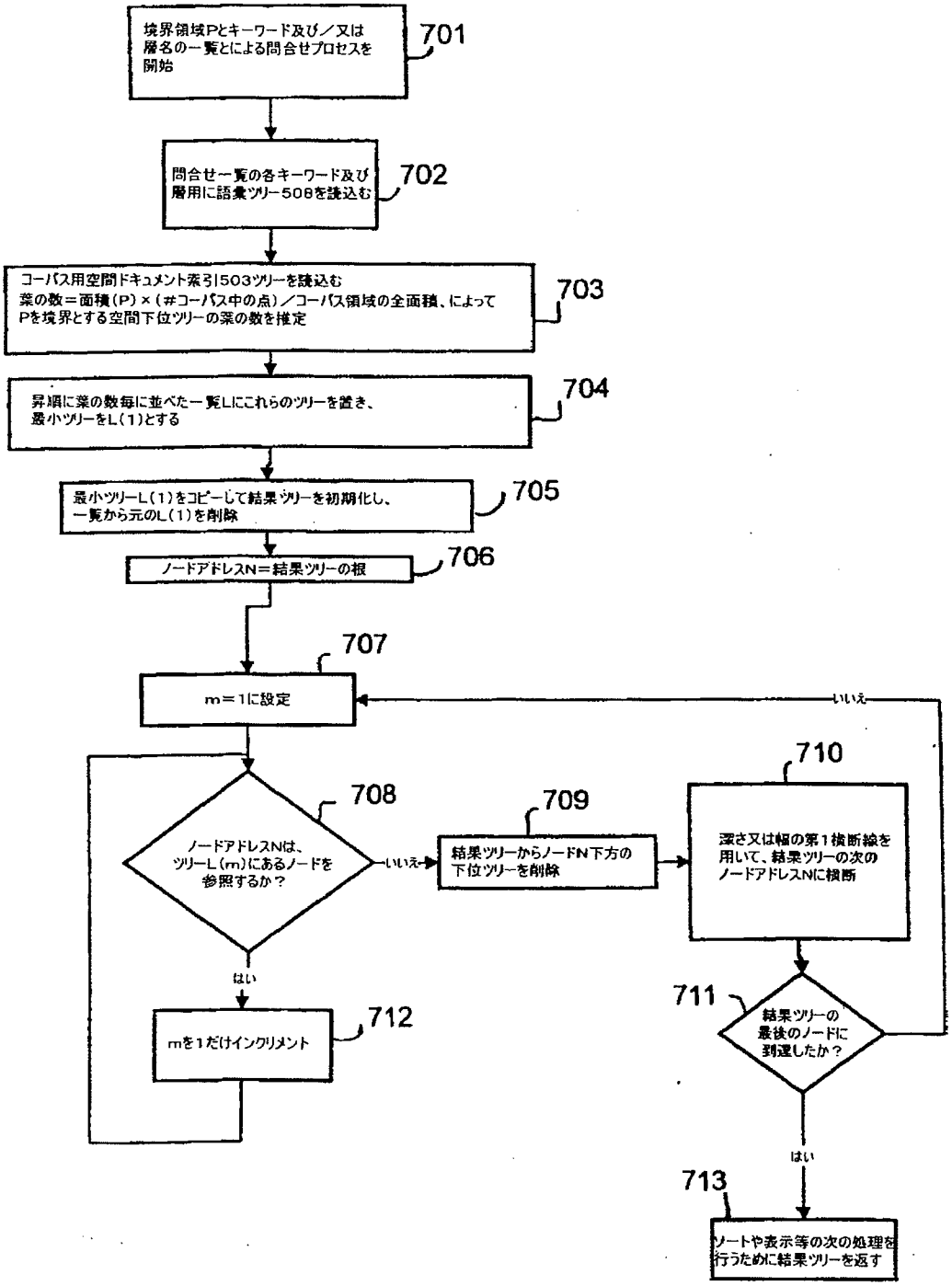
【図４】


42
40
データ分析
空間認識部
422
非構造化テキストの候補地理データの検索
424
構造決定のためのテキスト構文解析
426
関連度スコアの計当て
多重空間基準区画部
複部構成クラスタ測定
4262
4262
空間符号化部
43
住所符号化部432
地域符号化部434
電話番号符号化部436
空間意味推定438
不均衡測定部439
新規開拓部
48
キーワード構文解析部
タグ認識部442
44
空間ドキュメント順位付け
ドキュメント対場所関連度452
ページ中の位置4521
末尾からの距離4523
他のSIの数4525
文中4527
強調4529
45
ドキュメント対単語関連度454
抽出品質456
ネットワーク接続性4562
手入力更新4564
索引付け部
46
空間索引付け部462
点の集合4621
空間・キーワード索引付け部465
ツリー次数変換部466

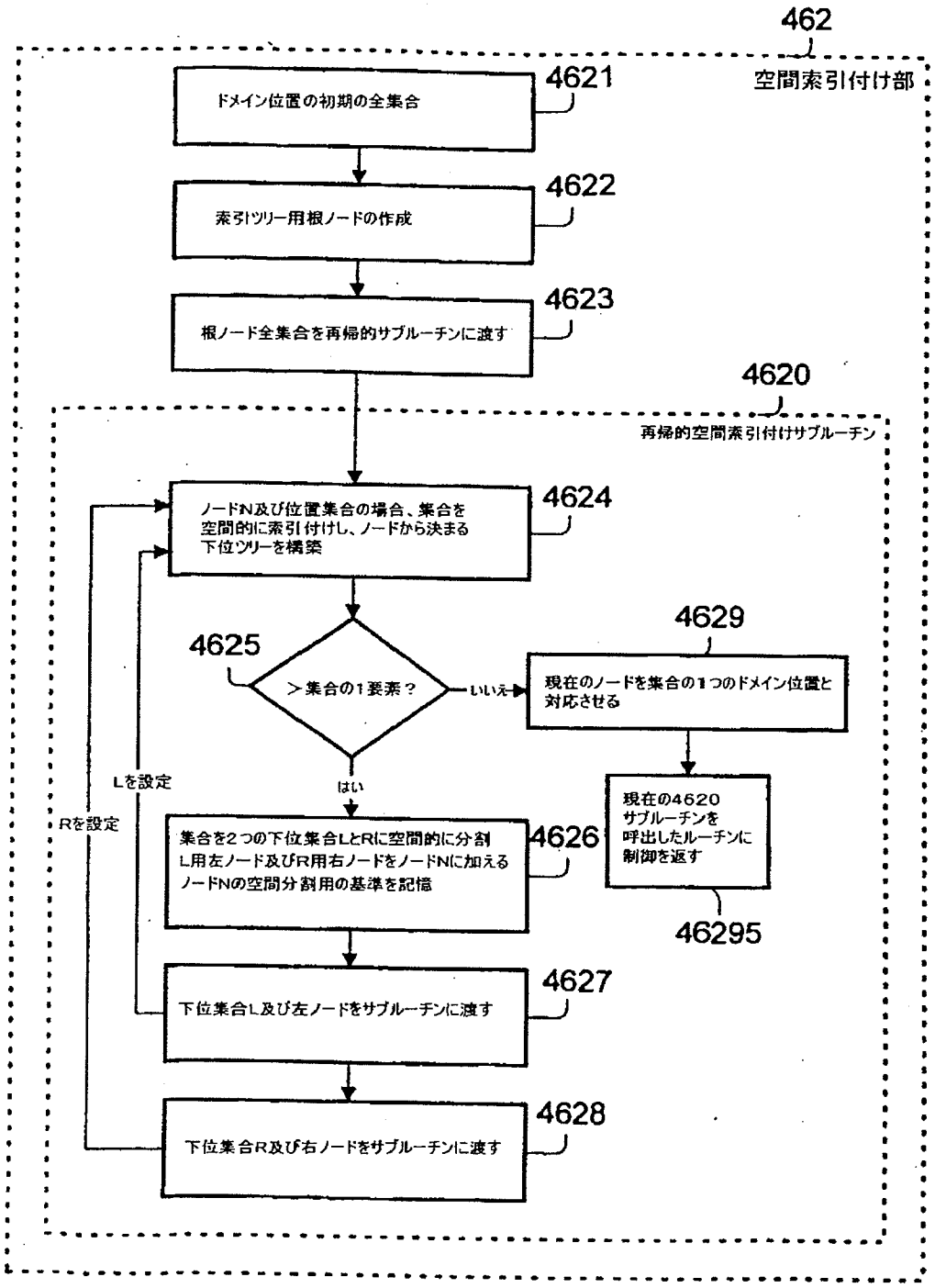
【図５】


50
検索
502
空間索引
503
空間ドキュメント索引
505
空間・キーワード・ドキュメント索引
506
キーワードツリー
507
空間ドキュメント下位ツリー
508
語彙ツリー
51
問合せインタフェース
56
ドキュメント順位付け部
57
アイコン順位付け部
59
ドメイン構文解析部

【図６】


境界領域Pとキーワード及び／又は層名の一覧とによる問合せプロセスを開始
701
問合せ一覧の各キーワード及び層用に語彙ツリー508を読込む
702
コーパス用空間ドキュメント索引503ツリーを読込む
葉の数＝面積(P)×(#コーパス中の点)／コーパス領域の全面積、によってPを境界とする空間下位ツリーの葉の数を推定
703
昇順に葉の数毎に並べた一覧Lにこれらのツリーを置き、最小ツリーをL(1)とする
704
最小ツリーL(1)をコピーして結果ツリーを初期化し、一覧から元のL(1)を削除
705
ノードアドレスN＝結果ツリーの根
706
m＝1に設定
707
いいえ
ノードアドレスNは、ツリーL(m)にあるノードを参照するか？
708
いいえ
結果ツリーからノードN下方の下位ツリーを削除
709
深さ又は幅の第1横断線を用いて、結果ツリーの次のノードアドレスNに横断
710
結果ツリーの最後のノードに到達したか？
711
はい
はい
mを1だけインクリメント
712
ソートや表示等の次の処理を行うために結果ツリーを返す
713

【図７】


462
空間索引付け部
4621
ドメイン位置の初期の全集合
4622
索引ツリー用根ノードの作成
4623
根ノード全集合を再帰的サブルーチンに渡す
4620
再帰的空間索引付けサブルーチン
4624
ノードN及び位置集合の場合、集合を空間的に索引付けし、ノードから決まる下位ツリーを構築
4625
＞集合の1要素？
いいえ
4629
現在のノードを集合の1つのドメイン位置と対応させる
はい
Lを設定
Rを設定
4626
集合を2つの下位集合LとRに空間的に分割
L用左ノード及びR用右ノードをノードNに加える
ノードNの空間分割用の基準を記憶
現在の4620サブルーチンを呼出したルーチンに制御を返す
46295
4627
下位集合L及び左ノードをサブルーチンに渡す
4628
下位集合R及び右ノードをサブルーチンに渡す

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/40173

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : G06F 17/30
US CL : 707/4, 104, 1-3, 5-8, 10; 345/348
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 707/4, 104, 1-3, 5-8, 10; 345/348

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,991,781 A (NIELSEN) 23 November 1999 (23.11.1999), ALL.	1-36
A	US 5,920,856 A (SYEDA-MAHMOOD) 06 July 1999 (06.07.1999), ALL.	1-36
A	US 5,802,361 A (WANG et al) 01 September 1998 (01.09.1998), ALL.	1-36

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 June 2001 (20.06.2001)	28 JUN 2001
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Uyen T Le Telephone No. 305-9000

Form PCT/ISA/210 (second sheet) (July 1998)

フロントページの続き

(51)Int.Cl.[7]	識別記号	F I	テーマコード（参考）
G09B 29/00		G09B 29/00	F

(81)指定国　EP(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), CA, JP

(72)発明者　ラウフ、エリック　エム.
アメリカ合衆国　07656－1103　ニュージャージー州　パーク　リッジ　サークル　ドライブ　40
(72)発明者　ドナヒュー、カレン
アメリカ合衆国　02474　マサチューセッツ州　アーリントン　レイク　ストリート　122

Fターム(参考)　2C032 HB06 HB22 HC16 HC22 HC23 HC24 HC27
5B075 ND08 NK02 NK31 NK43 PP03 PP13 PQ02 PQ12 PQ13 PQ46 PQ49 PQ60 PR06 PR08 UU14

Patent
Attorney's Docket No. 0026-0130

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of	)	
	)	
Daniel Egnor et al.	)	Group Art Unit: 2163
	)	
Application No.: 11/024,967	)	Examiner: S. Hwa
	)	
Filed: December 30, 2004	)	
	)	
For: AUTHORITATIVE DOCUMENT IDENTIFICATION	)	

INFORMATION DISCLOSURE STATEMENT UNDER 37 C.F.R. § 1.97(d)

U.S. Patent and Trademark Office
Customer Service Window, Mail Stop Amendment
Randolph Building
401 Dulany Street
Alexandria, VA 22314

Sir:

Pursuant to 37 C.F.R. §§ 1.56 and 1.97(d), applicant(s) bring(s) to the attention of the Examiner the documents listed on the attached PTO 1449. This Information Disclosure Statement is being filed after a: ☒ Final Action; or ☐ Notice of Allowance, but before payment of the issue fee - and is accompanied by the certification as specified under § 1.97(e). Applicant(s) respectfully request(s) that the Examiner consider the listed documents and indicate that they were considered by making appropriate notations on the attached PTO 1449 form.

☒ **Certification 1**: Each item of information contained in the information disclosure statement was first cited in a communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of the information disclosure statement.

☐ **Certification 2**: No item of information contained in the information disclosure statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to the knowledge of the person signing the certification after making reasonable inquiry, no item of information contained in the information disclosure statement was known to any individual designated in §1.56(c) more than three months prior to the filing of the information disclosure statement.

☒ This Information Disclosure Statement is accompanied by a fee of $180.00 as specified by Section 1.17(p).

☒ Copies of the listed documents are attached.

☐ Copies of the listed documents were previously submitted in a prior application, serial no. _____, filing date _____, upon which applicant(s) rely(ies) for the benefits provided in 35 U.S.C. § 120.

☐ The following is a concise statement of relevance of the non-English language documents.

1. _____ discloses _____.

2. _____ discloses _____.

☒ English translations of the non-English documents are enclosed.

☐ In lieu of a statement of relevance or translation of the non-English documents, an English language version of a search report from the _____ Patent Office in a corresponding application citing these documents and setting forth the relevance thereof is enclosed.

This submission does not represent that a search has been made or that no better art exists and does not constitute an admission that each or all of the listed documents are material or constitute "prior art." If the Examiner applies any of the documents as prior art against any claims in the application and applicant(s) determine(s) that the cited document(s) do not constitute "prior art" under United States law, applicant(s) reserve(s) the right to present to the office the relevant facts and law regarding the appropriate status of such documents.

Applicant(s) further reserve(s) the right to take appropriate action to establish the patentability of the disclosed invention over the listed documents, should one or more of the documents be applied against the claims of the present application.

If any copending application(s) is/are cited on the attached PTO 1449, the Examiner's attention is directed to the foregoing application(s) in compliance with § 2001.06(b) of the Manual of Patent Examining Procedure. By identifying the copending application(s), the assignee and/or applicant of the application(s) do not waive confidentiality of the application(s). Accordingly, the U.S. Patent and Trademark Office is requested to maintain the confidentiality of the copending application(s) under 35 U.S.C. § 122.

If there is any fee due in connection with the filing of this Statement, please charge the fee to our Deposit Account No. 50-1070.

Respectfully submitted,

HARRITY & HARRITY, LLP

By: /Michael S. Brooke, Reg. No. 41,641/
Michael S. Brooke
Reg. No. 41,641

11350 Random Hills Road
Suite 600
Fairfax, Virginia 22030
(571) 432-0800
CUSTOMER NUMBER: 44989
Date: July 6, 2010

Electronic Patent Application Fee Transmittal

Application Number:	11024967
Filing Date:	30-Dec-2004
Title of Invention:	Authoritative document identification
First Named Inventor/Applicant Name:	Daniel Egnor
Filer:	Michael S. Brooke/Sara Dodge
Attorney Docket Number:	0026-0130

Filed as Large Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
Submission- Information Disclosure Stmt	1806	1	180	180
		Total in USD ($)		**180**

Electronic Acknowledgement Receipt

EFS ID:	7938600
Application Number:	11024967
International Application Number:	
Confirmation Number:	7261
Title of Invention:	Authoritative document identification
First Named Inventor/Applicant Name:	Daniel Egnor
Customer Number:	44989
Filer:	Michael S. Brooke/Sara Dodge
Filer Authorized By:	Michael S. Brooke
Attorney Docket Number:	0026-0130
Receipt Date:	06-JUL-2010
Filing Date:	30-DEC-2004
Time Stamp:	16:44:11
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Credit Card
Payment was successfully received in RAM	$180
RAM confirmation Number	3491
Deposit Account	
Authorized User	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)

1	Information Disclosure Statement (IDS) Filed (SB/08)	0026-0130_1449_07-01-10.pdf	68322 70cc08485834f66ee593573110cf48fc9d14dc31	no	1
Warnings:					
Information:					
This is not an USPTO supplied IDS fillable form					
2	Foreign Reference	JP2000250931A.pdf	2412668 c8b4b19b666dce703cb676a9d34c15215bb6f63c	no	46
Warnings:					
Information:					
3	Foreign Reference	JP2000348041A.pdf	1596968 fb9282ac892e1ec22d57396720b39570572137b5	no	28
Warnings:					
Information:					
4	Foreign Reference	JP2003067419A.pdf	3433057 1feb9b0fbc2b8685022c944a105852d93188d62d	no	45
Warnings:					
Information:					
5	Foreign Reference	JP2003173280A.pdf	1125824 4da0fddc6faa16af9187fdac40722882d8ce3c2c	no	27
Warnings:					
Information:					
6	Foreign Reference	JP2004227165A.pdf	1220407 358db64fa02c5cca935c6bbdcfd01f852bf7140e	no	33
Warnings:					
Information:					
7	Foreign Reference	WO01063479.pdf	2891888 ff649888ad589cdda874f50702ca228097d664b6	no	63
Warnings:					
Information:					
8	Foreign Reference	JP2003-524259.pdf	3952385 0b902016c641d808b439142624e888898a5c4480	no	87
Warnings:					
Information:					
9	Information Disclosure Statement (IDS) Filed (SB/08)	0026-0130_IDS_07-01-10.pdf	73798 1d5cf9fe6a13ba543c6d9f1fff7fd376585b23b7a	no	4

Warnings:					
Information:					
This is not an USPTO supplied IDS fillable form					
10	Fee Worksheet (PTO-875)	fee-info.pdf	29756 b59d66e69ced2a88ed5058dd099857ceaf9b5d4a	no	2
Warnings:					
Information:					
			Total Files Size (in bytes):	16805073	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/024,967	12/30/2004	Daniel Egnor	0026-0130	7261

44989 7590 07/09/2009

HARRITY & HARRITY, LLP
11350 Random Hills Road
SUITE 600
FAIRFAX, VA 22030

EXAMINER	
HWA, SHYUE JIUNN	

ART UNIT	PAPER NUMBER
2163	

MAIL DATE	DELIVERY MODE
07/09/2009	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

United States Patent and Trademark Office

Under Secretary of Commerce for Intellectual Property and
Director of the United States Patent and Trademark Office
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

HARRITY & HARRITY, LLP
11350 RANDOM HILLS ROAD
SUITE 600
FAIRFAX, VA 22030

Appeal No: 2009-012635
Application: 11/024,967
Appellant: Daniel Egnor et al.

Board of Patent Appeals and Interferences
Docketing Notice

Application 11/024,967 was received from the Technology Center at the Board on June 29, 2009 and has been assigned Appeal No: 2009-012635.

A review of the file indicates that the following documents have been filed by appellant:

Appeal Brief filed on:	September 29, 2008
Reply Brief filed on:	June 03, 2009
Request for Hearing filed on:	NONE

In all future communications regarding this appeal, please include both the application number and the appeal number.

The mailing address for the Board is:

BOARD OF PATENT APPEALS AND INTERFERENCES
UNITED STATES PATENT AND TRADEMARK OFFICE
P.O. BOX 1450
ALEXANDRIA, VIRGINIA 22313-1450

The facsimile number of the Board is 571-273-0052. Because of the heightened security in the Washington D.C. area, facsimile communications are recommended. Telephone inquiries can be made by calling 571-272-9797 and should be directed to a Program and Resource Administrator.

By order of the Board of Patent Appeals and Interferences.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/024,967	12/30/2004	Daniel Egnor	0026-0130	7261

44989 7590 06/23/2009

HARRITY & HARRITY, LLP
11350 Random Hills Road
SUITE 600
FAIRFAX, VA 22030

EXAMINER	
HWA, SHYUE JIUNN	

ART UNIT	PAPER NUMBER
2163	

MAIL DATE	DELIVERY MODE
06/23/2009	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

UNITED STATES DEPARTMENT OF COMMERCE
U.S. Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450

APPLICATION NO./ CONTROL NO.	FILING DATE	FIRST NAMED INVENTOR / PATENT IN REEXAMINATION	ATTORNEY DOCKET NO.
11024967	12/30/2004	EGNOR ET AL.	0026-0130

HARRITY & HARRITY, LLP
11350 Random Hills Road
SUITE 600
FAIRFAX, VA 22030

EXAMINER	
JAMES HWA	
ART UNIT	**PAPER**
2163	20090619

DATE MAILED:

Please find below and/or attached an Office communication concerning this application or proceeding.

Commissioner for Patents

THE REPLY BRIEF FILED ON 06/03/2009 HAS BEEN CONSIDERED AND ENTERED BY THE EXAMINER, THE REPLY BRIEF HAS BEEN FORWARDED TO THE BOARD OF APPEAL

/Cam Y Truong/
Primary Examiner, Art Unit 2169

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/024,967	12/30/2004	Daniel Egnor	0026-0130	7261

44989 7590 06/18/2009

HARRITY & HARRITY, LLP
11350 Random Hills Road
SUITE 600
FAIRFAX, VA 22030

EXAMINER	
HWA, SHYUE JIUNN	

ART UNIT	PAPER NUMBER
2163	

MAIL DATE	DELIVERY MODE
06/18/2009	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE BOARD OF PATENT APPEALS
AND INTERFERENCES

Ex parte: DANIEL EGNOR and GEETA CHAUDHRY

Application No. 11/024,967
Technology Center 2100

Mailed: June 18, 2009

Before Deborah L. Perry, Supervisory Paralegal Specialist, Review Team.
Perry, Supervisory Paralegal Specialist, Review Team.

ORDER RETURNING UNDOCKETED APPEAL TO EXAMINER

This application was electronically received by the Board of Patent Appeals and Interferences on June 8, 2009. A review of the application revealed that it is not ready for docketing as an appeal. Accordingly, the application is herewith being returned to the Examiner to address the following matter(s) requiring attention prior to docketing.

EXAMINER'S CONSIDERATION OF REPLY BRIEF

A Reply Brief was filed in this application on June 3, 2009. Please note, this Reply Brief was not timely filed, however there is no evidence on the record indicating that the Examiner has considered and/or acknowledged the Reply Brief in accordance with 37 CFR § 41.43(a)(1) and MPEP § 1208, part II.

Consideration and/or acknowledgement of the reciept of the Reply Brief is required.

CONCLUSION

Accordingly, it is

ORDERED that the application is returned to the Examiner to:

1) consider and/or acknowledge receipt of the Reply Brief filed June 3, 2009, and

2) for such further action as may be appropriate.

If there are any questions pertaining to this Order, please contact the Board of Patent Appeals and Interferences at 571-272-9797.

DLP/bar

HARRITY & HARRITY, LLP
11350 RANDOM HILLS ROAD
SUITE 600
FAIRFAX, VA 22030

Patent
Attorney's Docket No. 0026-0130

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of	)	MAIL STOP Appeal Brief-Patents
	)	
Daniel EGNOR et al.	)	Group Art Unit: 2163
	)	
Application No.: 11/024,967	)	Examiner: S. Hwa
	)	
Filed: December 30, 2004	)	Confirmation No. 7261
	)	
For: AUTHORITATIVE DOCUMENT IDENTIFICATION	)	
	)	

U.S. Patent and Trademark Office
Customer Window, Mail Stop Appeal Brief - Patents
Randolph Building
401 Dulany Street
Alexandria, VA 22314

REPLY BRIEF UNDER 37 CFR § 41.41

Sir:

This Reply Brief is submitted in response to the Supplemental Examiner's Answer, dated March 3, 2009.

I. STATUS OF CLAIMS

Claims 1-29 are pending in this application. Claims 1-29 were rejected in the final Office Action, dated March 28, 2008, and are the subject of the present appeal. These claims were reproduced in the Claim Appendix of the Appeal Brief filed on September 29, 2009.

II. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. Claims 1-4, 6-8, 10, 12-18, 20-22, 24 and 26-28 have been rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over U.S. Patent No. 6,643,640 B1 to Getchius (hereinafter "GETCHIUS") in view of U.S. Patent Application No. 2002/0133374 to Agoni (hereinafter "AGONI").

B. Claims 5, 9, 11, 19, 23 and 25 have been rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over GETCHIUS in view of AGONI and further in view of U.S. Patent Application No. 2004/0064334 A1 to Nye (hereinafter "NYE").

C. Claim 29 has been rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over GETCHIUS in view of NYE.

III. ARGUMENTS

In the "Response to Arguments" section of the Examiner's Answer (pp. 22-44), the Examiner reiterates many of the allegations that are presented in the "Grounds of Rejection" section of the Examiner's Answer and in the final Office Action, dated March 28, 2008. Thus, Appellants' arguments presented in the Appeal Brief, filed September 29, 2008, are applicable to those allegations. Appellants submit the following additional remarks.

1. Claims 1-3 and 12-13.

Examiner's Point (1):

In providing reasons to combine GETCHIUS and AGONI, the Examiner alleges (Examiner's Answer, p. 23):

> As discussed above, a person of an ordinary skill in the art at the time the invention was made would recognize the advantage of Agoni to add the Agoni's teaching of case control panel identifies the client name for each case. The case control panel also displays case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144), because that would allow cross-referencing of other tables and facilitate speedy search (page 15, 0134) and genuinely improve and enhance the quality of service rendered by the professional and received by the client (page 2, paragraph 0012) as taught by Agoni.

Appellants respectfully submit that the Examiner's allegation does not address the features of claim 1.

Appellants' Response to Point (1):

Appellants argued that GETCHIUS and AGONI do not disclose or suggest determining a measure of authoritativeness of candidate documents (that were all associated with the same geographic location) for a business at that geographic location based on signals associated with the candidate documents, as recited in claim 1. The Examiner admits that GETCHIUS does not disclose "identifying signals associated with candidate documents," as recited in claim 1 (final Office Action, p. 4). The Examiner's

does not explain how AGONI could be combined with GETCHIUS to remedy this admitted deficiency of GETCHIUS. Instead, the Examiner alleges that the combination of GETCHIUS and AGONI would allow cross-referencing and facilitate a speedy search.

Furthermore, Appellants submit that one of ordinary skill in the art at the time of the invention would not find it obvious to combine GETCHIUS and AGONI, because such a combination is illogical. For example, the Examiner relies on the *control panel of attorney case records*, as disclosed in paragraph [0144] of AGONI, as allegedly corresponding to signals associated with the candidate document, as recited in claim 1 (see, for example, final Office Action, p. 4 and Examiner's Answer, p. 23). The Examiner also relies on *query terms*, as disclosed in col. 28, lines 7-11 of GETCHIUS, as allegedly corresponding to candidate documents, as recited in claim 1 (see, for example, final Office Action, pp. 3-4 and Examiner's Answer, p. 24). Therefore, keeping this interpretation in mind, combining the *control panel of attorney case records* of AGONI with the *query terms* of GETCHIUS, and applying this combination to the above-noted feature of claim 1, lead to the illogical feature of determining a measure of authoritativeness of *query terms* (that were all associated with the same geographic location) for a business at that geographic location based on a *control panel of attorney case records* associated with the *query terms*. Appellants submit that one of ordinary skill in the art at the time of the invention would not seek to modify GETCHIUS to obtain such a feature.

Examiner's Point (2):

The Examiner relies on Fig. 34 of GETCHIUS for allegedly disclosing "candidate *query terms* for a business" and points to the example of the "MA AND

RESTAURANTS AND FLOWERSHOPS" query disclosed in Fig. 34 of GETCHIUS (Examiner's Answer, p. 24).

Appellants' Response to Point (2):

Appellants respectfully submit that the "MA AND RESTAURANTS AND FLOWERSHOPS" query is not a candidate query for a business. In other words, the "MA AND RESTAURANTS AND FLOWERSHOPS" query is not a query for a specific business, but is rather a query to identify businesses that match the query. Therefore, Fig. 34 of GETCHIUS does not disclose or suggest determining a measure of authoritativeness of candidate *queries* (that were all associated with the same geographic location) for a business at that geographic location based on signals associated with the candidate *queries*, as would be required by claim 1 based on the Examiner's interpretation of the *query* of GETCHIUS as allegedly corresponding to the term "documents," as recited in claim 1.

Examiner's Point (3):

Appellants argued that *query terms* cannot be reasonably held to correspond to candidate documents for a business, as recited in claim 1, and that *categories of business listings* cannot be reasonably held to correspond to a business, as recited in claim 1. In response to this argument, the Examiner alleges that GETCHIUS teaches that various business listings may be grouped together in categories (Examiner's Answer, p. 25). The Examiner further alleges that each business listing may be represented as a document stored in the primary and secondary databases (Examiner's Answer, p. 25).

Appellants' Response to Point (3):

Appellants respectfully submit that these allegations do not address Appellants' arguments. For example, the Examiner did not explain why *query terms*, as disclosed by GETCHIUS, could be reasonably interpreted as candidate documents for a business, as recited in claim 1 or why *categories of business listings*, as disclosed by GETCHIUS, could be reasonably interpreted as a business, as recited in claim 1.

Examiner's Point (4):

In response to Appellants' argument that AGONI does not disclose or suggest "identifying signals associated with the candidate documents," as recited in claim 1, the Examiner relies on paragraph [0018] and paragraph [0144] of AGONI for allegedly disclosing this feature (Examiner's Answer, p. 27).

Appellants' Response to Point (4):

Paragraph [0144] of AGONI was addressed on p. 12 of the Appeal Brief. Appellants will additionally address paragraph [0018] of AGONI.

Paragraph [0018] of AGONI discloses a service system that includes profile data representing characteristics of service providers, a search engine responsive to search criteria to search the profile data, a communication module that makes the results available to a client, a case communication module to receive status data from a selected service provider, and a billing module. The search criteria can include a first importance level assigned to first profile criteria and a second importance level assigned to second profile criteria. The result data is presented such that a first group of service providers that match the first profile criteria is presented at the front of the list of results. The communication module receives candidate data representing a candidate set of service

providers comprising one or more service providers identified by the result data, the communication module receiving and storing the service summary data representing needed services and making the service summary data available to each of the candidate set of service providers.

The Examiner appears to be alleging that the *candidate set of service providers* disclosed by this section of AGONI corresponds to candidate documents, as recited in claim 1. However, this section of AGONI does not disclose or suggest identifying signals associated with the *candidate set of service providers*, as would be required by claim 1 based on the Examiner's interpretation of AGONI. Instead, this section of AGONI discloses receiving status data from a selected service provider selected from the candidate set of service providers.

If the Examiner interprets *receiving status data from a selected service provider selected from the candidate set of service providers*, as disclosed by AGONI, as allegedly corresponding to identifying signals associated with candidate documents, as recited in claim 1, (an interpretation Appellants' do not agree with), then the Examiner must maintain that interpretation for all features of claim 1 that include the term "signals." Therefore, since claim 1 recites determining a measure of authoritativeness of candidate documents (that were all associated with the same geographic location) for a business at that geographic location based on signals associated with the candidate documents, the combination of GETCHIUS and AGONI would have to disclose determining a measure of authoritativeness of a *candidate set of service providers* (that were all associated with the same geographic location) for a business at that geographic location based on *receiving status data from a selected service provider selected from the candidate set of*

service providers. The Examiner has not explained how a combination of GETCHIUS and AGONI would disclose such a feature.

For example, the Examiner relies on *query terms*, as disclosed by GETCHIUS, for allegedly corresponding to candidate documents, as recited in claim 1. The Examiner has not explained how the *query terms* of GETCHIUS could be replaced with the *candidate set of service providers* disclosed by AGONI or why one or ordinary skill in the art would find it obvious to do so.

Examiner's Point (5):

In response to Appellants' arguments that GETCHIUS does not disclose determining a measure of authoritativeness of a candidate document for a business at a location, the Examiner alleges GETCHIUS discloses that a subset of query terms is determined such that a string representing a particular query is uniquely mapped to a name corresponding to a data set (Examiner's Answer, p, 28). Thus, it appears that the Examiner is alleging that *uniquely mapping a query string to a name in a data set*, as disclosed by GETCHIUS, corresponds to determining a measure of authoritativeness of a candidate document for a business at a location, as recited in claim 1.

Appellants' Response to Point (5):

Appellants respectfully disagree with the Examiner's allegation.

A measure of authoritativeness of a particular document for a business at a location determines how authoritative the particular document is as a document for the business at the location. In other words, a document with a higher measure of authoritativeness is perceived as a more reliable document. Uniquely mapping a query

string to a name in a data set does not determine how authoritative the query string is as a query string for the name.

Furthermore, even if it is assumed, for the sake of argument, that "uniquely mapping a query string to a name in a data set" can be construed as determining a measure of authoritativeness for the query string (a point Appellants do not agree with), GETCHIUS and AGONI, whether taken alone or in any reasonable combination, do not disclose or suggest uniquely mapping a query string to a name in a data set based on a *candidate set of service providers*, as would be required by claim 1 based on the Examiner relying on the *candidate set of service providers* of AGONI as allegedly corresponding to the signals associated with the candidate documents, as recited in claim 1.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claims 1-3 and 12-13 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

2. Claim 4.

Examiner's Point (6):

Appellants' argued that GETCHIUS and AGONI do not disclose or suggest that identifying a set of documents further includes determining documents that are linked to by candidate documents, and identifying the determined document as candidate documents, as recited in claim 4.

In response, the Examiner additionally relies on col. 33, lines 25-53 of GETCHIUS, as well as Figs. 39 and 40 of GETCHIUS as allegedly disclosing the features of claim 4 (Examiner's Answer, p. 29).

Appellants' Response to Point (6):

Appellants submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 4.

Col. 33, lines 25-53 of GETCHIUS disclose:

> Once the system is initialized, the system may operate to obtain results that are to be displayed to the user. The steps for obtaining results may be seen in a flow chart 88 displayed in FIG. 41. Referring to FIG. 41, the parse driver 858 may at a step 20 parse a user query and deliver the parsed query in suitable form for handling by the query engine 862. The query engine may include the information retrieval software 908. At a step 22, the query engine 862 may operate the information retrieval software 908 to take the parsed user request and expand the query, turning the user request into a detailed query. Next, at a step 24, the information retrieval software may operate on the expanded term lists 836 by identifying documents associated with the terms identified in the expanded query. In an embodiment, the term lists 836 are the business listings described in connection with steps 8284 and 86 above, expanded to include synonyms and terms that are determined to be related to the words in the business listing. Identification of documents may be accomplished by a variety of information retrieval techniques. Documents may also be associated with queries by sorted relevancy ranking, clustering (automated grouping of related documents), automated document, summarization (creation of content abstracts, not simply the first few sentences of the document) and query-by-example (turning an individual document into a query in order to retrieve "more documents like this"). These functions may be accomplished by software techniques, such as having a table of pointers having as an argument a tokenized version of each possible term from the expanded user query from the step 22. The table of pointers may point to the location of a term list 836, for each such term. The term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the entire set of documents, the association of the document with other documents, the association of the document with categories, and the like.

This section of GETCHIUS discloses parsing and expanding a query and identifying documents associated with terms in the query. The terms lists are business listings and the document matching the terms of the query are sorted by relevancy ranking. This section of GETCHIUS further discloses that a table of pointers may point to the location of a term list and that the term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the set of documents, and the association of the document with other documents or categories.

A linked list is a data structure in which one record links to the next record in the sequence. This section of GETCHIUS does not disclose or suggest determining which records in the linked list link to any other records. Rather, this section of GETCHIUS merely discloses that a list of terms may be a linked list of documents that include the term. Furthermore, as the Examiner relies on the *query terms* of GETCHIUS as allegedly corresponding to the candidate document, as recited in claim 1, this section of GETCHIUS does not disclose or suggest determining documents in the linked list that are linked to *query terms*, as would be required by claim 4 based on the Examiner's interpretation of GETCHIUS. Therefore, this section of GETCHIUS does not disclose or suggest that identifying a set of documents further includes determining documents that are linked to by candidate documents, and identifying the determined document as candidate documents, as recited in claim 4.

Fig. 39 of GETCHIUS depicts a query engine coupled to a term list database and an advertisement banner term lists database. Fig. 40 of GETCHIUS depicts a flow graph for a process that includes accessing mark up language files, creating term lists, and expanding the term lists. Figs. 39 and 40 of GETCHIUS do not disclose or suggest that identifying a set of documents further includes determining documents that are linked to by candidate documents, and identifying the determined document as candidate documents, as recited in claim 4.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 4 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

3. Claim 6.

Examiner's Point (7):

Appellants argued that GETCHIUS does not disclose or suggest determining a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate documents that point to the candidate document, as recited in claim 6.

In response, the Examiner alleges that GETCHIUS discloses that each category is given a weight corresponding to the number of listings that are associated with the category and relies on col. 65, lines 10-25 and col. 33, lines 24-48 of GETCHIUS for allegedly disclosing the features recited in claim 6 (Examiner's Answer, p. 30).

Appellants' Response to Point (7):

Appellants submit that the Examiner's allegation does not address the features of claim 6.

A *weight given to a category*, as disclosed by GETCHIUS, cannot be reasonably held to correspond to an outlink from a document or a number of outlinks from a document, as recited in claim 6. An outlink is a hyperlink to another document. As demonstrated on pp. 17-19 of the Appeal Brief, none of the sections of GETCHIUS relied on by the Examiner in the rejection of claim 6 disclose, suggest, or even mention outlinks.

Appellants will further address the additional sections of GETCHIUS relied on by the Examiner.

Col. 65, lines 10-25 of GETCHIUS disclose:

> These statistics may be further improved by weighting other factors. For example, it is possible to weight each term that appears in one of the categories that is retrieved upon execution of a user query and to normalize the IDF and RTF statistics over the weights. Thus, if a particular category deserves a higher weight, then it might be accorded higher weight in ranking super-categories. For example, a category that is manually mapped to a super-category might be given a higher weight than a category that is automatically mapped. The user query might be given a higher or lower weight, than other information. Categories with a large number of listings may be given higher weight. In an embodiment, each category is given a weight corresponding to the number of listings that are associated with the category, normalized by dividing the total number of listings. In an embodiment, the user query terms are each given a weight of one. In the weighting process, the weight may be multiplied by the term element in performing the sum of the product of term frequency and inverse document frequency over all terms for all documents in the super-category linked list.

This section of GETCHIUS discloses weighing terms that appear in categories based on the number of listings in a category. A number of listings in a category cannot be reasonably held to correspond to a number of outlinks. As stated above, an outlink is a hyperlink to another document. This section of GETCHIUS does not disclose, suggest, or even mention a number of outlinks. Therefore, this section of GETCHIUS does not disclose or suggest determining a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate documents that point to the candidate document, as recited in claim 6.

Col. 33, lines 24-28 of GETCHIUS were reproduced above. This section of GETCHIUS discloses that a table of pointers may point to the location of a term list and that the term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the set of documents, and the association of the document with other document or categories.

This section of GETCHIUS does not disclose, suggest, or even mention a number of outlinks. Therefore, this section of GETCHIUS does not disclose or suggest determining a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate documents that point to the candidate document, as recited in claim 6.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 6 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

4. Claim 7.

Examiner's Point (8):

(8) Appellants argued that GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes identifying anchor text associated with links to the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location, as recited in claim 7.

In response, the Examiner additionally relies on col. 40, lines 25-37 and on col. 62, lines 49-60 of GETCHIUS for allegedly disclosing the features recited in claim 7 (Examiner's Answer, p. 31).

Appellants' Response to Point (8):

Appellants submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 7.

Col. 40, lines 25-37 of GETCHIUS disclose:

> At step 1008, the procedure "match phone number" is performed to produce a subset of one or more entries of the existing database which match the existing phone number. Control proceeds to step 1010 where the procedure "name match" is performed. Generally, "name match" will be described in paragraphs that follow to determine whether there is a business name match for a particular entry. Control proceeds to step 1012 where "derive score" is performed based on the zip code and the name match score. Generally, the result of step 1012 produces a score representing a statistic relative to determining whether two entries in a particular database and an updated version of the database match.

This section of GETCHIUS discloses a match phone number procedure, a name match procedure, and a derive score based on a zip code and the name match score. This section of GETCHIUS does not disclose, suggest, or even mention anchor text, let alone matching anchor text to a name of a business at a location. Anchor text is the visible, clickable text in a hyperlink. This section of GETCHIUS does not even mention hyperlinks. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes identifying anchor text associated with links to the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location, as recited in claim 7.

Col. 62, lines 49-60 of GETCHIUS disclose:

> Once control has returned to the flow chart 52 of FIG. 68, meaning that all yellow pages categories have been mapped to a super-category, at a step 77 the banner ad retrieval software 909 may index the various super-categories in a banner ad term list 837. The banner ad term list 837 may take the form of a linked list of the super-categories, with each element in the list consisting of all of the terms that appear in the super-category, as well as all of the terms that appear in each of the categories that was matched to the super-category.

This section of GETCHIUS discloses mapping yellow pages categories to a super-category and indexing the super-categories in a banner ad term list, in the form of a linked list. This section of GETCHIUS does not disclose, suggest, or even mention anchor text, let alone matching anchor text to a name of a business at a location. As stated above, anchor text is the visible, clickable text in a hyperlink. This section of GETCHIUS does not even mention hyperlinks. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes identifying anchor text associated with links to the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location, as recited in claim 7.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 7 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

5. Claim 8.

Examiner's Point (9):

Appellants argued that GETCHIUS does not disclose or suggest that wherein identifying signals associated with the candidate documents includes identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the

candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location.

In response, the Examiner additionally relies on col. 9, lines 37-43; Figs. 9 and 10; col. 47, lines 11-30; and Fig. 58 of GETCHIUS for allegedly disclosing the features of claim 8 (Examiner's Answer, pp. 32-33).

Appellants' Response to Point (9):

Appellants respectfully submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 8.

Col. 9, lines 37-43 of GETCHIUS, which describe Figs. 9 and 10 of GETCHIUS, disclose:

> Referring now to FIGS. 9 and 10, shown is one embodiment of a user interface for displaying a first page of the top query categories 1820. Generally, these categories are associated with the various business listings and are tags by which a user may perform queries. In this embodiment, for example, the user may select the "top categories" from the initial interface as included in the field 1802.

This section of GETCHIUS discloses a user interface for displaying a first page of top query categories that are associated with various business listings and are tags by which a user may perform queries. A user may select the top categories from the interface. This section of GETCHIUS does not disclose or suggest titles for documents, let alone determining whether a title of a document matches the name of a business. Therefore, this section of GETCHIUS does not disclose or suggest that wherein identifying signals associated with the candidate documents includes identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location, as recited in claim 8.

Col. 47, lines 11-30 of GETCHIUS, which describe Fig. 58 of GETCHIUS, disclose:

> Referring now to FIG. 58, shown is a flowchart of an embodiment of method steps for detecting duplicates in the category file. Generally, these steps are more detailed processing steps of step 1520 of FIG. 57. At step 1500, a first category name in the category file of the unfiltered database is tokenized. In other words, each word included in the heading or category name is associated with a token. Similarly, in step 1504, the next record of a category is examined and also tokenized. At step 1506, a comparison of the two tokenized names is performed to derive a score in accordance with the number of matching name components. This may also be normalized, as described in accordance with the foreign source update processing techniques. At step 1508, a determination is made as to whether or not the score is greater than a predetermined threshold. In this instance, the threshold is 75%. If the score is greater than the threshold, control proceeds to step 1512 where the categories are tagged as duplicates propagating any previous matching identifier tag. In other words, the transitive matching technique is used in marking matching categories. For example, if ID1,=ID2. Then, it is determined that ID2=ID5, ID5, is also marked as having ID1, as a matching identifier. Similarly, subsequent matches to ID5, f further propagate the value ID11. Subsequently, control proceeds to steps 1510 for advancement to the next record. If it is determined at step 1508 that the score is not greater than the threshold, no match is found and control proceeds to step 1510 where the next category is advanced to. At step 1514, a determination is made as to whether all the categories have been processed in the category file. If they have, control proceeds to step 1516 where processing stops. Otherwise, control proceeds to step 1504 for further comparisons and determinations of equivalent categories.

This section of GETCHIUS discloses a process that includes tokenizing a first category name and examining and tokenizing the next record of a category. The two tokenized records are compared and a score is derived and normalized. If the score is greater than a threshold, the categories are tagged as duplicates. If the score is not greater than the threshold, no match is found and control proceeds to the next category. This section of GETCHIUS does not disclose or suggest titles for documents, let alone determining whether a title of a document matches the name of a business. Therefore, this section of GETCHIUS does not disclose or suggest that wherein identifying signals associated with the candidate documents includes identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location, as recited in claim 8.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 8 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

6. Claim 10.

Examiner's Point (10):

Appellants argued that GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

In response, the Examiner additionally relies on col. 18, lines 18-26; col. 30, lines 62-67; and col. 43, lines 35-46 of GETCHIUS for allegedly disclosing the features of claim 8 (Examiner's Answer, pp. 33-34).

Appellants' Response to Point (10):

Appellants respectfully submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 10.

Col. 18, lines 18-26 of GETCHIUS disclose:

> In this particular embodiment, the databases may include business information, such as for specific businesses or classifications of businesses. Additionally, data queries may be performed based on characteristics of the various businesses, such as location, name, or category. Furthermore, the architecture described herein supports a flexible presentation of these businesses, based on business agreements and service offerings.

This section of GETCHIUS discloses that the databases may include business information, such as specific businesses or classification of businesses. Data queries may be performed based on the characteristics of the various businesses, such as location,

name, or category. This section of GETCHIUS does not disclose or suggest increasing the authoritativeness of a candidate document when the candidate document is associated with a single location. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

Col. 30, lines 59-67 of GETCHIUS disclose:

> Referring now to FIGS. 37 and 38 shown is a flowchart of an embodiment of a method for integrating total-city and multi-city cache results into "normal" cached search results. At step 260, a total-city cache name corresponding to the data query is formed. In one embodiment, the total city cache name is formed by starting with the string "SCOPE=T" to identify a total-city name. Additionally, the following information is extracted from the original query string, as formed by the parser: category, category id, business name, street address, keywords, longitude, latitude.

This section of GETCHIUS discloses forming a total-city cache name corresponding to a data query and extracting, from the original query string, a category, category identification, a business name, a street address, keywords, longitude, and latitude. This section of GETCHIUS does not disclose or suggest determining whether a document is associated with a single location, let alone increasing a measure of authoritativeness based on whether the candidate document is associated with a single location. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of

authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

Col. 43, lines 35-46 of GETCHIUS disclose:

> Referring now to FIG. 50, shown as a flow chart of the steps of one embodiment for performing the routine "derive score", as performed from step 1012 of FIG. 46. Generally, code. At step 1080, the score previously derived from name match for each entry is updated by one if the zip codes of an existing database entry match an updated entry. At step 1082 this score is normalized by taking the score computed thus far and dividing it by the number of tokens in produced a normalized score as in step 1082. At step 1084, control returns to the point of call. In this particular instance, control returns to step 1012 where processing resumes with step 1020 of FIG. 47.

This section of GETCHIUS discloses that a score derived from a name match is updated if one of the zip codes of an existing database entry matches an updated entry, and that the score is normalized by dividing it by a number of tokens. This section of GETCHIUS does not disclose or suggest determining whether a document is associated with a single location, let alone increasing a measure of authoritativeness based on whether the candidate document is associated with a single location. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 10 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

IV. CONCLUSION

In view of the foregoing arguments and those arguments presented in the Appeal Brief, Appellants respectfully solicit the Honorable Board to reverse the Examiner's rejections of claims 1-29.

To the extent necessary, a petition for an extension of time under 37 C.F.R. § 1.136 is hereby made. Please charge any shortage in fees due in connection with the filing of this paper, including extension of time fees, to Deposit Account No. 50-1070 and please credit any excess fees to such deposit account.

Respectfully submitted,

HARRITY & HARRITY, LLP

/Viktro Simkovic, Reg No. 56,012/
Viktor Simkovic
Reg. No. 56,012

Date: June 3, 2009
11350 Random Hills Road
Suite 600
Fairfax, Virginia 22030
(571) 432-0800 main
(571) 432-0899 direct

Customer No.: 44989

Electronic Acknowledgement Receipt

EFS ID:	5447691
Application Number:	11024967
International Application Number:	
Confirmation Number:	7261
Title of Invention:	Authoritative document identification
First Named Inventor/Applicant Name:	Daniel Egnor
Customer Number:	44989
Filer:	Viktor Simkovic/Sandra Stocklinski
Filer Authorized By:	Viktor Simkovic
Attorney Docket Number:	0026-0130
Receipt Date:	03-JUN-2009
Filing Date:	30-DEC-2004
Time Stamp:	16:34:00
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Reply Brief Filed	0026-0130_Reply_Brief_PTO.pdf	143919 7b41da2ea8cfe0af67e7279a3b8135d7c1d9157a	no	23

Warnings:	
Information:	

Total Files Size (in bytes):	143919

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

Patent
Attorney's Docket No. 0026-0130

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of	)	MAIL STOP Appeal Brief-Patents
	)	
Daniel EGNOR et al.	)	Group Art Unit: 2163
	)	
Application No.: 11/024,967	)	Examiner: S. Hwa
	)	
Filed: December 30, 2004	)	Confirmation No. 7261
	)	
For: AUTHORITATIVE DOCUMENT IDENTIFICATION	)	
	)	

U.S. Patent and Trademark Office
Customer Window, Mail Stop Appeal Brief - Patents
Randolph Building
401 Dulany Street
Alexandria, VA 22314

REPLY BRIEF UNDER 37 CFR § 41.41

Sir:

This Reply Brief is submitted in response to the Supplemental Examiner's Answer, dated March 3, 2009.

I. STATUS OF CLAIMS

Claims 1-29 are pending in this application. Claims 1-29 were rejected in the final Office Action, dated March 28, 2008, and are the subject of the present appeal. These claims were reproduced in the Claim Appendix of the Appeal Brief filed on September 29, 2009.

II. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. Claims 1-4, 6-8, 10, 12-18, 20-22, 24 and 26-28 have been rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over U.S. Patent No. 6,643,640 B1 to Getchius (hereinafter "GETCHIUS") in view of U.S. Patent Application No. 2002/0133374 to Agoni (hereinafter "AGONI").

B. Claims 5, 9, 11, 19, 23 and 25 have been rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over GETCHIUS in view of AGONI and further in view of U.S. Patent Application No. 2004/0064334 A1 to Nye (hereinafter "NYE").

C. Claim 29 has been rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over GETCHIUS in view of NYE.

III. ARGUMENTS

In the "Response to Arguments" section of the Examiner's Answer (pp. 22-44), the Examiner reiterates many of the allegations that are presented in the "Grounds of Rejection" section of the Examiner's Answer and in the final Office Action, dated March 28, 2008. Thus, Appellants' arguments presented in the Appeal Brief, filed September 29, 2008, are applicable to those allegations. Appellants submit the following additional remarks.

1. Claims 1-3 and 12-13.

Examiner's Point (1):

In providing reasons to combine GETCHIUS and AGONI, the Examiner alleges (Examiner's Answer, p. 23):

> As discussed above, a person of an ordinary skill in the art at the time the invention was made would recognize the advantage of Agoni to add the Agoni's teaching of case control panel identifies the client name for each case. The case control panel also displays case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144), because that would allow cross-referencing of other tables and facilitate speedy search (page 15, 0134) and genuinely improve and enhance the quality of service rendered by the professional and received by the client (page 2, paragraph 0012) as taught by Agoni.

Appellants respectfully submit that the Examiner's allegation does not address the features of claim 1.

Appellants' Response to Point (1):

Appellants argued that GETCHIUS and AGONI do not disclose or suggest determining a measure of authoritativeness of candidate documents (that were all associated with the same geographic location) for a business at that geographic location based on signals associated with the candidate documents, as recited in claim 1. The Examiner admits that GETCHIUS does not disclose "identifying signals associated with candidate documents," as recited in claim 1 (final Office Action, p. 4). The Examiner's

does not explain how AGONI could be combined with GETCHIUS to remedy this admitted deficiency of GETCHIUS. Instead, the Examiner alleges that the combination of GETCHIUS and AGONI would allow cross-referencing and facilitate a speedy search.

Furthermore, Appellants submit that one of ordinary skill in the art at the time of the invention would not find it obvious to combine GETCHIUS and AGONI, because such a combination is illogical. For example, the Examiner relies on the *control panel of attorney case records*, as disclosed in paragraph [0144] of AGONI, as allegedly corresponding to signals associated with the candidate document, as recited in claim 1 (see, for example, final Office Action, p. 4 and Examiner's Answer, p. 23). The Examiner also relies on *query terms*, as disclosed in col. 28, lines 7-11 of GETCHIUS, as allegedly corresponding to candidate documents, as recited in claim 1 (see, for example, final Office Action, pp. 3-4 and Examiner's Answer, p. 24). Therefore, keeping this interpretation in mind, combining the *control panel of attorney case records* of AGONI with the *query terms* of GETCHIUS, and applying this combination to the above-noted feature of claim 1, lead to the illogical feature of determining a measure of authoritativeness of *query terms* (that were all associated with the same geographic location) for a business at that geographic location based on a *control panel of attorney case records* associated with the *query terms*. Appellants submit that one of ordinary skill in the art at the time of the invention would not seek to modify GETCHIUS to obtain such a feature.

Examiner's Point (2):

The Examiner relies on Fig. 34 of GETCHIUS for allegedly disclosing "candidate *query terms* for a business" and points to the example of the "MA AND

RESTAURANTS AND FLOWERSHOPS" query disclosed in Fig. 34 of GETCHIUS (Examiner's Answer, p. 24).

Appellants' Response to Point (2):

Appellants respectfully submit that the "MA AND RESTAURANTS AND FLOWERSHOPS" query is not a candidate query for a business. In other words, the "MA AND RESTAURANTS AND FLOWERSHOPS" query is not a query for a specific business, but is rather a query to identify businesses that match the query. Therefore, Fig. 34 of GETCHIUS does not disclose or suggest determining a measure of authoritativeness of candidate *queries* (that were all associated with the same geographic location) for a business at that geographic location based on signals associated with the candidate *queries*, as would be required by claim 1 based on the Examiner's interpretation of the *query* of GETCHIUS as allegedly corresponding to the term "documents," as recited in claim 1.

Examiner's Point (3):

Appellants argued that *query terms* cannot be reasonably held to correspond to candidate documents for a business, as recited in claim 1, and that *categories of business listings* cannot be reasonably held to correspond to a business, as recited in claim 1. In response to this argument, the Examiner alleges that GETCHIUS teaches that various business listings may be grouped together in categories (Examiner's Answer, p. 25). The Examiner further alleges that each business listing may be represented as a document stored in the primary and secondary databases (Examiner's Answer, p. 25).

Appellants' Response to Point (3):

Appellants respectfully submit that these allegations do not address Appellants' arguments. For example, the Examiner did not explain why *query terms*, as disclosed by GETCHIUS, could be reasonably interpreted as candidate documents for a business, as recited in claim 1 or why *categories of business listings*, as disclosed by GETCHIUS, could be reasonably interpreted as a business, as recited in claim 1.

Examiner's Point (4):

In response to Appellants' argument that AGONI does not disclose or suggest "identifying signals associated with the candidate documents," as recited in claim 1, the Examiner relies on paragraph [0018] and paragraph [0144] of AGONI for allegedly disclosing this feature (Examiner's Answer, p. 27).

Appellants' Response to Point (4):

Paragraph [0144] of AGONI was addressed on p. 12 of the Appeal Brief. Appellants will additionally address paragraph [0018] of AGONI.

Paragraph [0018] of AGONI discloses a service system that includes profile data representing characteristics of service providers, a search engine responsive to search criteria to search the profile data, a communication module that makes the results available to a client, a case communication module to receive status data from a selected service provider, and a billing module. The search criteria can include a first importance level assigned to first profile criteria and a second importance level assigned to second profile criteria. The result data is presented such that a first group of service providers that match the first profile criteria is presented at the front of the list of results. The communication module receives candidate data representing a candidate set of service

providers comprising one or more service providers identified by the result data, the communication module receiving and storing the service summary data representing needed services and making the service summary data available to each of the candidate set of service providers.

The Examiner appears to be alleging that the *candidate set of service providers* disclosed by this section of AGONI corresponds to candidate documents, as recited in claim 1. However, this section of AGONI does not disclose or suggest identifying signals associated with the *candidate set of service providers*, as would be required by claim 1 based on the Examiner's interpretation of AGONI. Instead, this section of AGONI discloses receiving status data from a selected service provider selected from the candidate set of service providers.

If the Examiner interprets *receiving status data from a selected service provider selected from the candidate set of service providers*, as disclosed by AGONI, as allegedly corresponding to identifying signals associated with candidate documents, as recited in claim 1, (an interpretation Appellants' do not agree with), then the Examiner must maintain that interpretation for all features of claim 1 that include the term "signals." Therefore, since claim 1 recites determining a measure of authoritativeness of candidate documents (that were all associated with the same geographic location) for a business at that geographic location based on signals associated with the candidate documents, the combination of GETCHIUS and AGONI would have to disclose determining a measure of authoritativeness of a *candidate set of service providers* (that were all associated with the same geographic location) for a business at that geographic location based on *receiving status data from a selected service provider selected from the candidate set of*

service providers. The Examiner has not explained how a combination of GETCHIUS and AGONI would disclose such a feature.

For example, the Examiner relies on *query terms*, as disclosed by GETCHIUS, for allegedly corresponding to candidate documents, as recited in claim 1. The Examiner has not explained how the *query terms* of GETCHIUS could be replaced with the *candidate set of service providers* disclosed by AGONI or why one or ordinary skill in the art would find it obvious to do so.

Examiner's Point (5):

In response to Appellants' arguments that GETCHIUS does not disclose determining a measure of authoritativeness of a candidate document for a business at a location, the Examiner alleges GETCHIUS discloses that a subset of query terms is determined such that a string representing a particular query is uniquely mapped to a name corresponding to a data set (Examiner's Answer, p, 28). Thus, it appears that the Examiner is alleging that *uniquely mapping a query string to a name in a data set*, as disclosed by GETCHIUS, corresponds to determining a measure of authoritativeness of a candidate document for a business at a location, as recited in claim 1.

Appellants' Response to Point (5):

Appellants respectfully disagree with the Examiner's allegation.

A measure of authoritativeness of a particular document for a business at a location determines how authoritative the particular document is as a document for the business at the location. In other words, a document with a higher measure of authoritativeness is perceived as a more reliable document. Uniquely mapping a query

string to a name in a data set does not determine how authoritative the query string is as a query string for the name.

Furthermore, even if it is assumed, for the sake of argument, that "uniquely mapping a query string to a name in a data set" can be construed as determining a measure of authoritativeness for the query string (a point Appellants do not agree with), GETCHIUS and AGONI, whether taken alone or in any reasonable combination, do not disclose or suggest uniquely mapping a query string to a name in a data set based on a *candidate set of service providers*, as would be required by claim 1 based on the Examiner relying on the *candidate set of service providers* of AGONI as allegedly corresponding to the signals associated with the candidate documents, as recited in claim 1.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claims 1-3 and 12-13 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

2. Claim 4.

Examiner's Point (6):

Appellants' argued that GETCHIUS and AGONI do not disclose or suggest that identifying a set of documents further includes determining documents that are linked to by candidate documents, and identifying the determined document as candidate documents, as recited in claim 4.

In response, the Examiner additionally relies on col. 33, lines 25-53 of GETCHIUS, as well as Figs. 39 and 40 of GETCHIUS as allegedly disclosing the features of claim 4 (Examiner's Answer, p. 29).

Appellants' Response to Point (6):

Appellants submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 4.

Col. 33, lines 25-53 of GETCHIUS disclose:

> Once the system is initialized, the system may operate to obtain results that are to be displayed to the user. The steps for obtaining results may be seen in a flow chart 88 displayed in FIG. 41. Referring to FIG. 41, the parse driver 858 may at a step 20 parse a user query and deliver the parsed query in suitable form for handling by the query engine 862. The query engine may include the information retrieval software 908. At a step 22, the query engine 862 may operate the information retrieval software 908 to take the parsed user request and expand the query, turning the user request into a detailed query. Next, at a step 24, the information retrieval software may operate on the expanded term lists 836 by identifying documents associated with the terms identified in the expanded query. In an embodiment, the term lists 836 are the business listings described in connection with steps 8284 and 86 above, expanded to include synonyms and terms that are determined to be related to the words in the business listing. Identification of documents may be accomplished by a variety of information retrieval techniques. Documents may also be associated with queries by sorted relevancy ranking, clustering (automated grouping of related documents), automated document, summarization (creation of content abstracts, not simply the first few sentences of the document) and query-by-example (turning an individual document into a query in order to retrieve "more documents like this"). These functions may be accomplished by software techniques, such as having a table of pointers having as an argument a tokenized version of each possible term from the expanded user query from the step 22. The table of pointers may point to the location of a term list 836, for each such term. The term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the entire set of documents, the association of the document with other documents, the association of the document with categories, and the like.

This section of GETCHIUS discloses parsing and expanding a query and identifying documents associated with terms in the query. The terms lists are business listings and the document matching the terms of the query are sorted by relevancy ranking. This section of GETCHIUS further discloses that a table of pointers may point to the location of a term list and that the term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the set of documents, and the association of the document with other documents or categories.

A linked list is a data structure in which one record links to the next record in the sequence. This section of GETCHIUS does not disclose or suggest determining which records in the linked list link to any other records. Rather, this section of GETCHIUS merely discloses that a list of terms may be a linked list of documents that include the term. Furthermore, as the Examiner relies on the *query terms* of GETCHIUS as allegedly corresponding to the candidate document, as recited in claim 1, this section of GETCHIUS does not disclose or suggest determining documents in the linked list that are linked to *query terms*, as would be required by claim 4 based on the Examiner's interpretation of GETCHIUS. Therefore, this section of GETCHIUS does not disclose or suggest that identifying a set of documents further includes determining documents that are linked to by candidate documents, and identifying the determined document as candidate documents, as recited in claim 4.

Fig. 39 of GETCHIUS depicts a query engine coupled to a term list database and an advertisement banner term lists database. Fig. 40 of GETCHIUS depicts a flow graph for a process that includes accessing mark up language files, creating term lists, and expanding the term lists. Figs. 39 and 40 of GETCHIUS do not disclose or suggest that identifying a set of documents further includes determining documents that are linked to by candidate documents, and identifying the determined document as candidate documents, as recited in claim 4.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 4 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

3. Claim 6.

Examiner's Point (7):

Appellants argued that GETCHIUS does not disclose or suggest determining a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate documents that point to the candidate document, as recited in claim 6.

In response, the Examiner alleges that GETCHIUS discloses that each category is given a weight corresponding to the number of listings that are associated with the category and relies on col. 65, lines 10-25 and col. 33, lines 24-48 of GETCHIUS for allegedly disclosing the features recited in claim 6 (Examiner's Answer, p. 30).

Appellants' Response to Point (7):

Appellants submit that the Examiner's allegation does not address the features of claim 6.

A *weight given to a category*, as disclosed by GETCHIUS, cannot be reasonably held to correspond to an outlink from a document or a number of outlinks from a document, as recited in claim 6. An outlink is a hyperlink to another document. As demonstrated on pp. 17-19 of the Appeal Brief, none of the sections of GETCHIUS relied on by the Examiner in the rejection of claim 6 disclose, suggest, or even mention outlinks.

Appellants will further address the additional sections of GETCHIUS relied on by the Examiner.

Col. 65, lines 10-25 of GETCHIUS disclose:

> These statistics may be further improved by weighting other factors. For example, it is possible to weight each term that appears in one of the categories that is retrieved upon execution of a user query and to normalize the IDF and RTF statistics over the weights. Thus, if a particular category deserves a higher weight, then it might be accorded higher weight in ranking super-categories. For example, a category that is manually mapped to a super-category might be given a higher weight than a category that is automatically mapped. The user query might be given a higher or lower weight, than other information. Categories with a large number of listings may be given higher weight. In an embodiment, each category is given a weight corresponding to the number of listings that are associated with the category, normalized by dividing the total number of listings. In an embodiment, the user query terms are each given a weight of one. In the weighting process, the weight may be multiplied by the term element in performing the sum of the product of term frequency and inverse document frequency over all terms for all documents in the super-category linked list.

This section of GETCHIUS discloses weighing terms that appear in categories based on the number of listings in a category. A number of listings in a category cannot be reasonably held to correspond to a number of outlinks. As stated above, an outlink is a hyperlink to another document. This section of GETCHIUS does not disclose, suggest, or even mention a number of outlinks. Therefore, this section of GETCHIUS does not disclose or suggest determining a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate documents that point to the candidate document, as recited in claim 6.

Col. 33, lines 24-28 of GETCHIUS were reproduced above. This section of GETCHIUS discloses that a table of pointers may point to the location of a term list and that the term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the set of documents, and the association of the document with other document or categories.

This section of GETCHIUS does not disclose, suggest, or even mention a number of outlinks. Therefore, this section of GETCHIUS does not disclose or suggest determining a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate documents that point to the candidate document, as recited in claim 6.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 6 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

4. Claim 7.

Examiner's Point (8):

(8) Appellants argued that GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes identifying anchor text associated with links to the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location, as recited in claim 7.

In response, the Examiner additionally relies on col. 40, lines 25-37 and on col. 62, lines 49-60 of GETCHIUS for allegedly disclosing the features recited in claim 7 (Examiner's Answer, p. 31).

Appellants' Response to Point (8):

Appellants submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 7.

Col. 40, lines 25-37 of GETCHIUS disclose:

> At step 1008, the procedure "match phone number" is performed to produce a subset of one or more entries of the existing database which match the existing phone number. Control proceeds to step 1010 where the procedure "name match" is performed. Generally, "name match" will be described in paragraphs that follow to determine whether there is a business name match for a particular entry. Control proceeds to step 1012 where "derive score" is performed based on the zip code and the name match score. Generally, the result of step 1012 produces a score representing a statistic relative to determining whether two entries in a particular database and an updated version of the database match.

This section of GETCHIUS discloses a match phone number procedure, a name match procedure, and a derive score based on a zip code and the name match score. This section of GETCHIUS does not disclose, suggest, or even mention anchor text, let alone matching anchor text to a name of a business at a location. Anchor text is the visible, clickable text in a hyperlink. This section of GETCHIUS does not even mention hyperlinks. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes identifying anchor text associated with links to the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location, as recited in claim 7.

Col. 62, lines 49-60 of GETCHIUS disclose:

> Once control has returned to the flow chart 52 of FIG. 68, meaning that all yellow pages categories have been mapped to a super-category, at a step 77 the banner ad retrieval software 909 may index the various super-categories in a banner ad term list 837. The banner ad term list 837 may take the form of a linked list of the super-categories, with each element in the list consisting of all of the terms that appear in the super-category, as well as all of the terms that appear in each of the categories that was matched to the super-category.

This section of GETCHIUS discloses mapping yellow pages categories to a super-category and indexing the super-categories in a banner ad term list, in the form of a linked list. This section of GETCHIUS does not disclose, suggest, or even mention anchor text, let alone matching anchor text to a name of a business at a location. As stated above, anchor text is the visible, clickable text in a hyperlink. This section of GETCHIUS does not even mention hyperlinks. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes identifying anchor text associated with links to the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location, as recited in claim 7.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 7 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

5. Claim 8.

Examiner's Point (9):

Appellants argued that GETCHIUS does not disclose or suggest that wherein identifying signals associated with the candidate documents includes identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the

candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location.

In response, the Examiner additionally relies on col. 9, lines 37-43; Figs. 9 and 10; col. 47, lines 11-30; and Fig. 58 of GETCHIUS for allegedly disclosing the features of claim 8 (Examiner's Answer, pp. 32-33).

Appellants' Response to Point (9):

Appellants respectfully submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 8.

Col. 9, lines 37-43 of GETCHIUS, which describe Figs. 9 and 10 of GETCHIUS, disclose:

> Referring now to FIGS. 9 and 10, shown is one embodiment of a user interface for displaying a first page of the top query categories 1820. Generally, these categories are associated with the various business listings and are tags by which a user may perform queries. In this embodiment, for example, the user may select the "top categories" from the initial interface as included in the field 1802.

This section of GETCHIUS discloses a user interface for displaying a first page of top query categories that are associated with various business listings and are tags by which a user may perform queries. A user may select the top categories from the interface. This section of GETCHIUS does not disclose or suggest titles for documents, let alone determining whether a title of a document matches the name of a business. Therefore, this section of GETCHIUS does not disclose or suggest that wherein identifying signals associated with the candidate documents includes identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location, as recited in claim 8.

Col. 47, lines 11-30 of GETCHIUS, which describe Fig. 58 of GETCHIUS, disclose:

> Referring now to FIG. 58, shown is a flowchart of an embodiment of method steps for detecting duplicates in the category file. Generally, these steps are more detailed processing steps of step 1520 of FIG. 57. At step 1500, a first category name in the category file of the unfiltered database is tokenized. In other words, each word included in the heading or category name is associated with a token. Similarly, in step 1504, the next record of a category is examined and also tokenized. At step 1506, a comparison of the two tokenized names is performed to derive a score in accordance with the number of matching name components. This may also be normalized, as described in accordance with the foreign source update processing techniques. At step 1508, a determination is made as to whether or not the score is greater than a predetermined threshold. In this instance, the threshold is 75%. If the score is greater than the threshold, control proceeds to step 1512 where the categories are tagged as duplicates propagating any previous matching identifier tag. In other words, the transitive matching technique is used in marking matching categories. For example, if ID1,=ID2. Then, it is determined that ID2=ID5, ID5, is also marked as having ID1, as a matching identifier. Similarly, subsequent matches to ID5, f further propagate the value ID11. Subsequently, control proceeds to steps 1510 for advancement to the next record. If it is determined at step 1508 that the score is not greater than the threshold, no match is found and control proceeds to step 1510 where the next category is advanced to. At step 1514, a determination is made as to whether all the categories have been processed in the category file. If they have, control proceeds to step 1516 where processing stops. Otherwise, control proceeds to step 1504 for further comparisons and determinations of equivalent categories.

This section of GETCHIUS discloses a process that includes tokenizing a first category name and examining and tokenizing the next record of a category. The two tokenized records are compared and a score is derived and normalized. If the score is greater than a threshold, the categories are tagged as duplicates. If the score is not greater than the threshold, no match is found and control proceeds to the next category. This section of GETCHIUS does not disclose or suggest titles for documents, let alone determining whether a title of a document matches the name of a business. Therefore, this section of GETCHIUS does not disclose or suggest that wherein identifying signals associated with the candidate documents includes identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location, as recited in claim 8.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 8 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

6. Claim 10.

Examiner's Point (10):

Appellants argued that GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

In response, the Examiner additionally relies on col. 18, lines 18-26; col. 30, lines 62-67; and col. 43, lines 35-46 of GETCHIUS for allegedly disclosing the features of claim 8 (Examiner's Answer, pp. 33-34).

Appellants' Response to Point (10):

Appellants respectfully submit that these additional sections of GETCHIUS also do not disclose or suggest the features recited in claim 10.

Col. 18, lines 18-26 of GETCHIUS disclose:

> In this particular embodiment, the databases may include business information, such as for specific businesses or classifications of businesses. Additionally, data queries may be performed based on characteristics of the various businesses, such as location, name, or category. Furthermore, the architecture described herein supports a flexible presentation of these businesses, based on business agreements and service offerings.

This section of GETCHIUS discloses that the databases may include business information, such as specific businesses or classification of businesses. Data queries may be performed based on the characteristics of the various businesses, such as location,

name, or category. This section of GETCHIUS does not disclose or suggest increasing the authoritativeness of a candidate document when the candidate document is associated with a single location. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

Col. 30, lines 59-67 of GETCHIUS disclose:

> Referring now to FIGS. 37 and 38 shown is a flowchart of an embodiment of a method for integrating total-city and multi-city cache results into "normal" cached search results. At step 260, a total-city cache name corresponding to the data query is formed. In one embodiment, the total city cache name is formed by starting with the string "SCOPE=T" to identify a total-city name. Additionally, the following information is extracted from the original query string, as formed by the parser: category, category id, business name, street address, keywords, longitude, latitude.

This section of GETCHIUS discloses forming a total-city cache name corresponding to a data query and extracting, from the original query string, a category, category identification, a business name, a street address, keywords, longitude, and latitude. This section of GETCHIUS does not disclose or suggest determining whether a document is associated with a single location, let alone increasing a measure of authoritativeness based on whether the candidate document is associated with a single location. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of

authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

Col. 43, lines 35-46 of GETCHIUS disclose:

> Referring now to FIG. 50, shown as a flow chart of the steps of one embodiment for performing the routine "derive score", as performed from step 1012 of FIG. 46. Generally, code. At step 1080, the score previously derived from name match for each entry is updated by one if the zip codes of an existing database entry match an updated entry. At step 1082 this score is normalized by taking the score computed thus far and dividing it by the number of tokens in produced a normalized score as in step 1082. At step 1084, control returns to the point of call. In this particular instance, control returns to step 1012 where processing resumes with step 1020 of FIG. 47.

This section of GETCHIUS discloses that a score derived from a name match is updated if one of the zip codes of an existing database entry matches an updated entry, and that the score is normalized by dividing it by a number of tokens. This section of GETCHIUS does not disclose or suggest determining whether a document is associated with a single location, let alone increasing a measure of authoritativeness based on whether the candidate document is associated with a single location. Therefore, this section of GETCHIUS does not disclose or suggest that identifying signals associated with the candidate documents includes determining locations with which ones of the candidate documents are associated; and wherein determining a measure of authoritativeness of the candidate documents further includes increasing the measure of authoritativeness of one of the candidate documents based on whether the candidate document is associated with a single location, as recited in claim 10.

For at least the reasons given above and for those reasons given in the Appeal Brief, Appellants respectfully request that the rejection of claim 10 under 35 U.S.C. § 103(a) based on GETCHIUS and AGONI be reversed.

IV. CONCLUSION

In view of the foregoing arguments and those arguments presented in the Appeal Brief, Appellants respectfully solicit the Honorable Board to reverse the Examiner's rejections of claims 1-29.

To the extent necessary, a petition for an extension of time under 37 C.F.R. § 1.136 is hereby made. Please charge any shortage in fees due in connection with the filing of this paper, including extension of time fees, to Deposit Account No. 50-1070 and please credit any excess fees to such deposit account.

Respectfully submitted,

HARRITY & HARRITY, LLP

/Viktro Simkovic, Reg No. 56,012/
Viktor Simkovic
Reg. No. 56,012

Date: June 3, 2009
11350 Random Hills Road
Suite 600
Fairfax, Virginia 22030
(571) 432-0800 main
(571) 432-0899 direct

Customer No.: 44989

Electronic Acknowledgement Receipt

EFS ID:	5447480
Application Number:	11024967
International Application Number:	
Confirmation Number:	7261
Title of Invention:	Authoritative document identification
First Named Inventor/Applicant Name:	Daniel Egnor
Customer Number:	44989
Filer:	Viktor Simkovic/Sandra Stocklinski
Filer Authorized By:	Viktor Simkovic
Attorney Docket Number:	0026-0130
Receipt Date:	03-JUN-2009
Filing Date:	30-DEC-2004
Time Stamp:	16:27:31
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Reply Brief Filed	0026-0130_Reply_Brief_PTO.pdf	143919 7b41da2ea8cfe0af67e7279a3b8135d7c1d9157a	no	23

Warnings:	
Information:	

Total Files Size (in bytes):	143919

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PATENT
Attorney's Docket No. 0026-0130

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:	)	
	)	**Mail Stop APPEAL BRIEF - PATENTS**
Daniel EGNOR et al.	)	
	)	Group Art Unit: 2163
Application No.: 11/024,967	)	
	)	Examiner: S. Hwa
Filed: December 30, 2004	)	
	)	
For: AUTHORITATIVE DOCUMENT IDENTIFICATION	)	

PETITION FOR ONE MONTH EXTENSION OF TIME

U.S. Patent and Trademark Office
Customer Service Window, Mail Stop APPEAL BRIEF - PATENTS
Randolph Building
401 Dulany Street
Alexandria, VA 22314

Sir:

Applicants hereby respectfully petition for a one (1) month extension of time for the filing of a Reply Brief. Applicants require additional time to prepare an appropriate response to the Supplemental Examiner's Answer mailed on March 3, 2009 (i.e., to prepare a Reply Brief). The required fee is attached hereto:

☐ $65.00 ☒ $130.00

☐ An extension fee in the amount of $_____ is enclosed.

☒ Charge $130.00 to the credit card number provided.

The Commissioner is hereby authorized to charge any other appropriate fees that may be required by this paper that are not accounted for above, and to credit any overpayment, to Deposit Account No. 50-1070.

Respectfully submitted,

HARRITY & HARRITY, LLP

By: /Viktor Simkovic, Reg. No. 56,012/
Viktor Simkovic
Reg. No. 56,012

Date: May 4, 2009

11350 Random Hills Road
Suite 600
Fairfax, Virginia 22030
(571) 432-0800

Customer Number: 44989

Electronic Patent Application Fee Transmittal

Application Number:	11024967
Filing Date:	30-Dec-2004
Title of Invention:	Authoritative document identification
First Named Inventor/Applicant Name:	Daniel Egnor
Filer:	Viktor Simkovic/Brooke Fredrick
Attorney Docket Number:	0026-0130

Filed as Large Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Extension - 1 month with $0 paid	1251	1	130	130

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
			Total in USD ($)	**130**

Electronic Acknowledgement Receipt

EFS ID:	5271190
Application Number:	11024967
International Application Number:	
Confirmation Number:	7261
Title of Invention:	Authoritative document identification
First Named Inventor/Applicant Name:	Daniel Egnor
Customer Number:	44989
Filer:	Viktor Simkovic/Brooke Fredrick
Filer Authorized By:	Viktor Simkovic
Attorney Docket Number:	0026-0130
Receipt Date:	04-MAY-2009
Filing Date:	30-DEC-2004
Time Stamp:	17:34:39
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Credit Card
Payment was successfully received in RAM	$130
RAM confirmation Number	3752
Deposit Account	
Authorized User	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)

1	Extension of Time	0026-0130_Petition_for_1M_EOT.pdf	76896 e64e023bf035390be10558ae153268b57a8a5337	no	2
Warnings:					
Information:					
2	Fee Worksheet (PTO-875)	fee-info.pdf	29755 aaba3ceac17570769f07d733385ea86efffd5146	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	106651		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/024,967	12/30/2004	Daniel Egnor	0026-0130	7261

44989 7590 03/03/2009

HARRITY & HARRITY, LLP
11350 Random Hills Road
SUITE 600
FAIRFAX, VA 22030

EXAMINER	
HWA, SHYUE JIUNN	

ART UNIT	PAPER NUMBER
2163	

MAIL DATE	DELIVERY MODE
03/03/2009	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

UNITED STATES DEPARTMENT OF COMMERCE
U.S. Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450

APPLICATION NO./ CONTROL NO.	FILING DATE	FIRST NAMED INVENTOR / PATENT IN REEXAMINATION	ATTORNEY DOCKET NO.
11024967	12/30/2004	EGNOR ET AL.	0026-0130

HARRITY & HARRITY, LLP
11350 Random Hills Road
SUITE 600
FAIRFAX, VA 22030

EXAMINER	
JAMES HWA	
ART UNIT	**PAPER**
2163	20090203

DATE MAILED:

Please find below and/or attached an Office communication concerning this application or proceeding.

Commissioner for Patents

The New Ground of Rejection regarding 35 U.S.C. 101 is withdrawn from Examiner Answer on 12/24/2008.

/don wong/
Supervisory Patent Examiner, Art Unit 2163

UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

Application Number: 11/024,967
Filing Date: December 30, 2004
Appellant(s): EGNOR ET AL.

Viktor Simkovic
For Appellant

EXAMINER'S ANSWER

This is a supplemental Examiner’s Answer to the appeal brief filed 9/29/2008 appealing from the Office Action mailed 3/28/2008. This supplemental Examiner's Answer effectively vacates the Examiner's Answer mailed December 24, 2008.

(1) Real Party in Interest

A statement identifying by name the real party in interest is contained in the brief.

(2) Related Appeals and Interferences

The Examiner is not aware of any related appeals, interferences, or judicial proceedings, which directly affect or be affected by or have a bearing on the Board's decision in the pending appeal.

(3) Status of Claims

The statement of the status of claims contained in the brief is correct.

(4) Status of Amendments After Final

There are no unentered amendments.

(5) Summary of Claimed Subject Matter

The summary of claimed subject matter contained in the brief is correct.

(6) Grounds of Rejection to be Reviewed on Appeal

The appellant's statement of the grounds of rejection to be reviewed on appeal is correct.

(7) Claims Appendix

The copy of the appealed claims contained in the Appendix to the brief is correct.

(8) Evidence Relied Upon

US Patent No. 6,643,640 B1	Getchius	11/04/2003
US Patent Application No. 2002/0133374 A1	Agoni et al.	09/19/2002
US Patent Application No. 2004/0064334 A1	Nye	04/01/2004

(9) Grounds of Rejection

The following ground(s) of rejection are applicable to the appealed claims:

Claim Rejections - 35 USC § 103

The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

This application currently names joint inventors. In considering patentability of the claims under 35 U.S.C. 103(a), the examiner presumes that the subject matter of the various claims was commonly owned at the time any inventions covered therein were made absent any evidence to the contrary. Applicant is advised of the obligation under 37 CFR 1.56 to point out the inventor and invention dates of each claim that was not commonly owned at the time a later invention was made in order for the examiner to consider the applicability of 35 U.S.C. 103(c) and potential 35 U.S.C. 102(e), (f) or (g) prior art under 35 U.S.C. 103(a).

3. Claims 1-4, 6-8, 10, 12-18, 20-22, 24 and 26-28 are rejected under 35 U.S.C. 103(a) as being unpatentable over Getchius (US Patent No. 6,643,640 B1, hereinafter “Getchius”) in view of Agoni et al. (US Patent Application No. 2002/0133374 A1, hereinafter “Agoni”).

As to claim 1, Getchius teaches the claimed limitations:

“A method” as a method for performing data query caching in a computer system (column 1, lines 58-59).

"Identifying a set of documents, as candidate documents, that are all associated with a same geographic location" as a query for a particular location in the city field or the state field, or for a business name in the business name field, the information retrieval software retrieves documents from the term lists that correspond to a ranking of an expansion of the user-entered query (column 36, lines 62-67).

All searchable fields have a tag, such as a business name or city. Identifiers are generally produced by the information retrieval software, for example, if the field zip code includes a tag as included in the mark-up language file which indicates that this particular field is searchable, it may be desired that whenever a user wishes to do a search for zip code what is actually retrieved or displayed to the user is the city and the state (column 14, lines 2-16).

"Determining a measure of authoritativeness of the candidate documents for a business at the location based on the signals" as a subset of query terms is determined such that a string representing a particular query is uniquely mapped to a name corresponding to a data set. The subset of keys that are used in mapping a string corresponding to a query to a name of a data set include: Proximity, City, State, Street, Zip, Category, Category Identifier, Business name, Area code, Phone number, Keywords, and National Account (column 28, lines 7-11). It may be desirable to use a logarithm or related measure of the term frequency and the inverse document frequency, rather than the raw numbers (column 64, lines 52-55).

Data queries may be performed based on characteristics of the various businesses, such as location, name, or category. Furthermore, the architecture

described herein supports a flexible presentation of these businesses, based on business agreements and service offering (column 18, lines 21-26).

"Processing the candidate documents based on their authoritativeness for the location" as a determination is made that such a data set exists in the multi-city cache, control proceeds to step where the data is moved to the hot cache, if not all ready located there. Additionally, a reference to this location in the hot cache is saved for use in later processing steps (column 31, lines 42-47).

Getchius does not explicitly teach the claimed limitation "identifying signals associated with the candidate documents"

Agoni teaches the case control panel identifies the client name for each case. The case control panel also displays case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius and Agoni before him/her, to modify Getchius identifying signals associated with the candidate documents because that would allow cross-referencing of other tables and facilitate speedy search as taught by Agoni (page 15, paragraph 0134).

As to claim 2, Getchius teaches the claimed limitations:

"Analyzing documents in a document corpus to identify snippets of text that include information associated with the location, and identifying documents that include

the snippets of text as candidate documents" as (column 14, lines 27-30) and (column 43, lines 30-34).

As to claims 3 and 17, Getchius teaches the claimed limitations:

"The information associated with the location includes at least one of a full or partial address of the location, a full or partial telephone number associated with the location, or a full or partial name of business at the location" as (column 15, lines 1-5).

As to claims 4 and 18, Getchius teaches the claimed limitations:

"Determining documents that are linked to by the candidate documents, and identifying the determined documents as candidate documents" as (column 32, lines 55-58).

As to claims 6 and 20, Getchius teaches the claimed limitations:

"Identifying a number of outlinks from ones of the candidate documents that point to other ones of the candidate documents; determining a measure of authoritativeness of the candidate documents includes: generating an authoritative score for one of the candidate documents based on the number of outlinks from other ones of the candidate documents that point to the candidate document" as (column 32, lines 55-58), (column 64, lines 52-55) and (column 40, lines 28-37).

As to claims 7 and 21, Getchius teaches the claimed limitations:

"Identifying anchor text associated with links to the candidate documents" as (column 33, lines 46-49).

"Determining a measure of authoritativeness of the candidate documents" as (column 64, lines 52-55).

"Generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location" as (column 40, lines 28-37).

As to claims 8 and 22, Getchius teaches the claimed limitations:

"Identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents; generating an authoritative score for one of the candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location" as category (e.g. a title) represents a classification as associated with each business, such as representing a type of business service (column 28, lines 20-22), (column 14, lines 1-9), (column 64, lines 52-55) and (column 40, lines 28-37).

As to claims 10 and 24, Getchius teaches the claimed limitations:

"Wherein determining a measure of authoritativeness of the candidate documents further includes: increasing the measure of authoritativeness of one of the

candidate documents based one whether the candidate document is associated with a single location" as (column 64, lines 36-43).

As to claims 12 and 26, Getchius teaches the claimed limitations:

"Wherein the signals are associated with a plurality of different types of data associated with the candidate documents; wherein the method further comprises: weighting the different types of data; Combining the weighted data for ones of the candidate documents; assigning authoritative scores to the ones of the candidate documents based on the combined, weighted data" as (column 14, lines 26-30), (column 4, line 65 to column 5, line 2) and (column 40, lines 34-37).

As to claims 13 and 27, Getchius teaches the claimed limitations:

"Ranking one of the candidate documents based on the authoritative score for the one of the candidate documents" as (column 31, lines 60-65) and (column 33, lines 34-39).

As to claim 14, Getchius teaches the claimed limitations:

"A system" as a single user system performing data queries and searches upon a local database (column 5, lines 21-23).

"Means for identifying a set of documents, as candidate documents, that are associated with a business" as category represents a classification as associated with

each business, such as representing a type of business service (column 28, lines 20-22).

Retrievable as used herein generally means fields or categories with associated data that may be retrieved. All searchable fields have a tag, such as a business name or city. Identifiers are generally produced by the information retrieval software, in this particular embodiment, produces term lists in which there exists a list for each particular key word, term or category followed by a chain of identifiers that indicate the record number in the denormalized data store (column 14, lines 1-9).

"Means for determining a measure of authoritativeness of the candidate documents for the business based on the signals" as the subset of keys that are used in mapping a string corresponding to a query to a name of a data set include: Proximity, City, State, Street, Zip, Category, Category Identifier, Business name, Area code, Phone number, Keywords, and National Account (column 28, lines 7-11). It may be desirable to use a logarithm or related measure of the term frequency and the inverse document frequency, rather than the raw numbers (column 64, lines 52-55).

Getchius does not explicitly teach the claimed limitation "means for identifying a plurality of signals associated with each of the candidate documents".

Agoni teaches the case control panel identifies the client name for each case. The case control panel also displays case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius and Agoni before him/her, to modify Getchius identifying a plurality of signals associated with each of the candidate documents because that would allow cross-referencing of other tables and facilitate speedy search as taught by Agoni (page 15, paragraph 0134).

As to claim 15, Getchius teaches the claimed limitations:

"A system" as a single user system performing data queries and searches upon a local database (column 5, lines 21-23).

"A memory to store instructions" as in the memory of a data query cache (column 7, lines 30-31). A browser with additional embedded processing instructions (column 15, lines 49-50).

"A processor to execute the instructions in the memory" as the parsing results may be stored in the PHTML execution tree accessed by the particular processor (column 11, lines 65-67).

"Determine documents that are all associated with a same geographic location" as a query for a particular location in the city field or the state field, or for a business name in the business name field, the information retrieval software retrieves documents from the term lists that correspond to a ranking of an expansion of the user-entered query (column 36, lines 62-67).

"Assign authoritative scores to the documents based on the signals, the authoritative scores indicating measures of authoritativeness of the documents for a

business at the location, and process the documents based on the authoritative scores" as a score is computed for each name comparison of the existing database entry with a record of the updated version of the database. The score is computed as one point per matching component (column 43, lines 22-26). It may be desirable to use a logarithm or related measure of the term frequency and the inverse document frequency, rather than the raw numbers (column 64, lines 52-55).

Control proceeds to step where derive score is performed based on the zip code and the name match score. Generally, the result produces a score representing a statistic relative to determining whether two entries in a particular database and an updated version of the database match (column 40, lines 32-37).

Getchius does not explicitly teach the claimed limitation "identify a plurality of signals associated with each of the documents"

Agoni teaches the case control panel identifies the client name for each case. The case control panel also displays case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius and Agoni before him/her, to modify Getchius identify a plurality of signals associated with each of the documents because that would allow cross-referencing of other tables and facilitate speedy search as taught by Agoni (page 15, paragraph 0134).

As to claim 16, Getchius teaches the claimed limitations:

"When determining documents, the processor is configured to analyze documents in a document corpus to detect documents that include snippets of text with information associated with the location" as (column 28, lines 7-11).

As to claim 28, Getchius teaches the claimed limitations:

The limitations therein have substantially the same scope as claim 1. In addition, Getchius teaches a computer-readable medium that stores instructions that when executed by a computer system causes the computer system to perform a method for processing a data query (claim 23 of Getchius). Therefore this claim is rejected for at least the same reasons as claim 1.

4. Claims 5, 9, 11, 19, 23 and 25 are rejected under 35 U.S.C. 103(a) as being unpatentable over Getchius (US Patent No. 6,643,640 B1) as applied to claims 1 and 15 above, and further in view of Agoni et al. (US Patent Application No. 2002/0133374 A1) and Nye (US Patent Application No. 2004/0064334 A1, hereinafter "Nye").

As to claims 5 and 19, Getchius teaches the claimed limitations:

Getchius does not explicitly teach the claimed limitation "Determining additional documents by stripping portions of addresses of the candidate documents, and identifying the additional documents as candidate documents".

Nye teaches a geographic search on the authenticated digital certificate database is performed using the possibly modified keyword query. Both term

removal/stripping process and the term enhancement process are optional based (page 9, paragraph 0081; see also element 806 of figure 8).

The geographic location of the electronic documents is determined by a specialized keyword-parsing algorithm designed to identify addresses within electronic documents. On the World Wide Web these electronic documents are referenced by electronic document addresses in the form of a Uniform Resource Locator (URL). Yet another approach is the creation of localized portals and city pages that restrict their content to that associated with a specific, predetermined, geographic area (page 1, paragraph 0005-0006).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius, Agoni and Nye before him/her, to modify Getchius stripping portions of addresses of the candidate documents because that would allow users to restrict searches to a preselected geographic region as taught by Nye (page 1, paragraph 0006).

As to claims 9 and 23, Getchius does not explicitly teach the claimed limitation "identifying domain names associated with ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents; generating an authoritative score for one of the candidate documents based on whether a domain name associated with the candidate document matches all or part of a name of the business at the location".

Nye teaches in the case of city pages, the domains are different from city to city, so there is no way to know the URL of the city page if you are visiting another city without asking someone or using a search engine (page 2, paragraph 0012)

The various types of authentications can be related to electronic documents stored in a directory/domain name structure. A type-1 authentication indicates a primary domain where whois.org information matches website and the geographic location and business name are verified from two independent sources (page 8, paragraph 0072).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius, Agoni and Nye before him/her, to modify Getchius a domain name associated with the candidate document matches all or part of the business name because that would allow users to restrict searches to a preselected geographic region as taught by Nye (page 1, paragraph 0006).

As to claims 11 and 25, Getchius teaches the claimed limitations:

"A number of outlinks from ones of the candidate documents that point to another one of the candidate documents, anchor text associated with links that point to ones of the candidate documents that matches all or part of a name of the business at the location, titles of ones of the candidate documents that match all or part of the business name, or domain names associated with ones of the candidate documents that match all or part of the business name" as (column 33, lines 24-48), (column 33, lines 33-47), (column 37, lines 1-10), (column 28, lines 20-22) and (column 14, lines 1-9).

Getchius does not explicitly teach the claimed limitation "domain names associated with ones of the candidate documents that match all or part of the business name".

Nye teaches in the case of city pages, the domains are different from city to city, so there is no way to know the URL of the city page if you are visiting another city without asking someone or using a search engine (page 2, paragraph 0012).

The various types of authentications can be related to electronic documents stored in a directory/domain name structure. A type-1 authentication indicates a primary domain where whois.org information matches website and the geographic location and business name are verified from two independent sources (page 8, paragraph 0072).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius, Agoni and Nye before him/her, to modify Getchius a domain name associated with the candidate document matches all or part of the business name because that would allow users to restrict searches to a preselected geographic region as taught by Nye (page 1, paragraph 0006).

5. Claim 29 are rejected under 35 U.S.C. 103(a) as being unpatentable over Getchius (US Patent No. 6,643,640 B1) in view of Nye (US Patent Application No. 2004/0064334 A1).

As to claim 29, Getchius teaches the claimed limitations:

"A method" as a method for performing data query caching in a computer system (column 1, lines 58-59).

"Identifying a set of documents, as candidate documents, that are associated with a same geographic location" as a series of steps takes place if the user enters a query for a particular location in the city field or the state field or for a business name in the business name field, the information retrieval software retrieves documents from the term lists that correspond to a ranking of an expansion of the user-entered query (column 36, lines 62-67).

All searchable fields have a tag, such as a business name or city. Identifiers are generally produced by the information retrieval software, for example, if the field zip code includes a tag as included in the mark-up language file which indicates that this particular field is searchable, it may be desired that whenever a user wishes to do a search for zip code what is actually retrieved or displayed to the user is the city and the state (column 14, lines 2-16).

"Identifying, for each of the candidate documents, a first signal based on a number of outlinks from one or more of the candidate documents that point to the candidate document" as the information retrieval software may operate on the expanded term lists by identifying documents associated with the terms identified in the expanded query. The term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the entire

set of documents, the association of the document with other documents, the association of the document with categories, and the like (column 33, lines 24-48).

"Identifying, for each of the candidate documents, a second signal based on whether there is anchor text, which matches all or part of a name of a business associated with the location, associated with a link that points to the candidate document" as identification of documents may be accomplished by a variety of information retrieval techniques. Documents may also be associated with queries by sorted relevancy ranking, clustering, automated document, and summarization and query-by-example. The table of pointers may point to the location of a term list, for each such term. The term list may be a linked list of documents that include the term (column 33, lines 33-47).

When both a category and a location or a business name, or all three, are entered by the user, then the information retrieval software may retrieve term lists that correspond to each of the terms of the query, such as a list corresponding to the category restaurant (e.g. anchor text) and a list corresponding to the city field Boston. The information retrieval software could then perform an intersection of the two sets and perform a ranking of the related categories (column 37, lines 1-10).

"Identifying, for each of the candidate documents, a third signal based on whether a title of the candidate document matches all or part of the business name" as category (e.g. a title) represents a classification as associated with each business, such as representing a type of business service (column 28, lines 20-22).

Retrievable as used herein generally means fields or categories with associated data that may be retrieved. All searchable fields have a tag, such as a business name or city. Identifiers are generally produced by the information retrieval software, in this particular embodiment, produces term lists in which there exists a list for each particular key word, term or category followed by a chain of identifiers that indicate the record number in the denormalized data store (column 14, lines 1-9).

"Identifying, for each of the candidate documents, a fifth signal based on whether the candidate document is associated with a single location" as a subset of query terms is determined such that a string representing a particular query is uniquely mapped to a name corresponding to a data set. The subset of keys that are used in mapping a string corresponding to a query to a name of a data set include: Proximity, City, State, Street, Zip, Category, Category Identifier, Business name, Area code, Phone number, Keywords, and National Account (column 28, lines 7-11).

"Combining the first, second, third, fourth, and fifth signals to identify a score for each of the candidate documents" as the combined search results are sorted such that any redundant listings are removed, any additional processing is performed, as in accordance with the user query, for example, as producing the listings which begin with B, or only listing the top ranked fifteen (15) listings as ranked in accordance with other user specified criteria (column 31, lines 60-65).

Control proceeds to step where derive score is performed based on the zip code and the name match score. Generally, the result produces a score representing a

statistic relative to determining whether two entries in a particular database and an updated version of the database match (column 40, lines 32-37).

"Processing the candidate documents based on the scores" as if there is more than one matching entry in the database for a record in the current updated version of the database, control proceeds to where a determination is made whether there is only one entry with a maximum score. If there are multiple entries with the same maximum score, control proceeds to where additional processing is required to determine which is the matching entry, if any (column 40, lines 47-58).

Getchius does not explicitly teach the claimed limitation "Identifying determining, for each of the candidate documents, a fourth signal based on whether a domain name associated with the candidate document matches all or part of the business name".

Nye teaches in the case of city pages, the domains are different from city to city, so there is no way to know the URL of the city page if you are visiting another city without asking someone or using a search engine (page 2, paragraph 0012)

The various types of authentications can be related to electronic documents stored in a directory/domain name structure. A type-1 authentication indicates a primary domain where whois.org information matches website and the geographic location and business name are verified from two independent sources (page 8, paragraph 0072).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius and Nye before him/her, to modify Getchius a domain name associated with the candidate document matches all

or part of the business name because that would allow users to restrict searches to a preselected geographic region as taught by Nye (page 1, paragraph 0006).

(10) Response to Argument

Appellant Argues:

(1) In rejecting a claim under 35 U.S.C. § 103, the Examiner must provide a factual basis to support the conclusion of obviousness.

Examiner Responds:

Applicant argued that it would not be obvious to combine the two references and no motivation to combine.

In response to applicant's argument, The examiner respectfully submits that to establish a prima facie case of obviousness under 35 USC 103, references must provide motivation or suggestion either in the references themselves, or in the knowledge generally available to one of ordinary skill in the art; must be analogous; and must teach all the claimed limitations.

In this case, the instant application is concerned to a system determines documents that are associated with a location, identifies a group of signals associated with each of the documents, and determines authoritativeness of the documents for the location based on the signals.

As discussed in the office action, Getchius provides data queries may be performed based on characteristics of the various businesses, such as location, name, or category (column 18, lines 29-45).

Similarly, Agoni teaches a search engine responsive to search criteria to search the profile data for portions of the profile data at least approximately matching the search criteria and to generate result data identifying service providers corresponding to the at least approximately matched portions of the profile data (page 3, paragraph 0018).

Importantly, Agoni provides an search engine searching the profile data for the profile criteria and automatically generating a candidate selection prompt for a client to select the service provider as a candidate for providing services to the client, the candidate selection prompt communicating a representation of the identification data identifying the service provider (page 2, paragraph 0015).

As discussed above, a person of an ordinary skill in the art at the time the invention was made would recognize the advantage of Agoni to add the Agoni's teaching of the case control panel identifies the client name for each case. The case control panel also displays case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144), because that would allow cross-referencing of other tables and facilitate speedy search (page 15, paragraph 0134) and genuinely improve and enhance the quality of service rendered by the professional and received by the client (page 2, paragraph 0012) as taught by Agoni.

Therefore, the 103 rejection for claims is proper and make the record clear.

Appellant Argues:

(2) Getchius does not disclose or suggest candidate query terms for a business (that were identified as a set of query terms associated with the same geographical location), as would be required by claim 1 (appeal page 9).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the databases may include business information, such as for specific businesses or classifications of businesses. Additionally, data queries may be performed based on characteristics of the various businesses, such as location, name, or category (column 18, lines 14-25). The data query cache may include different types of cached geographical data as may be used in performing different data queries. For example, the type of data cached described in the prior paragraphs is the normal business listing data as associated with a well-defined geographic area (column 30, lines 12-20).

FIG. 34, shown an example determined and apply the best derivation sequence. In this example, the query is for MA AND RESTAURANTS AND FLOWERSHOPS. As represented, it has been determined that MA is the starting data set (e.g. same geographical location in state MA) which is located in the data query cache. In this example, the parentage has been extended to grandparents, and MA has been determined to be the first ranking data set in terms of parentage and number of listings in the data set (column 26, lines 27-40). Other techniques for weighting may also be-used. For example, if a term is a high frequency term, it may not make much difference

in logical significance whether the term occurs, for example, one thousand times, in the search, or whether the term occurs one million times. In order to collapse the significance of such high frequency terms, it may be desirable to use a logarithm or related measure of the term frequency and the inverse document frequency, rather than the raw numbers (column 64, lines 47-56).

Appellant Argues:

(3) Getchius does not disclose or suggest candidate documents for a business. Terms of a query cannot be reasonably held to be equivalent to candidate documents for a business, and categories of business listings cannot be reasonably held to be equivalent to a business (appeal page 10).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches various business listings may be grouped together in categories. In this example, relating to shoes, are business listings included in thirteen categories. From this listing of thirteen categories, the user may select one of these relating to shoes. For example, selection, as by using a mouse of custom made shoes (column 10, lines 1-11; see also element 1862 of figure 14 and figure 22).

Each business listing may be represented as a document stored in the primary and secondary databases. The documents may be manipulated as generic objects. As more detail, representing each business listing as a generic object facilitates subsequent handling of the business listings (column 19, lines 61-67). Relevance

information is Verity-specific information as it relates to the query. For example, this generally represents the frequency of words or terms in a document. The advertiser priority indicates a service level that may be used in presenting business listings (column 29, lines 21-55; see also figure 36).

Appellant Argues:

(4) Agoni does not disclose or suggest "identifying signals associated with the candidate documents," as recited in claim 1 (appeal page 12).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the markup language files include one file or document per business for which there is an advertisement, for example, in this particular embodiment. Each of the markup language files includes markup language statements, such as SGML-like statements, with tags identifying key data items (e.g. signals) in the document for each business (column 13, line 54 to column 14, lines 17). The query engine may operate the information retrieval software to take the parsed user request and expand the query, turning the user request into a detailed query. Next, the information retrieval software may operate on the expanded term lists by identifying documents associated with the terms identified in the expanded query. An indexing architecture such as that provided by Verity allows for incremental indexing, so that only new, updated or deleted documents require changes, avoiding the need for a complete re-index each time a document changes. Online identifiers may be

provided, so that searches can continue while the identifiers are modified. This function is also provided by the Verity software (column 33, lines 23-61; see also figure 41).

A subset of query terms is determined such that a string representing a particular query is uniquely mapped to a name corresponding to a data set. The subset of keys that are used in mapping a string corresponding to a query to a name of a data set include: Proximity, City, State, Street, Zip, Category, Category Identifier, Business name, Area code, Phone number, Keywords (column 28, lines 7-11). It may be desirable to use a logarithm or related measure of the term frequency and the inverse document frequency, rather than the raw numbers (column 64, lines 52-55). Data queries may be performed based on characteristics of the various businesses, such as location, name, or category (column 18, lines 21-26).

Getchius does not explicitly teach the claimed limitation “the candidate documents”

Agoni teaches a communication module making available to a client the result data, the communication module receiving candidate data representing a candidate set of service providers comprising one or more of the service providers identified by the result data, the communication module receiving and storing the service summary data representing needed services and making the service summary data available to each of the candidate set of service providers (page 3, paragraph 0018). The case control panel identifies the client name for each case. The case control panel also displays case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius and Agoni before him/her, to modify Getchius identifying signals associated with the candidate documents because that would allow cross-referencing of other tables and facilitate speedy search as taught by Agoni (page 15, paragraph 0134).

Appellant Argues:

(5) there is no act of determining how much authoritativeness the one thing has with respect to the other thing in such a mapping in claim 1 (appeal page 13).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches steps for forming a name associated with a data set, as may be stored in the data query cache or page cache. A subset of query terms is determined such that a string representing a particular query is uniquely mapped to a name corresponding to a data set. In this embodiment, the subset of keys that are used in mapping a string corresponding to a query to a name of a data set include: Proximity, City, State, Street, Zip, Category, Category Identifier, Business name, Area code, Phone number, Keywords, and National Account (column 28, lines 4-28; see also figure 35).

Appellant Argues:

(6) Getchius and Agoni, whether taken alone or in any reasonable combination, do not disclose or suggest the combination of features of “determining documents that

are linked to by the candidate documents, and identifying the determined documents as candidate documents" in claim 4 (appeal page 15).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the information retrieval software may operate on the expanded term lists by identifying documents associated with the terms identified in the expanded query. In an embodiment, the term lists are the business listings, expanded to include synonyms and terms that are determined to be related to the words in the business listing. The term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the entire set of documents, the association of the document with other documents, the association of the document with categories, and the like (column 33, lines 25-53; see also figures 39, 40).

Getchius does not explicitly teach the claimed limitation "the candidate documents"

Agoni teaches a communication module making available to a client the result data, the communication module receiving candidate data representing a candidate set of service providers comprising one or more of the service providers identified by the result data, the communication module receiving and storing the service summary data representing needed services and making the service summary data available to each of the candidate set of service providers (page 3, paragraph 0018). The case control panel identifies the client name for each case. The case control panel also displays

case ID, number of days since last communication, and also whether there are any unread message (e.g. signal) associated with the case (page 16, paragraph 0144).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made, having the teachings of Getchius and Agoni before him/her, to modify Getchius identifying signals associated with the candidate documents because that would allow cross-referencing of other tables and facilitate speedy search as taught by Agoni (page 15, paragraph 0134).

Appellant Argues:

(7) Getchius does not disclose or suggest determining a number of outlinks from terms of a search query in claim 6 (appeal page 17).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the user query might be given a higher or lower weight, than other information. Categories with a large number of listings may be given higher weight. In an embodiment, each category is given a weight corresponding to the number of listings that are associated with the category, normalized by dividing the total number of listings. In an embodiment, the user query terms are each given a weight of one. In the weighting process, the weight may be multiplied by the term element in performing the sum of the product of term frequency and inverse document frequency over all terms for all documents in the super-category linked list (column 65, lines 10-25). The information retrieval software may operate on the expanded term lists by identifying documents associated with the terms identified in

the expanded query. The term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the entire set of documents, the association of the document with other documents, the association of the document with categories, and the like (column 33, lines 24-48).

Appellant Argues:

(8) Getchius and Agoni, whether taken alone or in any reasonable combination, do not disclose or suggest the combination of "determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether the candidate document is pointed to by one or more links whose anchor text matches all or part of a name of the business at the location" in claim 7 (appeal page 19).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the procedure Match Phone Number is performed to produce a subset of one or more entries of the existing database which match the existing phone number. Control proceeds to step where the procedure Name Match is performed. Generally, Name Match will be described in paragraphs that follow to determine whether there is a business name match for a particular entry. Control proceeds to step where Derive Score is performed based on the zip code and the name match Score. Generally, the result produces a score

representing a statistic relative to determining whether two entries in a particular database and an updated version of the database match (column 40, lines 25-37).

Once control has returned that all yellow pages categories have been mapped to a super-category, at a step the banner ad retrieval software may index the various super-categories in a banner ad term list. The banner ad term list may take the form of a linked list of the super-categories, with each element in the list consisting of all of the terms that appear in the super-category, as well as all of the terms that appear in each of the categories that was matched to the super-category (column 62, lines 49-60).

Appellant Argues:

(9) Getchius and Agoni do not disclose or suggest the combination of features "identifying titles of ones of the candidate documents; and wherein determining a measure of authoritativeness of the candidate documents includes generating an authoritative score for one of the candidate documents based on whether a title associated with the candidate document matches all or part of a name of the business at the location" in claim 8 (appeal page 21).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches one embodiment of a user interface for displaying a first page of the top query categories. Generally, these categories are associated with the various business listings and are tags (e.g. title) by which a user may perform queries (column 9, lines 37-43; see also figures 9 and 10).

A first category name (e.g. title) in the category file of the unfiltered database is tokenized. In other words, each word included in the heading or category name (e.g. title) is associated with a token. Similarly, the next record of a category is examined and also tokenized. A comparison of the two tokenized names is performed to derive a score in accordance with the number of matching name components. This may also be normalized, as described in accordance with the foreign source update processing techniques. A determination is made as to whether or not the score is greater than a predetermined threshold. If the score is greater than the threshold, control proceeds to step where the categories are tagged as duplicates propagating any previous matching identifier tag (e.g. title). In other words, the transitive matching technique is used in marking matching categories (column 47, lines 11-30; see also figure 58).

Appellant Argues:

(10) Getchius and Agoni do not disclose or suggest increasing a measure of authoritativeness of one of the candidate documents based one whether the candidate document is associated with a single location, as recited in claim 10 (appeal page 24).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the databases may include business information, such as for specific businesses or classifications of businesses. Additionally, data queries may be performed based on characteristics of the various businesses, such as location, name, or category (column 18, lines 18-26). A similar series of steps takes place if the user enters a query for a particular location in the city

field or the state field, or for a business name in the business name field. The information retrieval software retrieves documents from the term lists that correspond to a ranking of an expansion of the user-entered query (column 30, lines 62-67). A flow chart of the steps of one embodiment for performing the routine Derive Score, as performed from step 1012 of figure 46. Generally, at step 1080 of figure 50, the score previously derived from name match for each entry is updated by one if the zip codes (e.g. geographical location) of an existing database entry match an updated entry. At step 1082 this score is normalized by taking the score computed thus far and dividing it by the number of tokens in produced a normalized score as in step 1082 (column 43, lines 35-46).

Also, Agoni teaches the CaseSmart search engine compares an attorney profile record with a search condition. If an attorney profile record matches and/or at least partially matches a search condition, then, the attorney's search score is increased by one unit number multiplied by the respective weight coefficient (page 10, paragraph 0092).

Appellant Argues:

(11) Getchius and Agoni do not disclose or suggest means for determining a measure of authoritativeness of candidate documents for a business (with which the candidate documents are all associated) based on signals associated with each of the candidate documents, as recited in claim 14 (appeal page 26).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches data queries may be performed based on characteristics of the various businesses, such as location, name, or category. Furthermore, the architecture supports a flexible presentation of these businesses, based on business agreements and service offering (column 18, lines 21-26). A subset of query terms is determined such that a string representing a particular query is uniquely mapped to a name corresponding to a data set. The subset of keys that are used in mapping a string corresponding to a query to a name of a data set include: Proximity, City, State, Street, Zip, Category, Category Identifier, Business name, Area code, Phone number, Keywords (column 28, lines 7-11). It may be desirable to use a logarithm or related measure of the term frequency and the inverse document frequency, rather than the raw numbers (column 64, lines 52-55).

Each of the markup language files includes markup language statements, such as SGML-like statements, with tags identifying key data items (e.g. signals) in the document for each business (column 13, line 54 to column 14, lines 17). The query engine may operate the information retrieval software to take the parsed user request and expand the query, turning the user request into a detailed query. Online identifiers may be provided, so that searches can continue while the identifiers are modified. This function is also provided by the Verity software (column 33, lines 23-61; see also figure 41).

Appellant Argues:

(12) Getchius and Agoni do not disclose or suggest instructions to assign authoritative scores to documents (that are all associated with the same geographical location) based on signals associated with each of the documents, the authoritative scores indicating measures of authoritativeness of the documents for a business at that geographical location, as recited in claim 15 (appeal page 31).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches Name Match will be described in paragraphs that follow to determine whether there is a business name match for a particular entry. Control proceeds to step where Derive Score is performed based on the zip code (e.g. geographical location) and the name match score. Generally, the result produces a score representing a statistic relative to determining whether two entries in a particular database and an updated version of the database match (column 40, lines 32-37; see also element 1012 of figure 46). A score is computed for each name comparison of the existing database entry with a record of the updated version of the database. The score is computed as one point per matching component (column 43, lines 22-26). The ranking may be performed by a variety of techniques. One such technique obtains a number for each term that appears in the user query and in the categories that consists of the product of the term frequency for that term and the inverse document frequency for that term. The sum of all the resulting numbers may be calculated for all super-categories, and the super-category with the

highest sum may be the highest ranked document. It may be desirable to use a logarithm or related measure of the term frequency and the inverse document frequency, rather than the raw numbers (column 64, lines 36-55).

Also, Agoni teaches the client is prompted to identify a weight coefficient for each entered search condition to indicate the importance of the respective search condition. In one embodiment, the client assigns a numeric value or a percentage value to each search condition as a weight coefficient. In another embodiment, the client selects a value from a range of values such as High Importance, Medium Importance and Low Importance to attach to each search condition, with each of such values having a predefined numeric value as a weight coefficient. In yet another embodiment, the client arranges the order of the entered search conditions, with a weight coefficient assigned to each search condition depending on its arranged order (page 10, paragraph 0091).

Appellant Argues:

(13) Getchius and Agoni do not disclose or suggest identifying documents that are linked to terms in a search query in claim 18 (appeal page 39).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the information retrieval software may produce term lists that are used by the information retrieval software to handle queries that are delivered to the query engine. The term lists may consist of a linked list for each term that appears in one of the business listings, with the elements of the linked list including a document identifier for the business listing and certain statistics

regarding the frequency of occurrence of the particular term in each document and in the document set as a whole (column 32, lines 55-67).

The assembling software on the receiver side integrates the data from temporary table into table. Additionally, a link is established in table to the data in table and the associated global identifier removed. Each entry in table is copied into table. In particular, the Id and Size fields are copied into table for identifiers. The integration software then uses the global Id obtained from temporary table to index into the repository in search for a matching global identifier entry. When a matching global identifier is found in table, the repository from table is copied into the blob pointer field of table (column 53, lines 16-30; see also figure 66).

Appellant Argues:

(14) Getchius does not disclose or suggest a number of outlinks from terms of a search query in claim 20 (appeal page 40).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the information retrieval software may operate on the expanded term lists by identifying documents associated with the terms identified in the expanded query. The term list may be a linked list of documents that include the term. The linked list may include information about each document, such as the number of occurrences of the term in the document, the inverse frequency of the term in the entire set of documents, the association of the document with other documents, the association of the document with categories, and the like

(column 33, lines 24-48). The user query might be given a higher or lower weight, than other information. Categories with a large number of listings may be given higher weight. In an embodiment, each category is given a weight corresponding to the number of listings that are associated with the category, normalized by dividing the total number of listings. In an embodiment, the user query terms are each given a weight of one. In the weighting process, the weight may be multiplied by the term element in performing the sum of the product of term frequency and inverse document frequency over all terms for all documents in the super-category linked list (column 65, lines 10-25).

Appellant Argues:

(15) Getchius and Agoni do not disclose or suggest identifying a plurality of signals, the processor is configured to identify anchor text associated with links to ones of the documents; and when assigning authoritative scores to the documents, the processor is configured to generate an authoritative score for one of the documents based on one or more links to the document whose anchor text matches all or part of a name of the business at the location in claim 21(appeal page 42).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the data transfer of identifiers; each identifier is associated with only blob data. It should be noted that this general technique and the data included in the data table may additionally include text data associated with each identifier or row in the table (column 50, lines 17-30). The

procedure Match Phone Number is performed to produce a subset of one or more entries of the existing database which match the existing phone number. Control proceeds to step where the procedure Name Match is performed. Generally, Name Match will be described in paragraphs that follow to determine whether there is a business name match for a particular entry. Control proceeds to step where Derive Score is performed based on the zip code and the name match Score. Generally, the result produces a score representing a statistic relative to determining whether two entries in a particular database and an updated version of the database match (column 40, lines 25-37).

Once control has returned that all yellow pages categories have been mapped to a super-category, at a step the banner ad retrieval software may index the various super-categories in a banner ad term list. The banner ad term list may take the form of a linked list of the super-categories, with each element in the list consisting of all of the terms that appear in the super-category, as well as all of the terms that appear in each of the categories that was matched to the super-category (column 62, lines 49-60).

Appellant Argues:

(16) Getchius and Agoni do not disclose or suggest “the processor is configured to identify titles of ones of the documents; and when assigning authoritative scores to the documents, the processor is configured to generate an authoritative score for one of the documents based on whether the document includes a title that matches all or part of a name of the business at the location” in claim 22 (appeal page 44).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches one embodiment of a user interface for displaying a first page of the top query categories. Generally, these categories are associated with the various business listings and are tags (e.g. title) by which a user may perform queries (column 9, lines 37-43; see also figures 9 and 10). All searchable fields have a tag (e.g. title), such as a business name or city. Identifiers are generally produced by the information retrieval software, in this particular embodiment, produces term lists in which there exists a list for each particular key word, term or category followed by a chain of identifiers that indicate the record number in the denormalized data store (column 14, lines 1-9).

Control proceeds to step where derive score is performed based on the zip code and the name match score. Generally, the result produces a score representing a statistic relative to determining whether two entries in a particular database and an updated version of the database match (column 40, lines 32-37). A first category name (e.g. title) in the category file of the unfiltered database is tokenized. In other words, each word included in the heading or category name (e.g. title) is associated with a token. Similarly, the next record of a category is examined and also tokenized. A comparison of the two tokenized names is performed to derive a score in accordance with the number of matching name components. This may also be normalized, as described in accordance with the foreign source update processing techniques. A determination is made as to whether or not the score is greater than a predetermined threshold. If the score is greater than the threshold, control proceeds to step where the

categories are tagged as duplicates propagating any previous matching identifier tag (e.g. title). In other words, the transitive matching technique is used in marking matching categories (column 47, lines 11-30; see also figure 58).

Appellant Argues:

(17) Getchius and Agoni do not disclose or suggest a processor configured to increase an authoritative score assigned to one of documents based one whether the document is associated with a single location, as recited in claim 24 (appeal page 47).

Examiner Responds:

Examiner respectfully disagrees. Getchius teaches the databases may include business information, such as for specific businesses or classifications of businesses. Additionally, data queries may be performed based on characteristics of the various businesses, such as location, name, or category (column 18, lines 18-26). A similar series of steps takes place if the user enters a query for a particular location in the city field or the state field, or for a business name in the business name field. The information retrieval software retrieves documents from the term lists that correspond to a ranking of an expansion of the user-entered query (column 30, lines 62-67).

A flow chart of the steps of one embodiment for performing the routine Derive Score, as performed from step 1012 of figure 46. Generally, at step 1080 of figure 50, the score previously derived from name match for each entry is updated by one if the zip codes (e.g. geographical location) of an existing database entry match an updated entry. At step 1082 this score is normalized by taking the score computed thus far and

dividing it by the number of tokens in produced a normalized score as in step 1082 (column 43, lines 35-46).

Also, Agoni teaches the CaseSmart search engine compares an attorney profile record with a search condition. If an attorney profile record matches and/or at least partially matches a search condition, then, the attorney's search score is increased by one unit number multiplied by the respective weight coefficient (page 10, paragraph 0092).

Appellant Argues:

(18) Getchius and Agoni do not disclose or suggest instructions for determining a measure of authoritativeness of documents for a business at a location with which all of the documents are associated based on signals associated with the documents, as recited in claim 28 (appeal page 49).

Examiner Responds:

Examiner respectfully disagrees. Getchius the markup language files include one file or document per business for which there is an advertisement, for example, in this particular embodiment. Each of the markup language files includes markup language statements, such as SGML-like statements, with tags identifying key data items in the document for each business (column 13, lines 55-65). FIG. 34, shown an example determined and apply the best derivation sequence. In this example, the query is for MA AND RESTAURANTS AND FLOWERSHOPS. As represented, it has been determined that MA is the starting data set (e.g. same geographical location in state MA) which is

located in the data query cache. In this example, the parentage has been extended to grandparents, and MA has been determined to be the first ranking data set in terms of parentage and number of listings in the data set (column 26, lines 27-40).

Appellant Argues:

(18) Getchius does not disclose or suggest "identifying signals associated with the candidate documents," as recited in claim 18 (appeal page 52).

Examiner Responds:

Examiner respectfully disagrees. This argument same as Applicant's Argues (4) in claim 1.

(11) Related Proceeding(s) Appendix

No decision rendered by a court or the Board is identified by the Examiner in the Related Appeals and Interferences section of this Examiner's Answer.

(12) Conclusion

For the above reasons, it is believed that the rejections should be sustained.

Respectfully submitted,

James Hwa

Conferees:

/James Hwa/
Examiner, Art Unit 2163

/don wong/
Supervisory Patent Examiner, Art Unit 2163

/Charles Rones/
Supervisory Patent Examiner, Art Unit 2164