(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 October 2011 (20.10.2011)

PCT

(10) International Publication Number
WO 2011/127655 A1

(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/CN2010/071758

(22) International Filing Date:
14 April 2010 (14.04.2010)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.** [US/US]; 11445 Compaq Center Drive West, Houston, Texas 77070 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **YANG, Sheng-Wen** [CN/CN]; 5/F, Block A, SP Tower, No. 1 Zhong Guan Cun East Road, Haidan District, Beijing 100084 (CN). **XIONG, Yuhong** [US/CN]; Room 504, Unit 133, Beijing Institute of Technology, Beijing 100081 (CN). **LIU, Wei** [US/CN]; Jinsong 311-1006, ChaoYang District, Beijing 100021 (CN).

(74) Agent: **SHANGHAI PATENT & TRADEMARK LAW OFFICE, LLC**; 435 Guiping Road, Shanghai 200233 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

WO 2011/127655 A1

(54) Title: METHOD FOR KEYWORD EXTRACTION

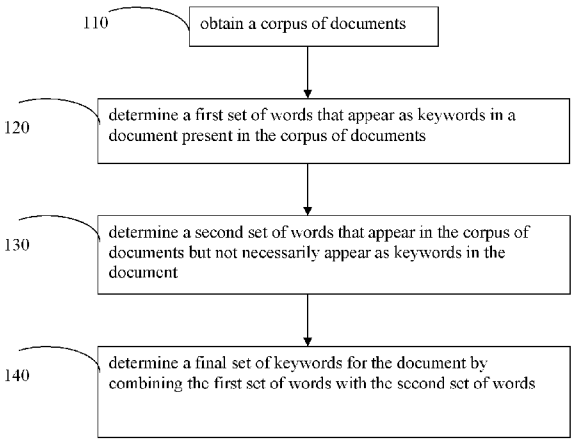

FIG. 1

(57) Abstract: Presented is a method of extracting keywords. The method includes obtaining a corpus of documents, determining a first set of words that appear as keywords in a document present in the corpus of documents, determining a second set of words that appear in the corpus of documents but not necessarily appear as keywords in the document, and determining a final set of keywords for the document by combining the first set of words with the second set of words.

METHOD FOR KEYWORD EXTRACTION

Background

With the advent of computers and the internet, the world has seen an information explosion like never before. Gone are the days, when print used to dominate the medium of expression. The internet has changed the way, people consume data. It's very common to find a digital version of almost every other document that is printed today. Such massive digitization, although immensely beneficial in many ways, has its own limitations. There is always this pressing problem of finding the right information or data. Therefore, document search remains one of the most challenging areas of research.

Keywords or keywords offer a valuable mechanism for characterizing text documents. They offer a meaningful way of searching for information in a document or corpus of documents. Traditionally, keywords are manually specified by authors, librarians, professional indexers and catalogers. However, with thousands of documents getting digitized everyday, manual specification is no longer possible. Computer-based automatic keyword extraction was a natural corollary of this problem. A number of keyword extraction methods have been proposed in the past several years. In some methods, the problem is formulated as a supervised classification problem and a classifier is trained based on a labeled training dataset. In some other methods, the keyword extraction is formulated as a ranking problem and candidate words are ranked according to some measures. The existing methods, however, have their own limitations. For example, they don't explicitly consider the semantic relationship between the candidate keywords and the document. Also, the extracted keywords are limited to the document content.

Brief Description of the Drawings

For a better understanding of the invention, embodiments will now be described, purely by way of example, with reference to the accompanying drawings, in which:

FIG. 1 shows a flow chart of a computer-implemented method of keyword extraction according to an embodiment

FIG. 2 shows a flowchart of a subroutine of the method of FIG. 1 according to an embodiment.

FIG. 3 shows a flowchart of another subroutine of the method of FIG. 1 according to an embodiment.

FIG 4. shows a block diagram of a computer system 400 upon which an embodiment may be implemented.

Detailed Description of the Invention

The following terms are used interchangeably through out the document including the accompanying drawings.

(a) "keyword" and "key phrase"
(b) "document" and "electronic document"

Embodiments of the present invention provide methods, computer executable code and computer storage medium for extracting keywords from a document which may be present in a corpus of documents. Specifically, the disclosed methods involve an in-document keyword extraction method and an in-corpus keyword extraction method. The former extracts keywords that appear in a single document; the latter extracts keywords that appear in a corpus (may not appear in the document).

FIG. 1 shows a flow chart of a method 100 of extracting keywords according to an embodiment. The method 100 may be performed on a computer system (or a computer readable medium).

The method begins in step 110. In step 110, a corpus of documents is obtained or accessed. The corpus of documents may be obtained from a repository, which could be an electronic database. The electronic database may be an internal database, such as, an intranet of a company, or an external database, such as, Wikipedia. Also, the electronic database may be stored on a standalone personal

computer, or spread across a number of computing machines, networked together, with a wired or wireless technology. For example, the electronic database may be hosted on a number of servers connected through a wide area network (WAN) or the internet.

In step 120, a document is selected from the corpus of documents, and a set of words that appear as keywords in the document is determined. The method steps involved in the selection of a set of words that appear as keywords in the document are described in further detail with reference to Fig. 2 below. At the present step, it is suffice to say that any document present in the corpus of documents may be selected and a first set of words that appear as keywords in the document may be determined. Further, the present step may be repeated for any number of documents present in the corpus of documents.

In step 130, a set of words that appear in the corpus of documents may be determined. Such set of words may not necessarily appear in the document selected in step 120. The method steps, involved in the determination of a second set of words that appear in the corpus of documents but may not necessarily appear as keywords in the document selected earlier, are described in further detail with reference to Fig. 3 below. The present step 130 is performed with regards to a corpus of documents.

In step 140, a final set of keywords for the document is determined. The step involves combining the first set of words, determined in step 120, with the second set of words, determined in step 120. Once the method steps outlined for step 120 and 130 are completed, a two set of keywords emerge that are used together to determine a final set of keywords for the document selected in step 120.

FIG. 2 shows a flowchart of a subroutine of the method of FIG. 1 according to an embodiment. The flowchart describes method step 120 in detail. The subroutine may be termed as in-document keyword extraction method. In an embodiment, the method involves following modules: learning of statistical topic modelling, inference of statistical topic modelling, noun phrase chunking, and topic-based noun phrase scoring. The main steps of the method are described as follows with notation used therein provided in Table 1 below.

Table 1

Notations

D: a corpus of documents	*d:* a document
W: a vocabulary of words	*w:* a word, $w \in W$
Z: a set of topics	*z:* a topic, $z \in Z$
W_d: a set of words in document, $W_d \subseteq W$	
$P(w\|z)$: probability of word *w* over topic *z*	$P(z\|d)$: probability of topic *z* over document *d*
$\{P(w\|z)\}_w$: a multinomial distribution of words $w \in W$ over topic $z, \Sigma_w P(w\|z)=1$	$\{P(z\|d)\}_z$, a multinomial distribution of topics $z \in Z$ over document $d, \Sigma_z P(z\|d)=1$
$\{P(w\|z)\}_{w,z}$: a set of multinomial distributions of words *W* over topics *Z*	$\{P(z\|d)\}_{z,d}$: a set of multinomial distributions of topics *Z* over documents *D*
$P(z\|d,w)$: posterior probability of topic *z* over word *w* in document *d*	$\{P(z\|d,w)\}_z$: a multinomial distribution of topic $z \in Z$ over word *w* in document *d*
$\{P(z\|d,w)\}_{z,w}$: a set of multinomial distributions of topics *Z* over words W_d in document *d*	

In step 210, a topic model is learned for a corpus of documents D, by utilizing a statistic topic modelling method. Any statistic topic modelling method, such as, but not limited to, Probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA), represented by {P(w|z)}w,z, a set of multinomial distributions of words W over topics Z and optionally {P(z|d)}z,d, a set of multinomial distributions of topics Z over documents D, may be used. Optionally, a pre-processing step may be performed, which may comprise of stop word removal, word stemming, and transformation of the corpus into a word by document matrix. Step 210 may be executed just one time for a corpus of documents. Once a model has been learnt, it may be directly applied in the following steps.

In step 220, for a given document, a multinomial distribution of topics over the document is inferred, according to the statistical topic model, to determine main topics of the document. To illustrate, in an embodiment, for a document d, the distribution of topics Z over the document d, i.e. {P(z|d)}z, is inferred according to the learnt model (in step 210), which is used to determine the main topics T of the

document by picking up the top k ones with the largest probabilities, i.e. T=argtopzP(z|d).

In step 230, posterior distributions of topics over words in the document is determined and used to assign topics to words in the document, resulting in a set of labeled words in triples. In an embodiment, the posterior distributions of topics over words in the document, i.e. {P(z|d,w)}z,w, are computed, which are used to assign topics to words by picking up the topic with the largest posterior probability for each word, i.e. z*d,w =argmaxz P(z|d,w), resulting in a set of labeled words in <w,z*,P(z*|d,w)> triples.

In step 240, a set of noun phrases are extracted from the same document by utilizing a noun phrase chunking method. The step may optionally include a post-processing step for filtering leading articles (e.g. "a", "an", "the") and pronouns (e.g. "his", "her", "your" "that", "those", etc).

In step 250, the extracted noun phrases are scored, according to occurrence of words labeled with the main topics T, and sorted in a descending order.

The scoring methods may be varied. For example, in one embodiment, the posterior probabilities of words labelled with the main topics of the document may be summed up as the score of a noun phrase. In another embodiment, the length of a noun phrase may be considered as a scoring factor by preferring bigram or trigram noun phrases.

In step 260, the top m noun phrases with highest scores are provided as an output. The output is the first set of words that appear as keywords of the document.

FIG. 3 shows a flowchart of another subroutine of the method of FIG. 1 according to an embodiment. The flowchart describes method step 130 in detail. The subroutine may be termed as in-corpus keyword extraction method. The method extracts keywords that may appear in the corpus may not necessarily appear in a particular document. The steps of the method are described as follows.

In step 310, a statistical topic model with respect to a corpus of documents is learnt. Any statistic topic modelling method, such as, but not limited to, Probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA) may be utilized for learning the statistical topic model.

Once a statistical topic model has been determined, the following steps are performed for each document in the corpus.

In step 320, for each document in the corpus, posterior distributions of topics over words is determined and used to assign topics to the words, resulting in a set of labeled words in <word, topic, probability> triples;

In step 330, for each document in the corpus, noun phrases are extracted from the document by utilizing a noun phrase chunking method. Optionally, a post-processing step of removing the articles and pronouns as described earlier may be performed, resulting in a set of noun phrases.

In step 340, each extracted noun phrase is labeled by associating each word with a topic and a weight according to the triples. This results in a sequence of triples. An output of labeled noun phrases is provided into a repository. The repository may be an electronic database.

In step 350, labeled noun phrases are read out from the repository and indexed with the help of an index engine. While indexing, the index engine may organize the sequence of triples in a way that supports the word-based search and the topic-based search, and supports the search result ranking by considering the probability as a scoring factor (step 360). Apache Lucene index engine, among others, may be customised to perform this task.

In step 370, for main topics of the document, a string query is composed. This may be done by concatenating the main topics of the document in a Boolean logic and then submitting the string query to the index engine. This results in a ranked list of matched noun phrases. The top n noun phrases are returned as keywords for the document. These are the second set of words that appear in the corpus of documents, but may not necessarily appear in the document.

FIG 4. shows a block diagram of a computer system 400 upon which an embodiment may be implemented. The computer system 400 includes a processor 410, a storage medium 420, a system memory 430, a monitor 440, a keyboard 450, a mouse 460, a network interface 420 and a video adapter 480. These components are coupled together through a system bus 490.

The storage medium 420 (such as a hard disk) stores a number of programs including an operating system, application programs and other program modules. A user may enter commands and information into the computer system 400 through input devices, such as a keyboard 450, a touch pad (not shown) and a mouse 460. The monitor 440 is used to display textual and graphical information.

An operating system runs on processor 410 and is used to coordinate and provide control of various components within personal computer system 400 in FIG. 4. Further, a computer program may be used on the computer system 400 to implement the various embodiments described above.

It would be appreciated that the hardware components depicted in FIG. 4 are for the purpose of illustration only and the actual components may vary depending on the computing device deployed for implementation of the present invention.

Further, the computer system 400 may be, for example, a desktop computer, a server computer, a laptop computer, or a wireless device such as a mobile phone, a personal digital assistant (PDA), a hand-held computer, etc.

The embodiment described provides an effective way of extracting keywords from a document by utilizing the noun phrase chunking technology to extract high-quality keyword candidates, and the statistic topic modelling technology to analyze the latent topics of text documents. The embodiment ranks the keyword candidates by considering the topic relevance between the candidate and the document as a scoring factor. By combining the in-document method and the in-corpus method, it generates a set of in-document keywords and a set of out-of-document keywords.

It will be appreciated that the embodiments within the scope of the present invention may be implemented in the form of a computer program product including computer-executable instructions, such as program code, which may be run on any

suitable computing environment in conjunction with a suitable operating system, such as, Microsoft Windows, Linux or UNIX operating system. Embodiments within the scope of the present invention may also include program products comprising computer-readable media for carrying or having computer-executable instructions or data structures stored thereon. Such computer-readable media can be any available media that can be accessed by a general purpose or special purpose computer. By way of example, such computer-readable media can comprise RAM, ROM, EPROM, EEPROM, CD-ROM, magnetic disk storage or other storage devices, or any other medium which can be used to carry or store desired program code in the form of computer-executable instructions and which can be accessed by a general purpose or special purpose computer.

It should be noted that the above-described embodiment of the present invention is for the purpose of illustration only. Although the invention has been described in conjunction with a specific embodiment thereof, those skilled in the art will appreciate that numerous modifications are possible without materially departing from the teachings and advantages of the subject matter described herein. Other substitutions, modifications and changes may be made without departing from the spirit of the present invention.

Claims:

1. A computer-implemented method of extracting keywords, comprising:

obtaining a corpus of documents;

determining a first set of words that appear as keywords in a document present in the corpus of documents;

determining a second set of words that appear in the corpus of documents but not necessarily appear as keywords in the document; and

determining a final set of keywords for the document by combining the first set of words with the second set of words.

2. A method according to claim 1, wherein the step of determining a first set of words that appear as keywords in a document, comprises:

learning a statistical topic model in respect of the corpus of documents;

inferencing, with respect to the document, a multinomial distribution of topics over the document according to the statistical topic model, to determine main topics of the document;

determining of posterior distributions topics over words in the document to assign topics to words in the document, resulting in a set of labeled words in triples;

extracting noun phrases from the document by utilizing a noun phrase chunking method;

scoring the noun phrases according to occurrence of words labeled with the main topics;

sorting the noun phrases in a descending order; and

outputting the top noun phrases with highest scores as the first set of words that appear as keywords of the document.

3. A method according to claim 2, further comprising, prior to the learning step, a preprocessing step, comprising:

removing of stop words;

stemming of words; and

transforming of the corpus of the documents into a word by a document matrix.

4. A method according to claim 2, wherein the statistical topic model is represented by a set of multinomial distributions of words over topics, and optionally a set of multinomial distributions of topics over the corpus of documents.

5. A method according to claim 2, wherein the statistical topic model is learned by Probabilistic latent semantic analysis (PLSA) or Latent Dirichlet Allocation (LDA) statistic topic modeling method.

6. A method according to claim 2, wherein determining the main topics of the document include selecting topics with largest probabilities.

7. A method according to claim 2, wherein the set of labeled words in triples is represented as <word,topic,probability>.

8. A method according to claim 2, further comprising, prior to the scoring step, a pre-processing step for filtering lead articles.

9. A method according to claim 1, wherein the step of determining a second set of words that appear in the corpus of documents, comprises:

learning a statistical topic model in respect of the corpus of documents;

determining, for each document in the corpus, posterior distributions of topics over words to assign topics to the words, resulting in a set of labeled words in triples;

extracting, for each document in the corpus, noun phrases from the document by utilizing a noun phrase chunking method;

labeling each extracted noun phrase by associating each word with a topic and a weight according to the triples; and

outputting the labeled noun phrases into a repository.

10. A method according to claim 9, further comprising reading out the labeled noun phrases from the repository and indexing the noun phrases with an index engine.

11. A method according to claim 10, further comprising:

composing, for main topics of the document, a string query by concatenating the main topics of the document in a Boolean logic; and

submitting the string query to the index engine, resulting in a ranked list of matched noun phrases, wherein top noun phrases are the second set of words that appear in the corpus of documents.

12. A method according to claim 1, wherein the corpus of documents is obtained from a repository.

13. A system, comprising:

a processor; and

a memory coupled to the processor, wherein the memory includes instructions for:

obtaining a corpus of documents;

determining a first set of words that appear as keywords in a document present in the corpus of documents;

determining a second set of words that appear in the corpus of documents but not necessarily appear as keywords in the document; and

determining a final set of keywords for the document by combining the first set of words with the second set of words.

14. A computer program comprising computer program means adapted to perform all of the steps of claim 1 when said program is run on a computer.

15. A computer program according to claim 14 embodied on a computer readable medium.

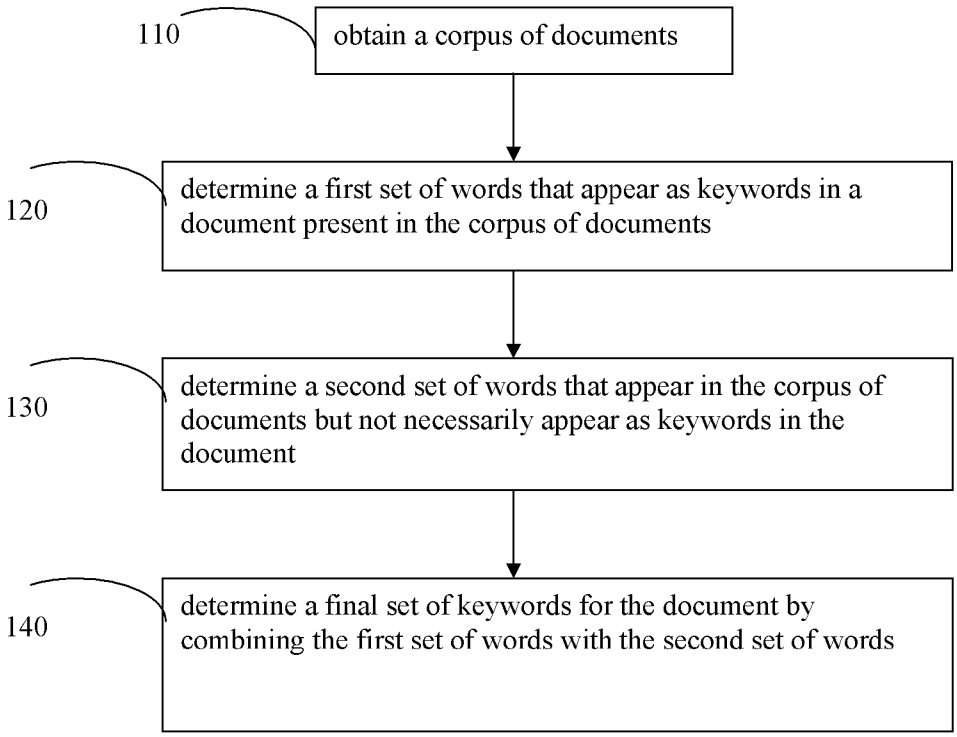
110
obtain a corpus of documents
120
determine a first set of words that appear as keywords in a document present in the corpus of documents
130
determine a second set of words that appear in the corpus of documents but not necessarily appear as keywords in the document
140
determine a final set of keywords for the document by combining the first set of words with the second set of words

FIG. 1 100

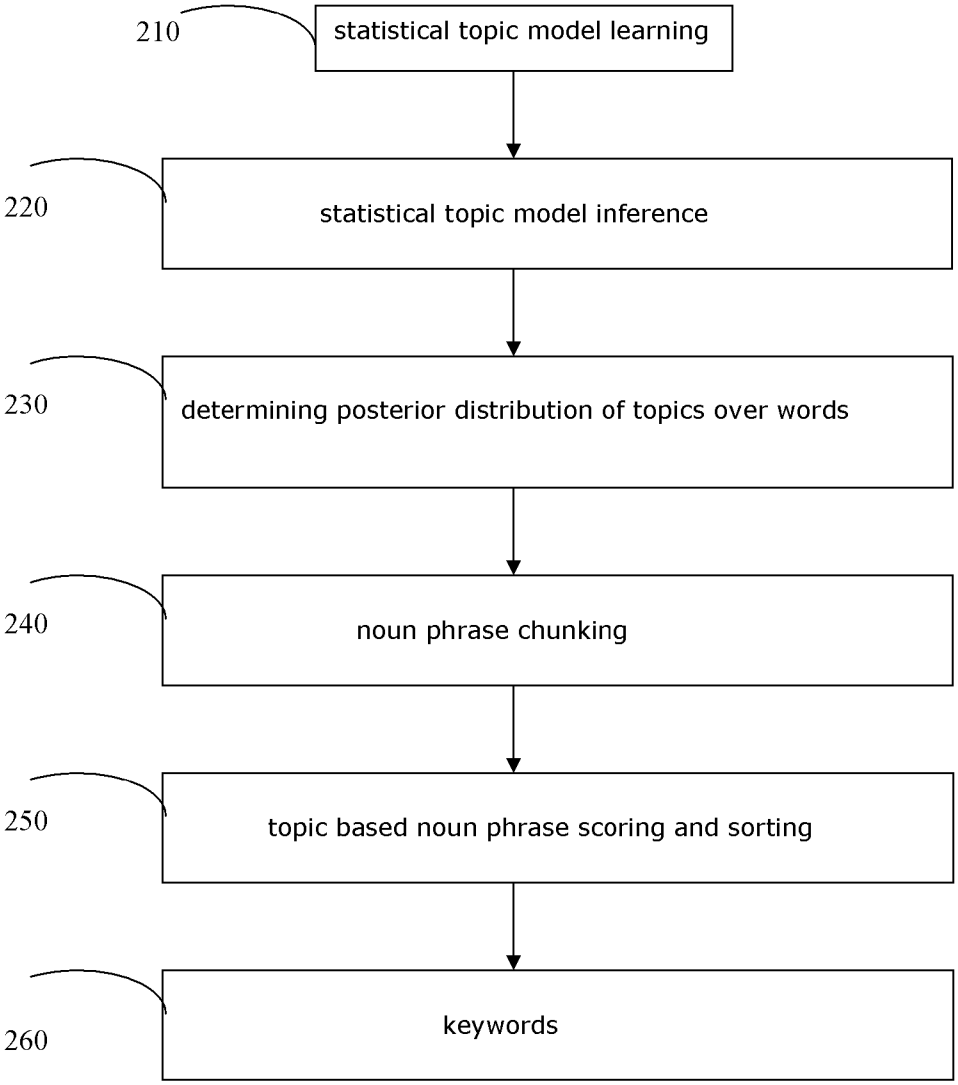
210
statistical topic model learning
220
statistical topic model inference
230
determining posterior distribution of topics over words
240
noun phrase chunking
250
topic based noun phrase scoring and sorting
260
keywords

FIG. 2 200

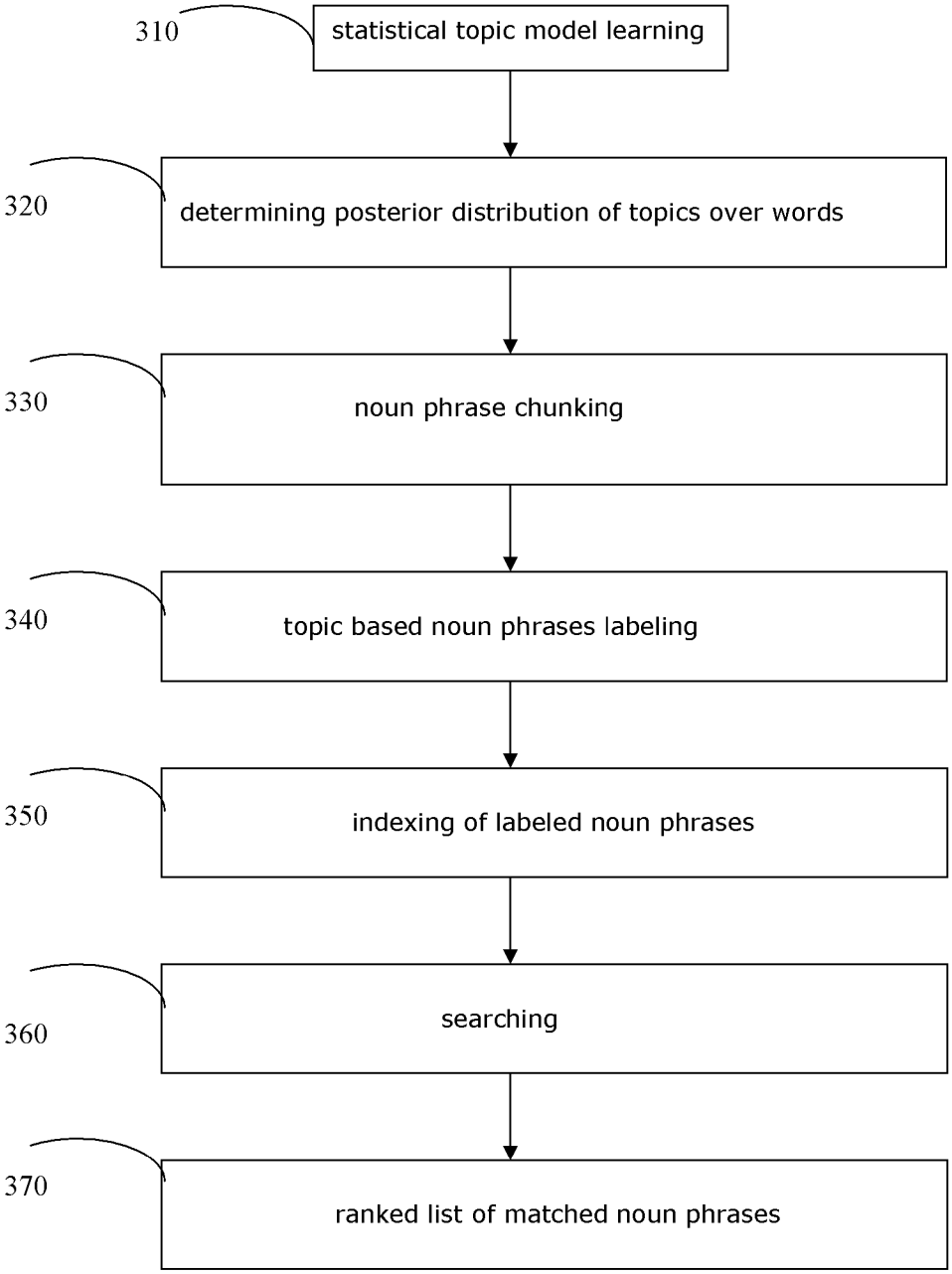
310
statistical topic model learning
320
determining posterior distribution of topics over words
330
noun phrase chunking
340
topic based noun phrases labeling
350
indexing of labeled noun phrases
360
searching
370
ranked list of matched noun phrases

FIG. 3 300

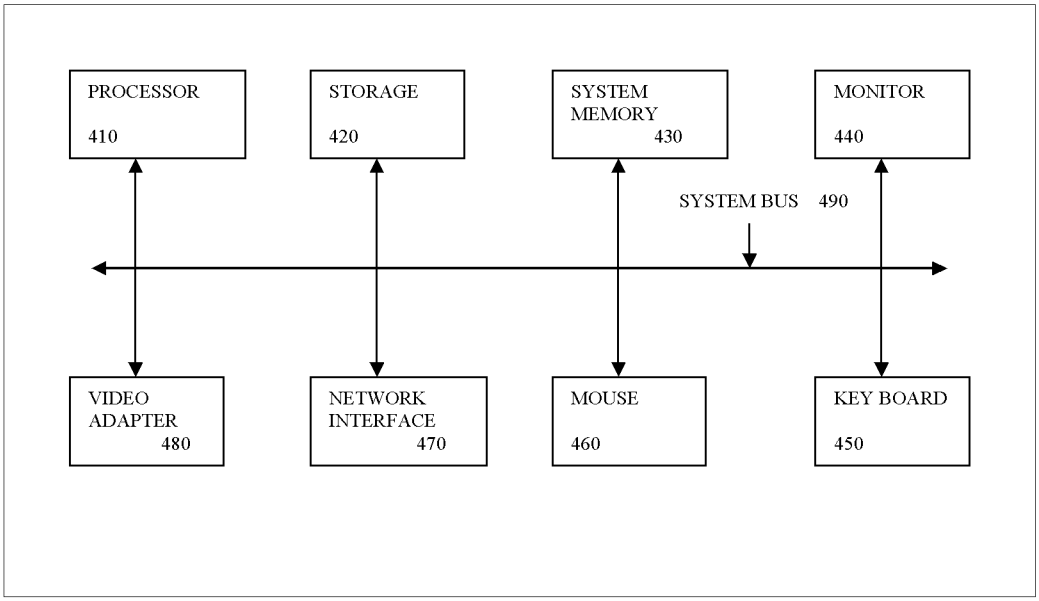
PROCESSOR
410
STORAGE
420
SYSTEM
MEMORY
430
MONITOR
440
SYSTEM BUS 490
VIDEO
ADAPTER
480
NETWORK
INTERFACE
470
MOUSE
460
KEY BOARD
450

FIG. 4 400

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2010/071758

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/30(2006.01)i

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)

IPC: G06F17/-, G06F15/-, G06K9/-

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)

WPI, EPODOC, CNPAT, CNKI: keyword, key, word, text, topic, document, file, extract, summary, search, attribute, property,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 1460947 A (INST COMPUTER SCI & TECHNOLOGY BEIJING UNIV et al.) 10 Dec. 2003(10.12.2003) see the whole document	1-15
A	CN 101183362 A (RICOH KK) 21 May 2008(21.05.2008) see the whole document	1-15

☐ 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 but 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
14 Jan. 2011(14.01.2011)	**20 Jan. 2011 (20.01.2011)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer TIAN, Jing Telephone No. (86-10)62411749

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2010/071758

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 1460947 A	10.12.2003	CN 1245696 C	15.03.2006
CN 101183362 A	21.05.2008	CN 100585594 C	27.01.2010
		US 2008114742 A1	15.05.2008
		JP 2008123526 A	29.05.2008