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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 January 2011 (20.01.2011)

PCT

(10) International Publication Number
WO 2011/006300 A1

(51) International Patent Classification:
G06F 17/30 (2006.01) *H04L 29/06* (2006.01)

(21) International Application Number:
PCT/CN2009/072782

(22) International Filing Date:
16 July 2009 (16.07.2009)

(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)*: **FENG, Shi-Cong** [CN/CN]; 5/F, Block A, SP Tower, No. 1 Zhong Guan Cun East Road, Haidan District, Beijing 100084 (CN). **XIONG, Yuhong** [US/US]; 1501 Page Mill Road, Palo Alto, California 94304-1100 (US). **LIU, Wei** [US/CN]; 112, JianGuo Road, ChaoYang District, HP Building, Beijing 100022 (CN).

(74) Agent: **SHANGHAI PATENT & TRADEMARK LAW OFFICE, LLC**; 435 Guiping Road, Xuhui, 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,

[Continued on next page]

(54) Title: ACRONYM EXTRACTION

(57) Abstract: Disclosed is a system and computer-i mplemented method for extracting an 5 acronym and one or more corresponding expansions of the acronym from a document represented in a markup language. The computer-implemented method comprises: identify ing at least one acronym contained in the document; determining one or more expansions of the at least one identified acronym based on a portion of document located proximat e the identified acronym; determining a 10 ranking for each determined expansion based attributes of the document; and selecting one or more expansions for an identified acronym using the determined rankings.

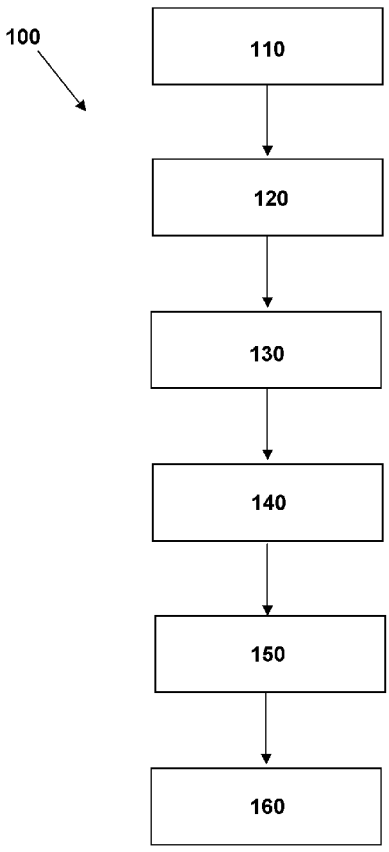

FIG. 1

WO 2011/006300 A1

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, 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, 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))*

ACRONYM EXTRACTION

An acronym is a word that is formed from the initial letter or letters of each component of a compound term or phrase (e.g., NATO, RADAR, SNAFU, etc.), while an abbreviation is a shortened form of a written word or phrase that is used or substituted for the whole word (e.g., "amt" is an abbreviation for amount). Acronyms and abbreviations tend to overlap and are frequently used in daily verbal discourse, in written documents and in electronic documents and web pages on the Internet. In certain communities (e.g., military, engineering, medicine, etc.), numerous acronyms are employed constantly. For example, a page of a military document commonly includes in excess of ten acronyms.

Acronyms present challenges to readers in several manners. In particular, individuals unfamiliar with a certain acronym may have difficulty understanding the acronym and using the acronym in vocabulary. For example, commonly known acronyms, such as "LASER" and "CDROM", are widely understood, while infrequently used or subject specific acronyms may be difficult for readers to understand (e.g., "AABFS" for Amphibious Assault Bulk Fuel System).

In many cases, a single acronym may correspond to a plurality of completely different terms, therefore resulting in confusion and ambiguity. Furthermore, acronyms are typically highly context dependent. For example, each company may have its own system of acronyms.

Various systems for acronym expansion are known. For example, the “AcronymFinder” system enables access to a manually compiled list of acronyms on the Internet. This system receives manual submissions of acronyms and corresponding expansions to update the list. However, such prior art systems suffer from several disadvantages. In particular, the AcronymFinder system is highly inefficient due to the list being generated by manual submissions. Further, the list is typically generic and static and may not suit or be tailored to various needs of particular organizations.

Although known systems extract acronyms and corresponding expansions, the results produced by these systems have limited accuracy. This tends to frustrate readers since the systems may omit acronyms within text or provide incorrect expansions for the acronyms, thereby requiring the reader to perform an additional task of ascertaining the correct expansion in another manner (e.g., manually).

BRIEF DESCRIPTION OF THE EMBODIMENTS

Embodiments are described in more detail and by way of non-limiting examples with reference to the accompanying drawings, wherein

FIG. 1 is flow diagram of a computer-implemented method according to an embodiment; and

FIG. 2 shows a data processing system in accordance with an embodiment.

DETAILED DESCRIPTION OF THE DRAWINGS

It should be understood that the Figures are merely schematic and are not drawn to scale. It should also be understood that the same reference numerals are used throughout the Figures to indicate the same or similar parts.

Proposed is a system and method for extracting an acronym and one or more corresponding expansions of the acronym from a document represented in a markup language. In particular, embodiments include: identifying at least one acronym contained in the document; determining one or more expansions of the at least one identified acronym based on a portion of document located proximate the identified acronym; determining a ranking for each determined expansion based attributes of the document; and selecting one or more expansions for an identified acronym using the determined rankings.

Embodiments can automatically and efficiently extract, normalize and rank acronym expansions from a plurality of documents such as web pages on a hosted on an organization's intranet. Different from plain text, web pages are created in a Markup Language, such Hypertext Markup Language (HTML) or

eXtensible Markup Language (XML), and contain a rich set of features such as tags and links to other web pages or documents.

By utilizing the rich set of features of web pages and explicit or implicit context cues, the extraction and ranking of acronym expansions may be improved. Valuable features of web pages and documents represented in a markup language, such as tags and links, can be used to provide information about an acronym and its corresponding expansion, thus enabling improved accuracy and efficiency performance.

Embodiments are context-aware and can leverage two types of context information: words proximate an acronym; and link structure among linked electronic documents such web pages.

Figure 1 shows a flow diagram of a computer-implemented acronym expansion method 100 according to an embodiment. Here the method is applied to web pages available on an organization's intranet network.

First, in step 110, a crawler (developed by Nutch and available to download at http://lucene.apache.org/nutch/), is used to download web pages in organization Intranet. Nutch is open source web-search software. It builds on Lucene Java, adding web-specifics, such as a crawler, a link-graph database, parsers for HTML and other document formats, etc. Next, in step 120, acronyms and their corresponding expansions are automatically extracted from the web pages using an extraction and expansion module. The expansions identified for each acronym are then normalized in step 130 using a normalization module. In step 140, the normalized expansions are then each assigned a score and ranked by score for each acronym. Low ranking expansions for an acronym are then discarded in step 150. Finally, in step 160, an acronym dictionary comprising acronyms and their corresponding expansions is generated from the remaining extracted and normalized expansions.

The above steps will now be described separately in greater detail.

Acronym Extraction (Step 120)

To automatically extract acronyms and associated expansions from web pages, candidate acronyms should be recognized with the text or markup

language of the web page. In this embodiment, a string is considered as a candidate acronym if it satisfies the following three conditions:

(i) its length is between 2 and 10 characters;

(ii) it includes at least 2 capital letters; and

(iii) its first character is alphabetic or numeric or one of only four special characters: "/" "-" "." "&".

According to the above constraints, HP (Hewlett-Packard), HPLC (HP Labs China), 3D (3 Dimensions), U.S.A (United States of America), R&D (Research and Development), CD-ROM (Compact Disk - Read Only Memory) and I/O (Input/Output) are all regarded as candidate acronyms. The conditions are applied to make the extraction precise and efficient, although some common acronyms in lower case such as "btw" and "aka" are excluded. Exceptions to the condition can be defined such that if a string is embedded in one of the two standard HTML tags <abbr></abbr> or <acronym></acronym> pairs, for example, the string is recognized as a candidate acronym without any constraints or conditions being imposed. Another exemplary condition may define that one or more predetermined terms are excluded from consideration as an acronym.

Once a candidate acronym is recognized, its associated expansion is extracted within its contexts (in other words, from text proximate the acronym). The context size defines how many words near the acronym are taken account of when determining an associated expansion and has a numerical value corresponding to the number of words before or after the acronym searched for the expansion. Here, the context size is the smaller value of: the length of the characters in the candidate acronym plus five; and the length of the text node that contains the candidate acronym.

Unlike conventional acronym extraction algorithms, each web page is processed twice in this embodiment.

In the first processing pass, a HTML web page is treated as a hypertext. A HTML parser parses the page content and a HTML DOM tree is built by the parser. For each text node in the DOM tree, if its XPath contains <abbr> or <acronym> tag, an acronym is extracted from this text node and associated

expansion is extracted from "title" attribute of its parent node. Otherwise, the text in the node is tokenized by the space character. For each token in this text node, if it is an acronym candidate, its expansion is extracted by the following extraction algorithm (presented in pseudo code):

```
Input: an acronym and its context
for each character in an acronym from left to right
        if this character is punctuation, ignore it and continue;
        scan the characters in the context from left to right;
        if no matched character is found in its context
                extracted result ← null and exit;
        else
                continue;
until all characters in an acronym is checked
if all characters in an acronym are found in its context in order
        extracted result ← extracted acronym-expansion pair;
else
        extracted result ← null;
Output: extracted result
```

The extraction algorithm above describes how to extract associated expansion from its context for an acronym. The context includes the words to the left and right of the acronym candidate. In other words, a bi-directional candidate search technique is used to select terms near an acronym in order to enhance identification of acronym expansions. Because the left context case is very similar to the right one, only the right case is presented in the algorithm above.

The extraction algorithm is simple and efficient, since it only requires a few character comparisons. However, for improved accuracy, the algorithm may be constrained using the following rules:

(i) The first character of an acronym and the first character of its expansion must be same.

(ii) If each character of an acronym is the first character of each token in the context in order, the context is regarded as the expansion of this acronym.

Once rule (ii) is matched, an expansion is extracted from this context. Otherwise, the extraction algorithm presented in pseudo code above is used to extract the expansion. Rule (ii) is applied above check whether the context contains an expansion before the extraction algorithm presented in pseudo code is applied. This is because the extraction algorithm does not move on to find the next one once it finds a matched character in its candidate expansion.

In the second pass, a web page is transformed to plain text by removing all HTML tags. The acronyms embedded in <abbr>, <acronym>, <alt> and other tags are discarded when these tags are removed. The text is also tokenized using the space character. After all tokens are processed using an extraction algorithm as above, all acronym-expansion pairs are recognized. Finally, the extracted results of the two passes are combined together to get as many acronyms as possible.

It will be appreciated that embodiments process each web page twice. In a first pass, some web pages may not be correctly parsed by the HTML parser. For example, some text nodes may be discarded due to closing tags being missing in the web page. The second pass then detects and extracts those discarded in the first pass since the markup tags are disregarded. This combined extraction approach can therefore discover acronyms which are otherwise discarded in an individual pass.

Normalization (Step 130)

Embodiments group expansions that have the same meaning to an acronym. For example, both “Research and Development” and “Research & Development” are the expansions of “R&D”. For this, two normalization algorithms are proposed.

The first normalization algorithm is based on direct string comparison and comprises the following steps for each acronym expansion:

(a) Change expansion to lower case.

(b) Replace all punctuation and stop words with a space character

(c) Tokenize the expansion by decomposing text into string tokens or words.

(d) Stem each token.

(e) Combine all stemmed tokens in sequence to generate a canonical string.

(f) Create an entry in a hash table using the canonical string as a key and the original expansion as its value. If the key already exists in the hash table, the current expansion is appended.

After all expansions are processed, the expansions having the same meaning are grouped in the result hash table. This string comparison based normalization algorithm has high computing efficiency.

For improved accuracy, a second normalization algorithm is proposed, which is based on K-medoids clustering and comprises the following steps for each acronym expansion:

(a) Change expansion to lower case.

(b) Replace all punctuation and stop words with a space character

(c) Tokenize the expansion by decomposing text into string tokens or words.

(d) Stem each token.

(e) Split each stemmed token into a set of three-character trigrams groups. For example, after “technical” is stemmed to “technic”, it is broken into “tec, ech, chn, hni, nic” trigram set.

(f) Create a vector from the tri-gram set by first sorting all trigrams in alphabetical order and then denoting this by a vector (for example, “tec, ech, chn, hni, nic” is first sorted to "chn, ech, hni, nic, tec" and then denoted by a vector <chn, ech, hni, nic, tec>).

(g) Assign the weight of each element in the vector to an integer, e.g. 1.

All expansions are processed in the same way and mapped to vectors. The K-medoids clustering algorithm, described in book "Data Mining Concepts and Techniques", J. Han and M. Kamber, Morgan Kaufmann Publishers, Inc., 2001, is then applied to group expansions.

The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoidshift algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize squared error, the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm k-medoids chooses datapoints as centers (medoids or exemplars).

Thus, k-medoid is a partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. A useful tool for determining k is the silhouette. It is more robust to noise and outliers as compared to k-means.

A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set.

In summary, the idea behind K-medoids is as follows: The initial representative objects (or seeds) are chosen arbitrarily. The iterative process of replacing representative objects by non-representative objects continues as long as the quality of the resulting clustering is improved. This quality is estimated using a cost function that measures the average dissimilarity between an object and the representative object of its cluster.

To reduce the number of similarity comparisons compared to the original K-medoids clustering algorithm, which computes similarity between every pair of expansions, a cosine measurement may be used to compute the similarity between two expansions. At the beginning of clustering, the expansion with the largest score is chosen as the initial cluster. Then, the other expansions are compared to this expansion by cosine measurement. If the similarity exceeds a threshold value, the expansion is clustered into the initial cluster. Otherwise, a new cluster is created. This process is then repeated until all expansions are grouped. Investigations have shown that 0.7 is a suitable threshold value.

Ranking (Step 140)

It is common for acronyms to have multiple expansions, even within a digital library or document database of a single organization. For example, 10

possible expansions for “ABC” were found from 656,350 intranet web pages of a large organization.

To discriminate the expansions for an acronym, a score is assigned to each expansion and then the expansions are ranked according to their score. After ranking the expansions, the expansion(s) having the most dominant meaning of an acronym can be identified. Lower ranked expansions may be incorrect due to typos or incorrect extraction and can be filtered out.

The ranking algorithm makes use of context cues implied in web pages while scoring expansions. There are two kinds of context; local context and global context. The local context is the text around or proximate an acronym within a page. The global context is the links among the web pages in a website or network.

First, a local score for every expansion is calculated based on the local context. A set of scoring rules can be used for defining how a local score is generated. An exemplary set of localscore scoring rules is listed in table 1 below:

Localscore Scoring Rules	Localscore
If an expansion is embedded in <abbr> or <acronym> tag	+10
If some special patterns are matched: “E(A)”, “A(E)”, “E:A”, “E=A” and “E-A”, where A = Acronym and E = Expansion	+5
If some special syntactic information is matched, such as “A stands for E”, “E is also known as A”, “A short for E”, “A is E”, “A refers to E”, and “A, see E”	+5
If the first letter in each token of an expansion match a letter in its acronym in order	+1
If expansion is embedded in or tag	+0.5
Each stop word in an expansion	-0.5
Each punctuation character in an expansion	-0.5
Compare the length of expansion (in term of the number of words) and its acronym (in term of the number of characters), each additional word	-0.5

Each extra word between the acronym and the expansion	-0.5

Table 1

Using the localscore scoring rules, such as those in Table 1 above, a localscore can be calculated.

The algorithm proposed by L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," in Technical report, Stanford University Database Group, 1998, is then used to compute a PageRank value of each web page for use as an indicator of the quality of a page. The quality of the page is in turn used to help decide the quality of expansions extracted from the web page. The potential expansions extracted from a high quality web page are typically more reliable than from ones with poor quality.

The final score (globalscore) of this expansion in the whole corpus is calculated using the following formula (1):

$$globalscore = \sum_{i=1}^{n} (localscore_i \times pr_i) \quad (1),$$

wherein pr is the PageRank value of the web page from which the acronym-expansion is extracted and n is the number of the web pages containing this expansion. Thus, it will be appreciated that both the local and the global cues are taken into account by the formula (1).

The local and the global context are then used together to rank the expansions in order of globalscore. By ranking the expansion in order, incorrect expansions such as those with the lowest globalscore can be filtered out (Step 150).

Turning now to Figure 2, a data processing system 400 in accordance with an embodiment is shown. A computer 410 has a processor (not shown) and a control terminal 420 such as a mouse and/or a keyboard, and has access to an electronic library or document database stored on a collection 440 of one or more storage devices, e.g. hard-disks or other suitable storage devices, and has access to a further data storage device 450, e.g. a RAM or ROM memory, a hard-disk, and so on, which comprises the computer program product

implementing a method according to an embodiment. The processor of the computer 410 is suitable to execute the computer program product implementing a method in accordance with an embodiment. The computer 410 may access the collection 440 of one or more storage devices and/or the further data storage device 450 in any suitable manner, e.g. through a network 430, which may be an intranet, the Internet, a peer-to-peer network or any other suitable network. In an embodiment, the further data storage device 450 is integrated in the computer 410.

It will be appreciated that embodiments provide advantages which can be summarized as follows:

Unlike human-edited systems, proposed embodiments are automatic and, hence, do not require excessive human effort in order to create and maintain acronym listings/dictionaries.

Embodiments utilize a rich set of features of hypertext and explicit or implicit context cues to help extract, normalize and rank acronym-expansions. Experimental results show that by using these features, significant performance improvements can be achieved.

Organization-specific or network-specific acronyms expansions can be collected and indexed. Many special acronyms used in an organization are not collected into the general-purpose acronym dictionaries because of their limited coverage.

The string-comparison-based normalization algorithm has improved efficiency because each expansion only need be scanned once. The computation complexity is O(n), where n is the number of expansions. Thus, embodiments may be suitable for application with large numbers of web pages that require huge computing resource.

It should be noted that the above-mentioned embodiments are illustrative, and that those skilled in the art will be able to design many alternative embodiments without departing from the scope of the appended claims. In the claims, any reference signs placed between parentheses shall not be construed as limiting the claim. The word "comprising" does not exclude the presence of

elements or steps other than those listed in a claim. The word "a" or "an" preceding an element does not exclude the presence of a plurality of such elements. Embodiments can be implemented by means of hardware comprising several distinct elements. In the device claim enumerating several means, several of these means can be embodied by one and the same item of hardware. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measures cannot be used to advantage.

CLAIMS

1. A computer-implemented method of extracting an acronym and one or more corresponding expansions of the acronym from a document represented in a markup language, the method comprising the steps of:

examining the document to identify at least one acronym contained therein;

retrieving a portion of the document for each identified acronym, wherein the portion is located within the document proximate the identified acronym;

determining one or more expansions of each identified acronym based on the corresponding retrieved document portion;

determining a ranking for each expansion based on attributes of the document; and

selecting one or more expansions for an identified acronym using the determined rankings.

2. The method of claim 1, wherein the step of examining comprises:

determining the existence of text between markup tags indicating the presence of an acronym so as to identify an acronym contained therein;

generating a text version of the document by removing the markup language tags from the document; and

examining the generated text version of the document to identify at least one acronym contained therein.

3. The method of claim 2, wherein the step of examining the generated text version of the document comprises determining a grammatical classification of each term within the text version to enable identification of an acronym.

4. The method of claim 3, wherein determining a grammatical classification comprises examining a term within the text version and determining compliance of text term attributes with conditions in order to identify that the text term is an acronym, and wherein said conditions relate to at least one of:

a length of a text term being within a particular range;

an amount of capitalization within a text term;

the presence of a special character within a text term; and

a text term being a predetermined term excluded from consideration as an acronym.

5. The method of claim 1, wherein the step of retrieving comprises selectively decomposing terms within at least one retrieved document portion into individual sub-terms to facilitate identification of an expansion within that retrieved document portion.

6. The method of claim 1, wherein the step of determining one or more expansions comprises identifying terms within each retrieved document portion that include an initial portion containing a first character of said acronym and producing a corresponding set of terms for each identified term including said identified term and following terms within the retrieved document portion.

7. The method of claim 1, wherein the step of determining a ranking utilizes an algorithm considering attributes of the markup language of the corresponding retrieved document portion and an indicator of the reliability of the document.

8. The method of claim 7, wherein the algorithm is adapted to determine compliance of attributes of the markup language with rules, and wherein said rules relate to at least one of:

the presence of predetermined markup tags;

the presence of one or more special characters;

the presence of predetermined text indicating an expansion of an acronym.

9. Apparatus for extracting an acronym and one or more corresponding expansions of the acronym from a document represented in a markup language, the method comprising the steps of:

an identification unit adapted to examine the document and to identify at least one acronym contained therein;

a context retrieval unit adapted to retrieve a portion of the document for each identified acronym, wherein the portion is located within the document proximate the identified acronym;

an expansion determination unit adapted to determine one or more expansions of each identified acronym based on the corresponding retrieved document portion;

a ranking unit adapted to determine a ranking for each expansion based on attributes of the document; and

an expansion selection unit adapted to select one or more expansions for an identified acronym using the determined rankings.

10. The apparatus of claim 9, wherein the identification unit comprises:

a markup tag identification unit adapted to determine the existence of text between markup tags indicating the presence of an acronym so as to identify an acronym contained therein;

a document conversion unit adapted to generating a text version of the document by removing the markup language tags from the document; and

a text identification unit adapted to examining the generated text version of the document to identify at least one acronym contained therein.

11. The apparatus of claim 9, wherein the an expansion determination unit is adapted to identify terms within each retrieved document portion that include an initial portion containing a first character of said acronym and to produce a corresponding set of terms for each identified term including said identified term and following terms within the retrieved document portion.

12. The apparatus of claim 9, wherein the ranking unit utilizes an algorithm considering attributes of the markup language of the corresponding retrieved document portion and an indicator of the reliability of the document.

13. A computer program product arranged to, if executed on a computer, cause the computer to execute the steps of:

establishing an application programming interface, API, connection between an application and a database;

establishing first and second socket connections between the application and the database in the API connection, wherein the first socket connection is secure and the second socket connection is non-secure;

communicating information through the first or second socket connection based on whether the information is identified as being confidential information or not.

14. A computer-readable data storage medium comprising the computer program product of claim 13.

15. A data processing system comprising:

data storage means for storing information;

a computer program memory comprising the computer program product of claim 13; and

a data processor having access to the computer program memory and the data storage means, the data processor being arranged to execute said computer program product.

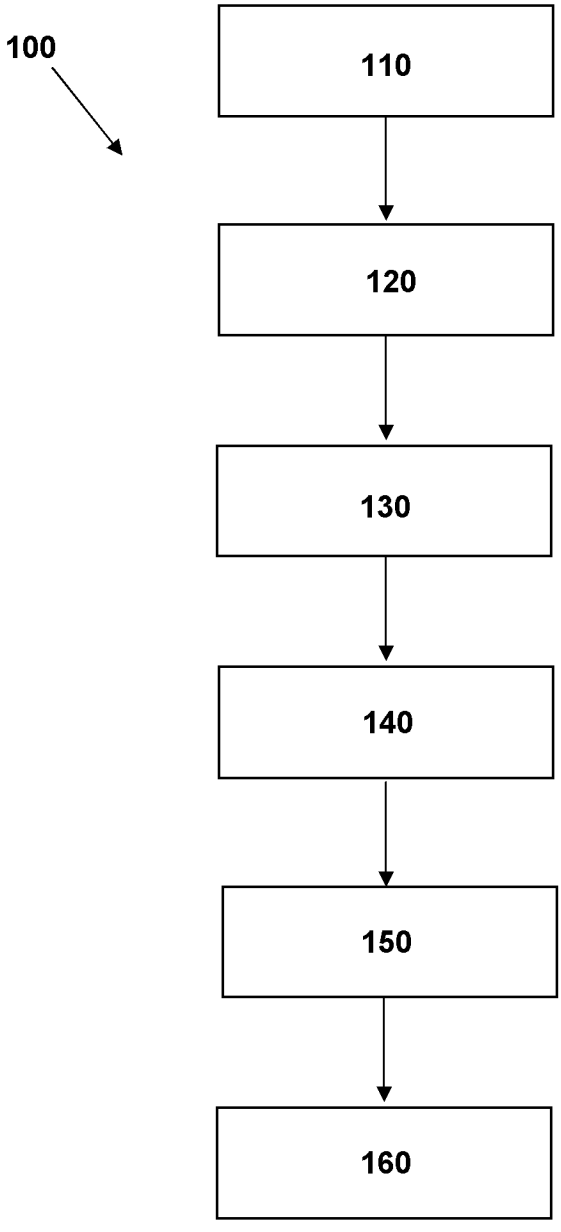

FIG. 1

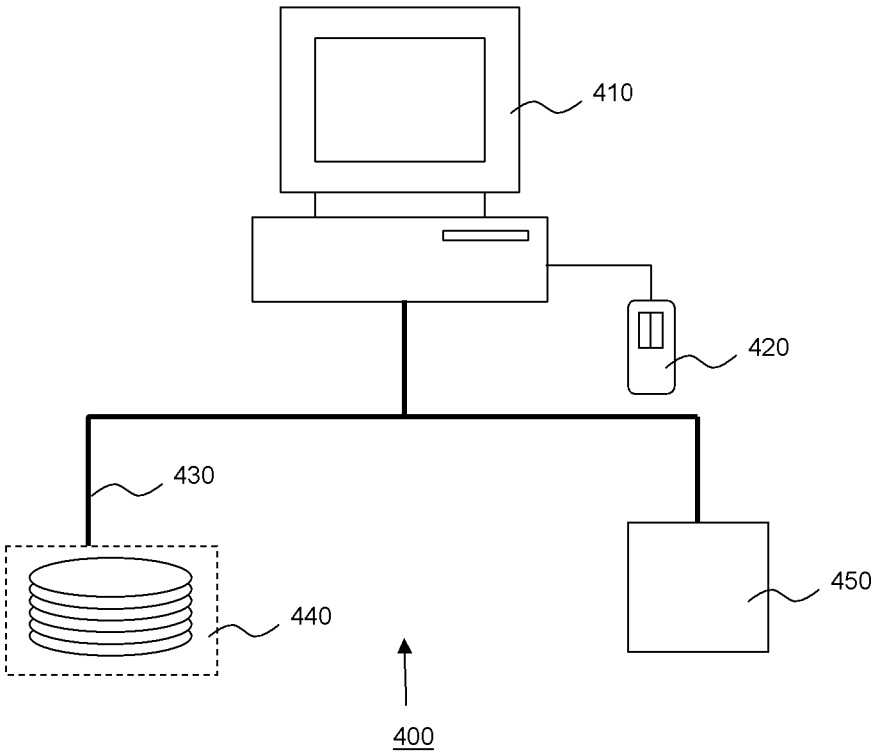

FIG. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2009/072782

A. CLASSIFICATION OF SUBJECT MATTER

See extra sheet

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: G06F ,H04L

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 CPRS CNKI(acronym expansion abbreviation term markup language rank search retrieve document API socket database)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2008033714 A1(ITT MFG ENTERPRISES INC) 07 Feb. 2008(07.02.2008) see the whole document	1-12
A	US 2002152064 A1(INT BUSINESS MACHINES CORP)17 Oct. 2002(17.10.2002) see the whole document	1-12
A	CN 1725757 A(SAMSUNG ELECTRONICS CO LTD)25 Jan. 2006(25.01.2006) see the whole document	13-15

☐ 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
31 March 2010(31.03.2010)	**15 Apr. 2010 (15.04.2010)**
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 WANG,Yankun Telephone No. (86-10)62411680

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2009/072782

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

This Authority considers that there are two inventions covered by the claims indicated as follows: I: Claims 1-12 directs to a computer-implemented method and apparatus for extracting an acronym and one or more corresponding expansions of the acronym from a document represented in a markup language. II: Claims 13-15 directs to a data processing system, a computer program product and a computer-readable data storage medium arranged to cause the computer to execute the steps of establishing API connection between an application and a database; establishing first and second socket connection between the application and the database in the API connection; and communicating information through the first or second socket connection based on whether the information is identified as being confidential information or not. The application, hence does not meet the requirements of unity of invention as defined in Rules 13.1 and 13.2 PCT.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☒ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2009/072782

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US2008033714A1	07.02.2008	US7236923B1	26.06.2007
US2002152064A1	17.10.2002	NONE	
CN1725757A	25.01.2006	JP2006031685A	02.02.2006
		EP1619855A1	25.01.2006
		US2006020705A1	26.01.2006
		KR20060007732A	26.01.2006
		KR100560752B1	13.03.2006
		EP1619855B1	07.05.2008

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2009/072782

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01)i

H04L 29/06 (2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC