



- (51) **International Patent Classification:**
G06F 17/30 (2006.01)
- (21) **International Application Number:**
PCT/MY2013/000254
- (22) **International Filing Date:**
12 December 2013 (12.12.2013)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
PI 2012701213 18 December 2012 (18.12.2012) MY
- (71) **Applicant:** MIMOS BERHAD [MY/MY]; Technology Park Malaysia, Kuala Lumpur, 57000 (MY).
- (72) **Inventors:** SADANANDAN, Arun Anand; c/o Mimos Berhad, Technology Park Malaysia, 57000 Kuala Lumpur (MY). YEEK, Tan Sieow; c/o Mimos Berhad, Technology Park Malaysia, 57000 Kuala Lumpur (MY). SEAN, Lim Ying; c/o Mimos Berhad, Technology Park Malaysia, 57000 Kuala Lumpur (MY). LUKOSE, Dickson; c/o Mimos Berhad, Technology Park Malaysia, 57000 Kuala Lumpur (MY).
- (74) **Agent:** H.A. RASHID, Ahmad Fadzlee; IPVOLUSI Sdn Bhd, A-6-6, Centrio Pantai Hillpark, No 1, Jalan Pantai Murni, 59200 Kuala Lumpur (MY).
- (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, BN, 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, IR, IS, JP, KE, KG, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, 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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *of inventorship (Rule 4.17(iv))*

Published:

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

[Continued on next page]

- (54) **Title:** A SEMANTIC QUERY SYSTEM AND METHOD THEREOF

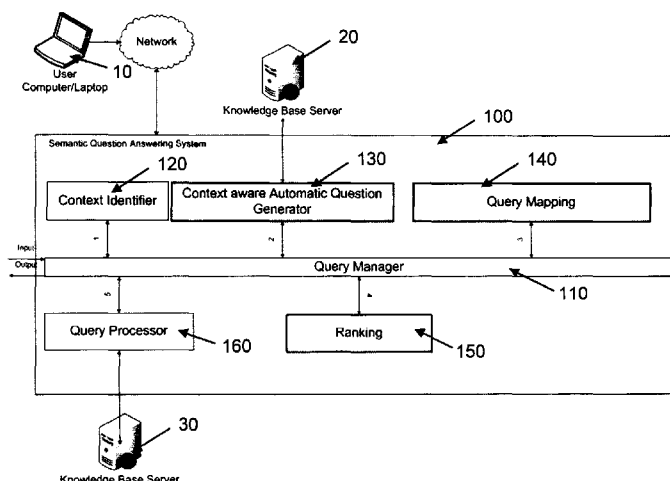


FIG. 1

(57) **Abstract:** The present invention relates to a semantic query system (100) and a method for processing a query. The semantic query system (100) comprises of a query manager (110), a context identifier (120), a question generator (130), a query mapper (140), a ranking component (150) and a query processor (160). When the semantic query system (100) receives a query, the context identifier (120) identifies all contexts related to the query. Thereon, the question generator (130) generates candidate questions based on the contexts. The query mapper (140) maps the query to the candidate questions by performing a syntactic analysis and a social network analysis on the candidate questions with respect to the query. The ranking component (150) produces a list of ranked candidate questions. Based on the list, the query processor (160) retrieves and ranks relevant information which is displayed as the query results.

- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

A SEMANTIC QUERY SYSTEM AND METHOD THEREOF

FIELD OF INVENTION

The present invention relates to an information retrieval system and more particularly to a semantic query system.

BACKGROUND OF THE INVENTION

Semantic query system is used to retrieve relevant information based on a query by understanding the contextual meaning of the query. Moreover, the semantic query system ranks relevancy of the retrieved information based on semantics. Thus, the semantic query system improves search accuracy compared to a typical search engine which retrieves information that matches the keywords in the query.

In order to provide an effective semantic query system, the semantic query system is evaluated based on precision and recall, wherein precision is defined as the number of relevant results retrieved by a search divided by the total number of results retrieved by that search, while recall is defined as the number of relevant results retrieved by a search divided by the total number of possible relevant results. High precision means that a semantic query system retrieves more relevant results than irrelevant results, while high recall means that a semantic query system retrieves most of the relevant results. Thus, the semantic query system needs to balance out between precision and recall.

Most of the existing semantic query systems conduct a search by mapping a query into internal query structures such as SPARQL and thereon, retrieving the relevant results. Such semantic query systems increase precision at the expense of recall. Moreover, such semantic query systems pose a mapping problem whereby the correct mapping of the query into the internal query may be difficult to find due to the complexity in natural language query. In order to resolve this, the semantic query systems use template based approach. However, such systems require a query to rigidly suit the structure of the templates in its database and thus, the systems may not be able to process any types of query unless it suits the templates. In order to expand the types of the query to be processed, the systems need to include more templates in its database.

Therefore, there is a need to provide a semantic query system that addresses the above-mentioned problems.

5 SUMMARY OF INVENTION

The present invention provides a semantic query system (100). The semantic query system (100) comprises a query manager (110) configured to control activities of the system (100); a context identifier (120) configured to identify at least one context related to a query, a question generator (130) configured to generate at least one candidate question based on the at least one context; a query mapper (140) configured to perform social network analysis on the at least one candidate question generated by the question generator (130) with respect to the query; a ranking component (150) configured to produce a list of ranked candidate question generated by the question generator (130); and a query processor (160) configured for retrieving relevant information based on the list of ranked candidate question generated by said question generator (130). The context identifier (120), the question generator (130), the query mapper (140), the ranking component (150) and the query processor (160) are connected to said query manager (110).

The present invention also provides a method for processing a query. The method is characterised by the steps of receiving a query; identifying at least one context related to the query; generating at least one candidate question based on the at least one context; mapping the query to the at least one candidate question to produce a list of ranked candidate question; and retrieving and ranking relevant information based on the list of ranked candidate question.

Preferably, mapping the query to the at least one candidate question includes the steps of performing a syntactic analysis between the query and the at least one candidate question; performing a social network analysis on the at least one candidate question with respect to the query; determining a semantic score, similarity score, and confidence score for each candidate question; normalizing the semantic score and similarity score for each candidate question; and producing a list of ranked candidate question by ranking the at least one candidate question based on the normalized semantic and similarity scores. The social network analysis is preferably performed by constructing a social network based on co-occurrence of concepts or

terms in the at least one candidate question; determining relevance value between the co-occurred terms for each term; and computing centrality value for each term.

BRIEF DESCRIPTION OF THE DRAWINGS

5 The accompanying drawings, which are incorporated in and constitute a part of the specification, illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention.

10 **FIG. 1** shows a block diagram of a semantic query system (**100**) according to an embodiment of the present invention.

FIG. 2 shows a flowchart of a method for processing a query according to an embodiment of the present invention.

15 **FIG. 3** shows a flowchart of a method mapping a query to questions generated by a question generator (**130**) according to an embodiment of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

20 A preferred embodiment of the present invention will be described herein below with reference to the accompanying drawings. In the following description, well known functions or constructions are not described in detail since they would obscure the description with unnecessary detail.

25 Referring to **FIG. 1**, there is shown a block diagram of a semantic query system (**100**) according to an embodiment of the present invention. The system (**100**) retrieves closest and most relevant information or results based on a natural language query provided by a user. The system (**100**) is connected to at least one user device (**10**) via a network. The user device (**10**) may include but not limited to laptop, mobile phone, personal computer, handheld communication device, and
30 handheld computing device. The system (**100**) comprises of a query manager (**110**), a context identifier (**120**), a question generator (**130**), a query mapper (**140**), a ranking component (**150**) and a query processor (**160**).

35 The query manager (**110**) is configured to control the flow of activities for the overall system (**100**). The query manager (**110**) is connected to the context identifier

(120), the question generator (130), the query mapper (140), the ranking component (150) and the query processor (160).

5 The context identifier (120) is configured to identify one or more contexts related to the query by using a classification algorithm such as k-nearest neighbour classification algorithm, Naive-Bayes classification algorithm or any other machine learning algorithms suitable for classifying a query.

10 The question generator (130) is configured to automatically generate one or more candidate questions based on the context identified by the context identifier (120). The question generator (130) generates the candidate questions by identifying concepts and relationships of each identified context and incorporating the concepts with one or more question stems. The question generator (130) is connected to a knowledge base server (20) to identify and analyse the concepts and relationships
15 based on the identified context during the process of generating the questions.

The query mapper (140) is configured to perform social network analysis on the candidate questions generated by the question generator (130) with respect to the query. The query mapper (140) performs the social network analysis by
20 constructing a social network based on co-occurrence of concepts or terms in the candidate questions, and performing co-occurrence analysis, centrality analysis, and relevance analysis. As a result, the query mapper (140) identifies multiple sub-graphs from the social network analysis, wherein each sub-graph shows how the query terms appear together and the closeness of query terms shows the similarity
25 between the query terms. Thus, the similar query terms can be identified based on the sub-graphs. Moreover, the relevance value and importance value of each sub-graph is determined, wherein the relevance value is computed using the relevance analysis and the importance value is computed using the centrality analysis. The relevance analysis and centrality analysis are based on graph theory and network
30 analysis to determine the relative importance of a node within each sub-graph. The query mapper (140) is also configured to perform syntactic analysis between the query and the candidate questions.

The ranking component (150) is configured to rank the sub-graphs which lead
35 to ranking the candidate question as the sub-graphs composed of concepts in the

candidate questions. This is to identify the candidate questions that are semantically closest to the query. The ranking is based on normalized semantic and similarity scores.

5 The query processor (160) is configured to retrieve and rank relevant results based on the candidate questions from the ranking component (150). The relevant information is displayed as the query results. The query processor (160) is also connected to a knowledge base server (30) to retrieve relevant information based on the candidate questions.

10

Referring to FIG. 2, there is shown a method for processing a query by using the semantic query system (100) of FIG. 1. Initially, as in step 201, the query manager (110) of the semantic query system (100) receives a query from a user device (10). The query manager (110) sends the query to the context identifier (120).
15 Based on the query, the context identifier (120) identifies all contexts related to the query by using a classification algorithm.

Thereon, the identified contexts are sent to the question generator (130). The question generator (130) automatically generates candidate questions based on the context identified by the context identifier (120) as in step 202.
20

In step 203, the query is mapped to the candidate questions. As a result, a list of ranked candidate questions is produced.

25 The list of ranked candidate questions is then transmitted to the query processor (160). The query processor (160) searches and retrieves relevant results based on the list of ranked candidate questions as in step 204. The results retrieved are answers to the candidate questions in the list.

30 In step 205, the results retrieved is transmitted and displayed as the query results to the user device (10). The query results are displayed according to the ranking of the candidate questions.

Referring now to FIG. 3, there is shown a flowchart of a method mapping a
35 query to candidate questions generated by the question generator (130). Initially, as

in step **310**, a syntactic analysis is performed between the query and the candidate questions generated by the question generator (**130**). The syntactic analysis includes extracting the terms in the query (step **311**), finding terms having similar meaning to the terms extracted from the query (step **312**), performing morphological analysis on all the terms (step **313**), and term matching or matching the terms in the query with the terms in the candidate questions (step **314**).

In step **320**, the query mapper (**140**) performs a social network analysis on the candidate questions generated by the question generator (**130**) with respect to the query. The social network analysis is initiated by constructing a social network based on co-occurrence of concepts or terms which indicates terms that appear together in the candidate questions (step **321**). For instance, assuming that term1 and term2 appear in a first candidate question while term3 and term4 appear in a second candidate question, the query mapper (**140**) constructs a social network of term1-term2-term3-term4 based on co-occurrence of the terms in the candidate questions. Based on the social network, the query mapper (**140**) determines the relevance value between the co-occurred terms (step **322**) and thereon, computes the centrality value for each term (step **323**).

In step **330**, each candidate question is scored based on the syntactic analysis and social network analysis to determine a semantic score, similarity score and confidence score for each candidate question (steps **331** to **333**). The semantic score is based on the relevance and centrality values for each candidate question. The similarity score is based on the number of matched terms based on the syntactic analysis performed. The confidence score is a relative measure of certainty of matched terms based on the syntactic analysis performed.

Thereon, the semantic and similarity scores for each candidate question are normalized as in step **340**.

In step **350**, the candidate questions are ranked according to the normalized scores and/or confidence scores by the ranking component (**150**). Thus, this produces a list of the ranked candidate questions.

For a further understanding of this invention, a specific example is provided herein below for illustration purpose only and is not intended to be limiting unless otherwise specified.

5 A user device (10) sends a query of "*cancer and its effect on today's youth*" to the semantic query system (100). Thereon, the query manager (110) sends the query to the context identifier (120).

10 Based on the query, the context identifier (120) identifies all contexts related to the query by using a classification algorithm. The contexts that have been identified are *cancer*, *disease*, and *malignant neoplasm*.

15 The identified contexts are then sent to the question generator (130) for generating candidate questions based on the identified context. The candidate questions generated are listed below:

Q1: *What are the symptoms of cancer?*

Q2: *Is cancer treatable?*

Q3: *Are youngsters taking to smoking more than before?*

Q4: *Studies on young people falling victims to cancer*

20 Q5: *Are youngsters more prone to diseases?*

Q6: *What are the factors that cause cancer to young people?*

Q7: *What are the dangerous hereditary diseases in today's world?*

25 Thereon, the query is mapped to the candidate questions based on syntactic analysis and social network analysis. Based on the syntactic analysis, the similarity score for each candidate question is obtained as follows: $Q6 = 0.8$, $Q4 = 0.8$, $Q2 = 0.73$, $Q1 = 0.75$, $Q3 = 0.4$, $Q5 = 0.35$, and $Q7 = 0.2$. On the other hand, the semantic score for each candidate question based on the social network analysis is obtained as follows: $Q6 = 78.54$, $Q7 = 70.23$, $Q4 = 65.23$, $Q1 = 55.26$, $Q2 = 52.31$, $Q3 = 45.11$,
30 and $Q5 = 40.37$. The semantic and similarity scores for each candidate question are normalized to produce a list of ranked candidate questions.

35 The list of ranked candidate questions is then transmitted to the query processor (160) to search and retrieve relevant results based on the ranked candidate questions. For instance, the query processor (160) retrieves the answer or

result for Q6 as “smoking, hereditary traits, working with chemicals, etc. are common factors contributing to cancer” The relevant results are transmitted and displayed as the query results to the user device (10).

- 5 While embodiments of the invention have been illustrated and described, it is not intended that these embodiments illustrate and describe all possible forms of the invention. Rather, the words used in the specifications are words of description rather than limitation and various changes may be made without departing from the scope of the invention.

CLAIMS

1. A semantic query system (100) comprising:

a query manager (110) configured to control activities of the system (100),

5 a context identifier (120) configured to identify at least one context related to a query, and

a query processor (160) configured for retrieving relevant information;

wherein said semantic query system (100) is characterised in that:

10 said semantic query system (100) further includes:

a question generator (130) configured to generate at least one candidate question based on the at least one context,

a query mapper (140) configured to perform social network analysis on the at least one candidate question generated by said question generator (130) with respect to the query, and

15 a ranking component (150) configured to produce a list of ranked candidate question generated by said question generator (130);

20 and wherein said query processor (160) is configured for retrieving relevant information based on the list of ranked candidate question generated by said question generator (130);

25 and wherein said context identifier (120), said question generator (130), said query mapper (140), said ranking component (150) and said query processor (160) are connected to said query manager (110)

2. A method for processing a query is characterised by the steps of:

30 a) receiving a query;

b) identifying at least one context related to the query;

c) generating at least one candidate question based on the at least one context;

d) mapping the query to the at least one candidate question to produce a list of ranked candidate question; and

- e) retrieving and ranking relevant information based on the list of ranked candidate question.

3. The method as claimed in claim 2, wherein mapping the query to the at least one candidate question includes the steps of:
- a) performing a syntactic analysis between the query and the at least one candidate question;
 - b) performing a social network analysis on the at least one candidate question with respect to the query;
 - c) determining a semantic score, similarity score, and confidence score for each candidate question;
 - d) normalizing the semantic score and similarity score for each candidate question; and
 - e) producing a list of ranked candidate question by ranking the at least one candidate question based on the normalized semantic and similarity scores.
4. The method as claimed in claim 3, wherein performing the social network analysis on the at least one candidate question with respect to the query includes the steps of:
- a) constructing a social network based on co-occurrence of concepts or terms in the at least one candidate question;
 - b) determining relevance value between the co-occurred terms for each term; and
 - c) computing centrality value for each term.

1/2

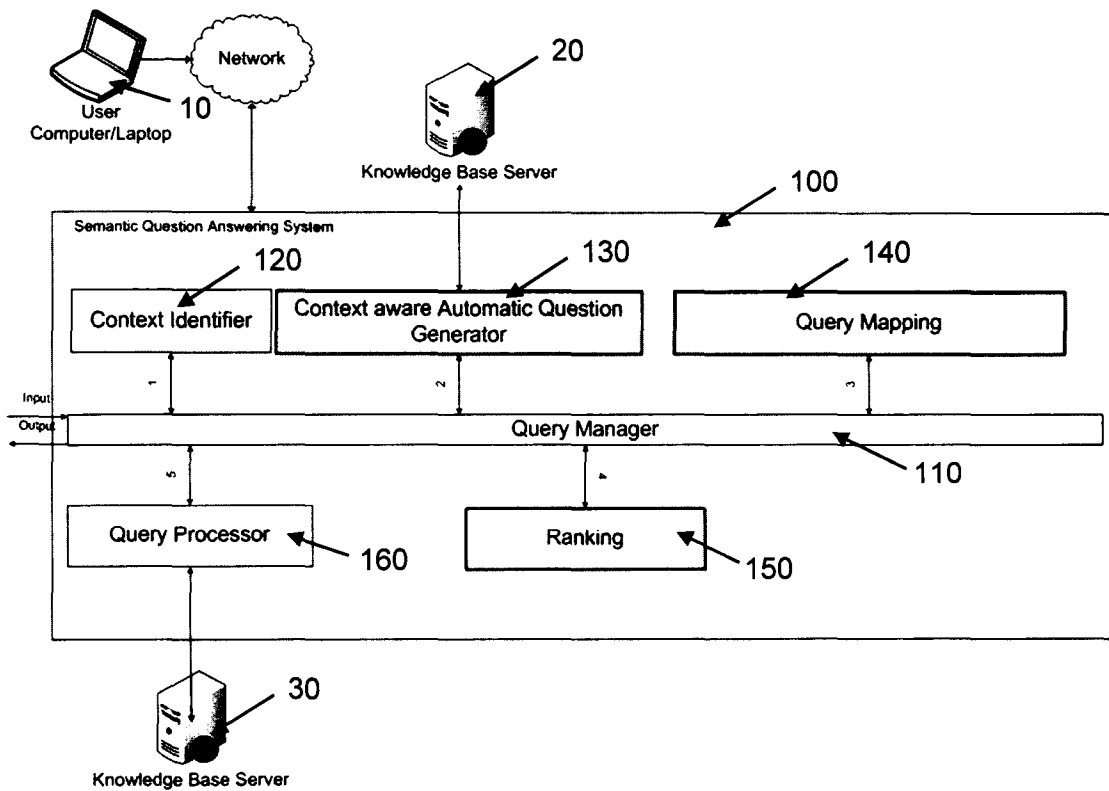


FIG. 1

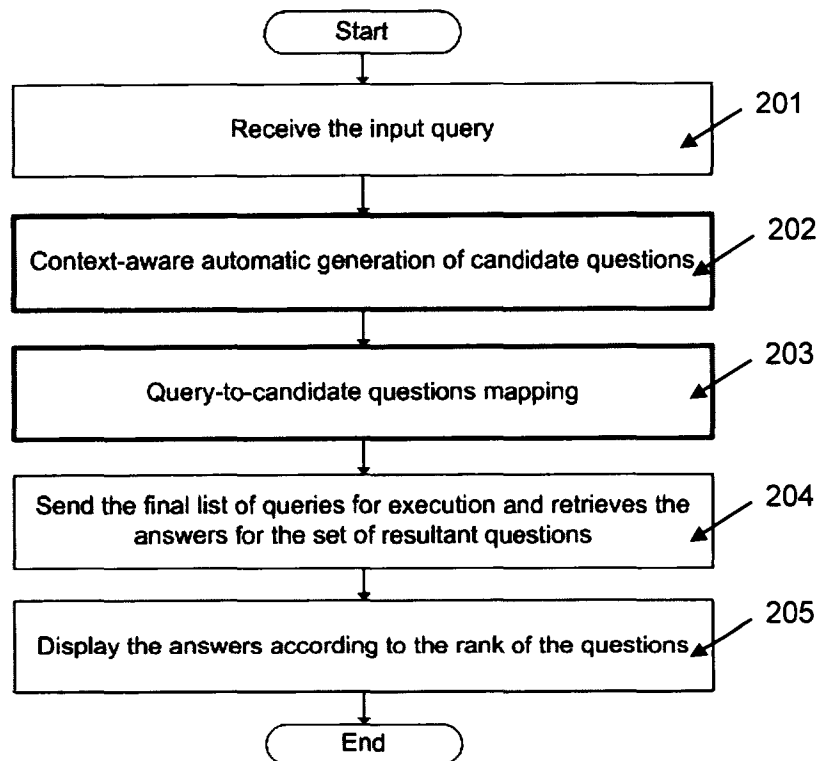


FIG. 2

2/2

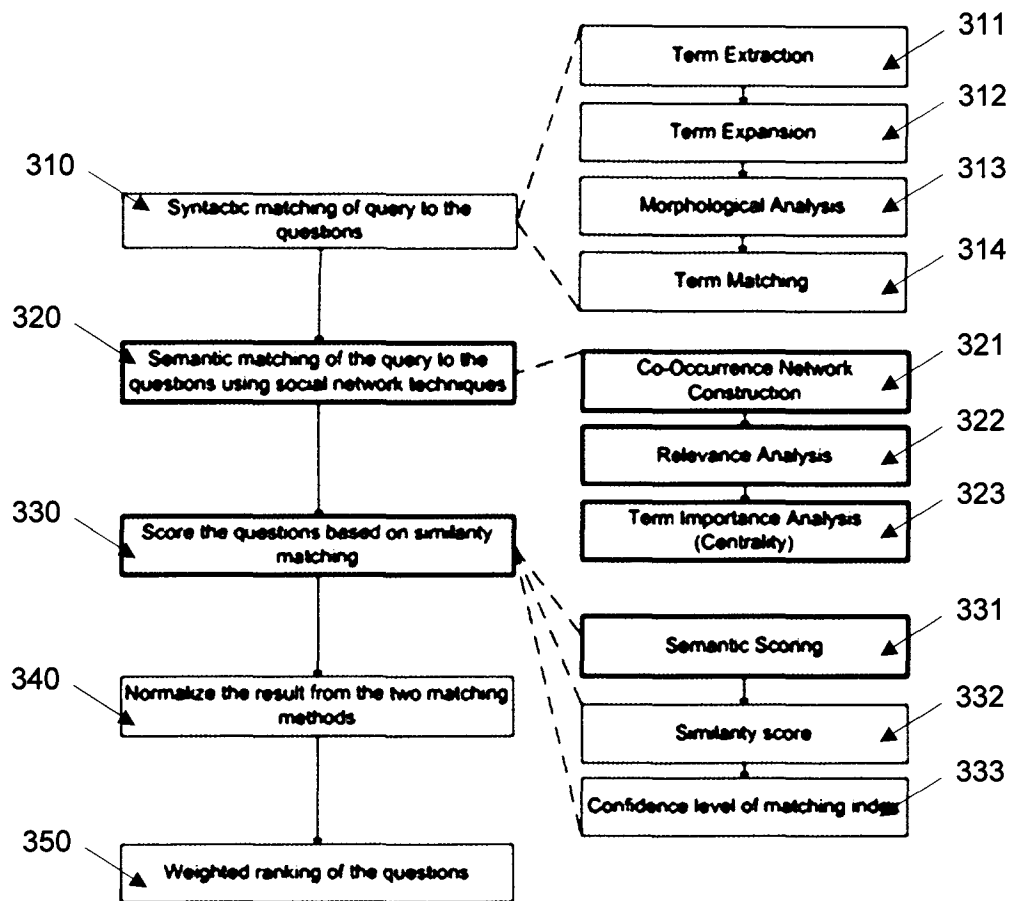


FIG. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/MY2013/000254A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/30
ADD.

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)
G06F

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)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	YUNBO CAO ET AL: "Recommending questions using the mdl-based tree cut model", PROCEEDING OF THE 17TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB , WWW '08, 21 April 2008 (2008-04-21), page 81, XP055112023, New York, New York, USA DOI: 10.1145/1367497.1367509 ISBN: 978-1-60-558085-2 the whole document ----- -/--	1-4



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

4 April 2014

Date of mailing of the international search report

14/04/2014

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Nazzaro, Antonio

INTERNATIONAL SEARCH REPORT

International application No

PCT/MY2013/000254

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	HUANHUAN CAO ET AL: "Context-aware query suggestion by mining click-through and session data", PROCEEDING OF THE 14TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, KDD 08, 24 August 2008 (2008-08-24), page 875, XP055112020, New York, New York, USA DOI: 10.1145/1401890.1401995 ISBN: 978-1-60-558193-4 abstract; figure 1 page 878, right-hand column, paragraph 5.1 - page 880, left-hand column, paragraph 5.3 page 881, paragraph 6.2 -----	1-4
X	APOSTOL (PAUL) NATSEV ET AL: "Semantic concept-based query expansion and re-ranking for multimedia retrieval", PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON MULTIMEDIA , MULTIMEDIA '07, 23 September 2007 (2007-09-23), page 991, XP055035435, New York, New York, USA DOI: 10.1145/1291233.1291448 ISBN: 978-1-59-593702-5	1,2
Y	abstract; figure 1 paragraph [03.2] - paragraph [04.3] -----	3,4
X	US 2009/171929 A1 (JING FENG [CN] ET AL) 2 July 2009 (2009-07-02)	1,2
Y	abstract; figures 2-3 paragraph [0040] - paragraph [0049] -----	3,4

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/MY2013/000254

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2009171929	A1	02-07-2009	NONE
