



(51) International Patent Classification:
G06Q 10/00 (2006.01)

(21) International Application Number:
PCT/EP2011/054797

(22) International Filing Date:
29 March 2011 (29.03.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/662,145 31 March 2010 (31.03.2010) US

(71) Applicant (for all designated States except US): **TECH-
NISCHE UNIVERSITÄT BERLIN** [DE/DE]; Straße
des 17. Juni 135, 10623 Berlin (DE).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **JUGEL, Leo**
[DE/DE]; Defreggerstraße 16, 12435 Berlin (DE).
BRAUN, Mikio [DE/DE]; Beckerstraße 11, 12157 Berlin
(DE). **MÜLLER, Klaus-Robert** [DE/DE]; Arno-Holz-
Straße 12 f, 12165 Berlin (DE).

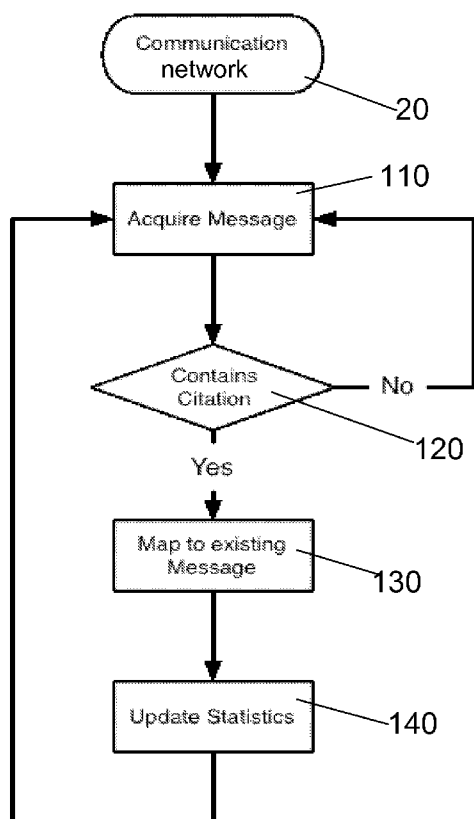
(74) Agent: **FISCHER, Uwe**; Moritzstraße 22, 13597 Berlin
(DE).

(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,

[Continued on next page]

(54) Title: METHOD AND SYSTEM FOR ANALYZING MESSAGES

Fig. 2



(57) Abstract: The invention relates to a method and system for analyzing messages transmitted in a communication network. An embodiment of the method comprises the steps of: determining an information-related citation index indicating how often information comprised in a message has been forwarded in consecutive messages to other users of the communication network, and providing an analysis result based on the information-related citation index.



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 (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, ML, MR, NE, SN, TD, TG).

Published:

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

Description

Method and system for analyzing messages

5 Background of the invention

The invention relates to a method and system for analyzing messages transmitted in a communication network.

Current instances of communication networks are for example
10 internet platforms like twitter (<http://twitter.com>), or facebook (<http://facebook.com>). On these sites, up to several hundred million users communicate through the exchange of status messages and direct messages.

15 A highly relevant problem in the analysis of such communication networks is the assessment of the influence of a user, and to extract emerging trends out of the whole body of ongoing communication and make them available for further use.

20 Objective of the present invention

Accordingly, the objective of the present invention is to provide a method and system for analyzing a communication network in order to determine the influence of users of the communication network.

25

Another objective of the present invention is to provide a method and system for analyzing a communication network in order to extract trends in ongoing communication.

30 Brief summary of the invention

An embodiment of the invention relates to a method for analyzing messages transmitted in a communication network wherein an information-related citation index is determined.

Said information-related citation index indicates how often information comprised in a message has been forwarded in consecutive messages to other users of the communication network. Then, the information-related citation index may be
5 analyzed in order to determine the influence of users and/or to extract trends.

The information may be considered to be forwarded if the respective message containing said information has been forwarded unchanged. However, according to a preferred embodiment of the invention, a text analysis is carried out in order to identify a predefined information and in order to determine whether the predefined information has been forwarded or not. A text analysis allows identifying information in
10 messages even if the message has been altered. In other words, the step of determining the impact of a particular user preferably includes at least one bibliometric procedure being based on a citation analysis and a content analysis of transmitted information.

Furthermore, by evaluating the information-related citation index of information sent by a particular user, an impact value indicating the impact of said particular user on the communication network may be determined. Preferably, the
20 evaluation of information is weighted taking the information's age into account. E. g., recent information is assigned a larger weight than older information. Alternatively, only recent information may be evaluated, and older information may be discarded. For instance, information may be considered "recent", if it is younger than a predefined age.
25
30

According to a preferred embodiment the bibliometric procedure includes a Hirsch-analysis, e. g. the information-related citation index may be Hirsch-index.

5 Another preferred embodiment comprises the steps of: analyzing a plurality of messages sent by said particular user, determining the information-related citation index of each information contained in said plurality of messages, determining an aggregating information-related statistics measuring
10 the spread of the citation histogram, and determining the user's impact value based on said weighted average information-related citation index value.

The method as described above may be used to provide lists
15 indicating users according to their impact on the communication network. To this end, an embodiment of the invention also allows for inputting a search query concerning user impact, and outputting users and/or their impact values in a ranked list, wherein the ranking corresponds to the users'
20 impact values.

Another preferred embodiment of the invention relates to a method for determining a relevance value for messages. For instance, the method may comprise the step of determining the
25 information-related citation index of information comprised in a particular message. This allows determining a relevance value which indicates the relevance of said particular message.

30 The relevance value may also consider the age of each message. Preferably younger messages are considered more relevant for the communication network than older messages.

The step of determining the relevance of a particular message may also include searching for one or more keywords in said particular message and determining a match factor indicating to what extent those keywords exist in said message. The calculation of the relevance value may include this match factor.

The relevance value of messages may also take the impact value of the sending user into account. According to this embodiment relevance values of messages are linked with the impact value of the sending users.

The method may also be used to provide message lists which indicate the impact of messages on the communication network. Thus, an embodiment of the invention may include the steps of inputting a search query concerning message relevance, and outputting messages and/or their relevance values in a ranked list, wherein the ranking corresponds to the messages' relevance values.

In order to determine relevant messages concerning a particular topic, an embodiment of the invention provides for inputting a search query identifying one or more keywords, and outputting all messages comprising at least one of those keywords, in a ranked list, wherein the ranking reflects the messages' relevance values.

Another embodiment of the invention allows generating a user specific data set characterizing topics preferred by the corresponding user. To this end, the method may comprise the steps of identifying characteristic words contained in information sent by the user. Then a weight factor may be assigned for each characteristic word, wherein the weight factor indi-

cates the occurrence of the characteristic words in the information sent by the user. The weight factor may also be weighted for each characteristic word with the information-related citation index corresponding to the information which
5 contains the respective word.

The method as described above may be used to provide a search tool for user specific data sets. To this end, an embodiment of the invention may also include the steps of inputting a
10 search query concerning a user specific data set, and outputting the user specific data set as a search result.

Furthermore, the method may be directed to evaluate neighborhood relations between users. For example, user specific data
15 sets may be generated for different users, and the user specific data sets may be compared. This allows calculating a neighborhood value indicating the similarity of communicated topics by said different users.

20 In order to identify neighborhood relations in a listed form, the method may comprise the steps of inputting a search query concerning neighborhood relations with respect to a particular user, and outputting neighbors in a ranked list, wherein the ranking corresponds to the neighborhood values.

25 Additionally or alternatively, the method may comprise the steps of inputting a search query identifying one or more keywords, and outputting neighbors in a ranked list, wherein the ranking reflects the match between said keywords and the
30 topics listed in the user specific data sets.

Additionally, the method may comprise the steps of inputting a search query concerning communicated topics, and outputting

neighbors and/or neighborhood values indicating the similarity of communicated topics.

The invention also relates to a system for analyzing messages transmitted in a communication network. An embodiment of the invention may comprise means for determining an information-related citation index indicating how often an information comprised in a message has been forwarded in consecutive messages to other users of the communication network, and means for storing information-related citation indices.

Brief description of the drawings

In order that the manner in which the above-recited and other advantages of the invention are obtained will be readily understood, a more particular description of the invention briefly described above will be rendered by reference to specific embodiments thereof which are illustrated in the appended drawings. Understanding that these drawings depict only typical embodiments of the invention and are therefore not to be considered to be limiting of its scope, the invention will be described and explained with additional specificity and detail by the use of the accompanying drawings in which

Figure 1 shows an exemplary embodiment of an inventive system;

Figure 2 shows an exemplary embodiment of a process flow which may be carried out by the system shown in Figure 1; and

Figure 3 shows in an exemplary fashion an embodiment of a trend analysis, wherein the ci-

tation activity related to individual messages is analyzed over a time window.

Detailed description of the preferred embodiment

- 5 The preferred embodiment of the present invention will be best understood by reference to the drawings, wherein identical or comparable parts are designated by the same reference signs throughout.
- 10 It will be readily understood that the process steps of the present invention, as generally described and illustrated in the figures herein, could vary in a wide range of different process steps. Thus, the following more detailed description of the exemplary embodiments of the present invention, as
- 15 represented in Figures 1 - 3 is not intended to limit the scope of the invention, as claimed, but is merely representative of presently preferred embodiments of the invention.

Figure 1 shows an exemplary embodiment of a system 10 for
20 analyzing messages transmitted in a communication network 20.

System 10 comprises a message analyzer 30 which intercepts messages transmitted over the communication network 20. Message analyzer 30 analyzes the messages and extracts information.
25 The extraction of information may be carried out by subjecting the messages to a text analysis or other prior art bibliometric procedures.

For identified information, message analyzer 30 determines an
30 information-related citation index which indicates how often the respective information comprised in a first message has been forwarded in consecutive messages to other users of the

communication network. The information-related citation indices are stored in database 40.

System 10 further comprises a search engine 50 and a trend analyzer 60. Search engine 50 allows a person, who is supervising the communication network 20, to enter search questions in order to extract data concerning

- (a) the impact of users on communication network 20,
- (b) the relevance of messages in communication network 20

10 and/or

- (c) neighborhood relations between users.

Figure 2 shows in an exemplary fashion how system 10 may run. In step 110, the message analyzer 30 intercepts the communication network 20 and acquires messages. The message analyzer analyzes the messages and identifies cited information (step 120). If cited information is found, the information is mapped to existing prior messages containing the same information and a citation index is calculated (step 130). Then, in step 140, database 40 is updated. This allows a person who is supervising communication network 20, to extract updated data related to the impact of users on communication network 20, the relevance of messages in communication network 20 and/or neighborhood relations between users.

25

The exemplary embodiment as described above allows the sustained, exact definition and measurement of the impact of users in communication network 20, for example, the communication hierarchies in companies, or in the armed forces, online communication, for example instant messaging (chat), email, blogs, or social networks, for example facebook and twitter.

30

The exemplary embodiment as described above may also use the actual communication patterns between users, where electronically available, as the basis for impact measurement and further analyses using bibliographic measures. The actual communication may then be analyzed in real-time based on impact estimation in order to detect current trends (that is, dominating topics) within communication network 20, to enable search queries and topic-related navigation, and to improve the search results of existing methods.

The exemplary embodiment as described above is preferably connected through appropriate interfaces to historical or real-time communication data. Communication patterns where one user refers to the messages of another user may then be extracted.

The extracted information may be stored in database 40 and linked to the sending users: the user who created the message, and the user who created the original message. The cited information may be mapped to already identified information. Here, the use of text analysis is advantageous, because messages can also occur in slightly edited or commented forms.

In this way, system 10 may obtain a temporally changing network of interaction between users, which may be the basis for further analysis.

The impact of a user may be computed by analyzing how often each of his messages are cited or commented. The impact may be computed using a bibliometric procedure, for example the so-called Hirsch-index (J. E. Hirsch "An index to quantify an

individual's scientific research output", PNAS November 15, 2005 vol. 102 no. 46 16569-16572).

The analysis as discussed above primarily considers actual
5 interactions and not only the static linkage structure of users. The static linkage structure of users may already be outdated, for example, because users have ceased to participate in the network without deleting their accounts.

10 For the trend analysis, the citation activity related to individual messages is preferably analyzed over a time window (see Figure 3).

Based on impact factors, both for single users but also for
15 single messages, the ranking of results of search queries for certain keywords by relevance can be significantly improved. For example, the number of citations of a message allows to directly derive the relevance of a message, in contrast to existing approaches which only use how recent a message is
20 and whether it contains the keywords.

For improving the search results, a text index may be stored in database 40, which allows to efficiently searching for messages containing certain key words. The results of the
25 initial search query may be ranked based on an adjustable mixture of recency and citation frequency. In order to bring these quantities on the same scale, the age of a message may be rescaled such that the most recent messages have score 1 and infinitely old messages have score 0, with an adjustable
30 decay rate. The citation count is also normalized to the interval 0 to 1 over all found results.

Further, two impact measures may also be used to provide content based neighborhood analysis. E. g., such an analysis may be used to find other users which are communicating over similar topics. Using the impact measures as a weighting, a much more accurate representation of the topics of a user can be estimated because the weighting ensures that only messages which other users found relevant is used to form the topic representation of a user. First, word frequencies may be computed over the whole message corpus. Then, individual weights for single words may be computed, for example based on the term-frequency/inverse-document-frequency measure (tf-idf) (Spärck Jones, Karen, "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation 28 (1): 11-21, 1972). Terms which lie below a certain threshold (and therefore occur in almost all messages) may be removed. Moreover, the weights may be scaled using the citation count of messages normalized by user. The weights can then be used for further similarity search, for example based on scalar products or related measures.

Claims

1. Method for analyzing messages transmitted in a communication network, the method comprising the steps of:

- 5 - determining an information-related citation index indicating how often an information comprised in a message has been forwarded in consecutive messages to other users of the communication network, and
- 10 - providing an analysis result based on the information-related citation index.

2. Method of claim 1, wherein the information is considered to be forwarded if the respective message containing said information has been forwarded

15 unchanged.

3. Method of claim 1, wherein messages are subjected to a text analysis in order to determine whether they include a specific information.

20

4. Method of claim 1, further comprising the step of: determining an impact value indicating the impact of a particular user on said communication network by evaluating the information-related citation index of at least one information

25 sent by said particular user.

5. Method of claim 1, further comprising the step of: determining an impact value indicating the impact of a particular user on said communication network by evaluating the information-related citation index of recent information sent

30 by said particular user, wherein information is considered recent, if it is younger than a predefined age.

6. Method of claim 4, wherein the step of determining the impact of the particular user includes a bibliometric procedure.

5 7. Method of claim 6, wherein said bibliometric procedure is a Hirsch-analysis and the information-related citation index is a Hirsch-index.

8. Method of claim 4, further comprising the steps of:

- 10 - analyzing a plurality of messages sent by said particular user,
- determining the information-related citation index of each information contained in said plurality of messages,
- determining an aggregating information-related statistics
15 measuring the spread of the citation histogram, and
- determining the user's impact value based on said weighted average information-related citation index value.

9. Method of claim 4, further comprising the steps of:

- 20 - inputting a search query concerning user impact, and
- outputting users and/or their impact values in a ranked list, wherein the ranking corresponds to the users' impact values.

25 10. Method of claim 1, further comprising the steps of:

- determining the information-related citation index of information comprised in a particular message, and
- determining a relevance value indicating the relevance of said particular message based on the determined information-related citation index of the information comprised
30 therein.

11. Method of claim 10, further comprising the steps of:

- determining the age of said particular message, and
- determining said relevance value further based on the age of said particular message.

5 12. Method of claim 10, further comprising the steps of:

- searching for one or more keywords in said particular message and determining a match factor indicating to what extent those keywords exist in said message, and
- determining said relevance value further based on said
10 match factor.

13. Method of claim 10, further comprising the steps of:

- inputting a search query concerning message relevance, and
- outputting messages and/or their relevance values in a
15 ranked list, wherein the ranking corresponds to the messages' relevance values.

14. Method of claim 1, wherein at least one user specific data set characterizing topics preferred by the at least one
20 user, is generated, the method comprising the steps of:

- identifying characteristic words contained in information sent by the at least one user,
- assigning a weight factor for each characteristic word, said weight factor indicating the occurrence of the characteristic word in the information sent by the at least
25 one user, and
- weighting said weight factor for each characteristic word with the information-related citation index corresponding to the information which contains the respective word.

30

15. Method of claim 14, further comprising the steps of:

- inputting a search query concerning at least one user specific data set, and

- outputting the at least one user specific data set as a search result.

16. Method of claim 14, further comprising the steps of:

- 5 - generating user specific data sets for at least two different users,
- comparing said user specific data sets, and
- generating a neighborhood value indicating the similarity of communicated topics by said at least two different users.

17. Method of claim 16, further comprising the steps of:

- inputting a search query concerning neighborhood relations with respect to a particular user, and
- 15 - outputting neighbors in a ranked list, wherein the ranking corresponds to the neighborhood values.

18. Method of claim 16, further comprising the steps of:

- inputting a search query concerning communicated topics,
- 20 and
- outputting neighbors and/or neighborhood values indicating the similarity of communicated topics.

19. System for analyzing messages transmitted in a communication network, the system comprising:

- 25 - means for determining an information-related citation index indicating how often an information comprised in a message has been forwarded in consecutive messages to other users of the communication network, and
- 30 - means for storing information-related citation indices.

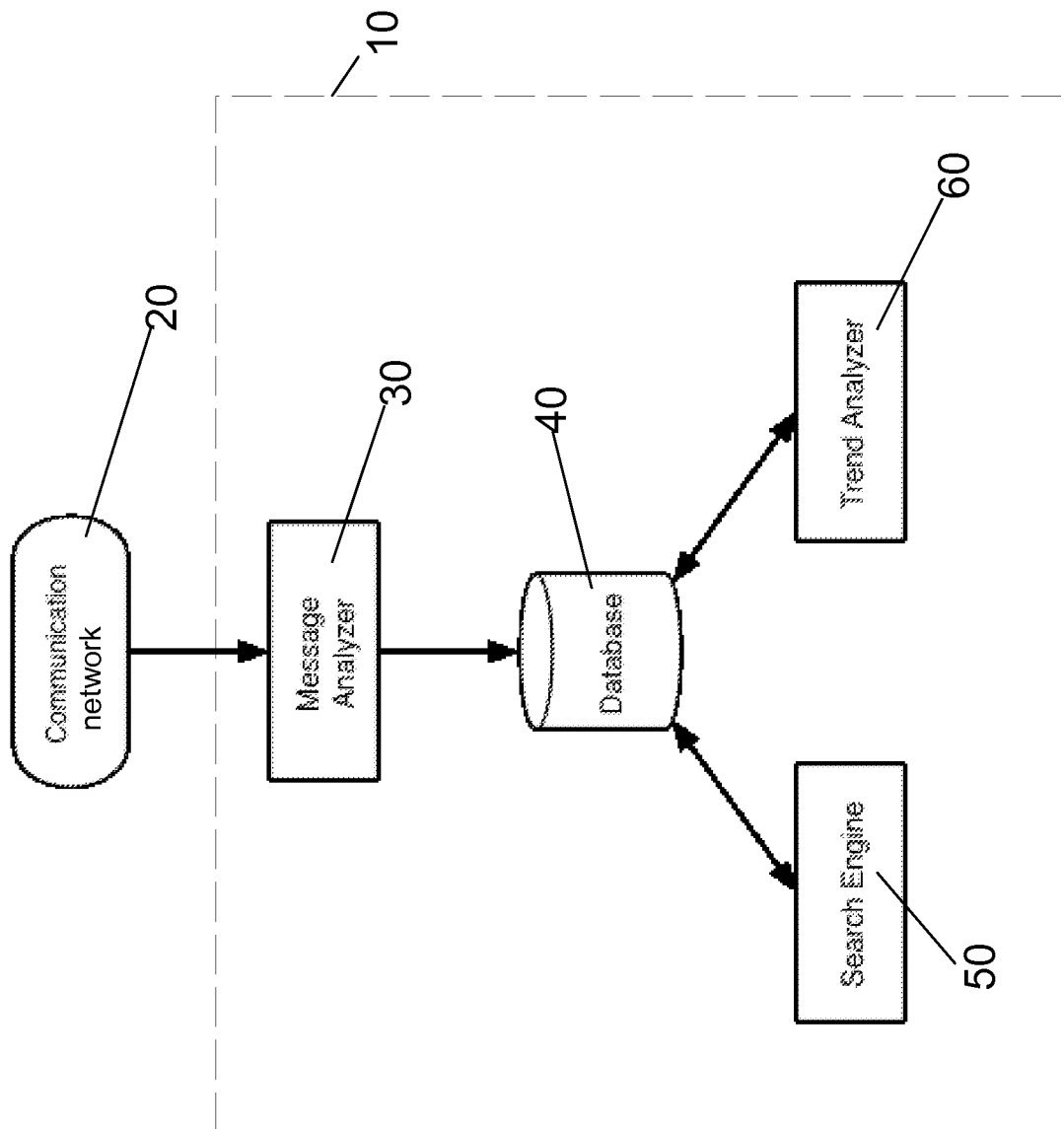


Fig. 1

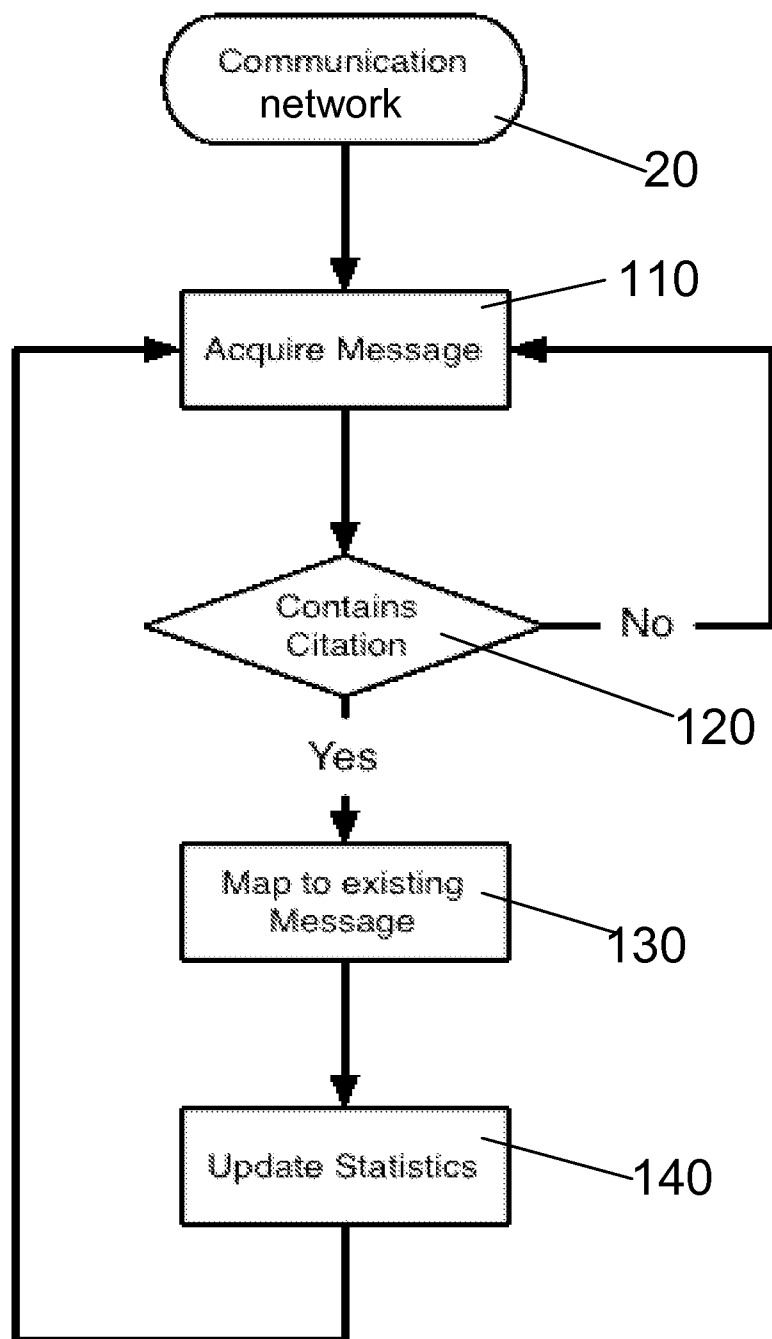


Fig. 2

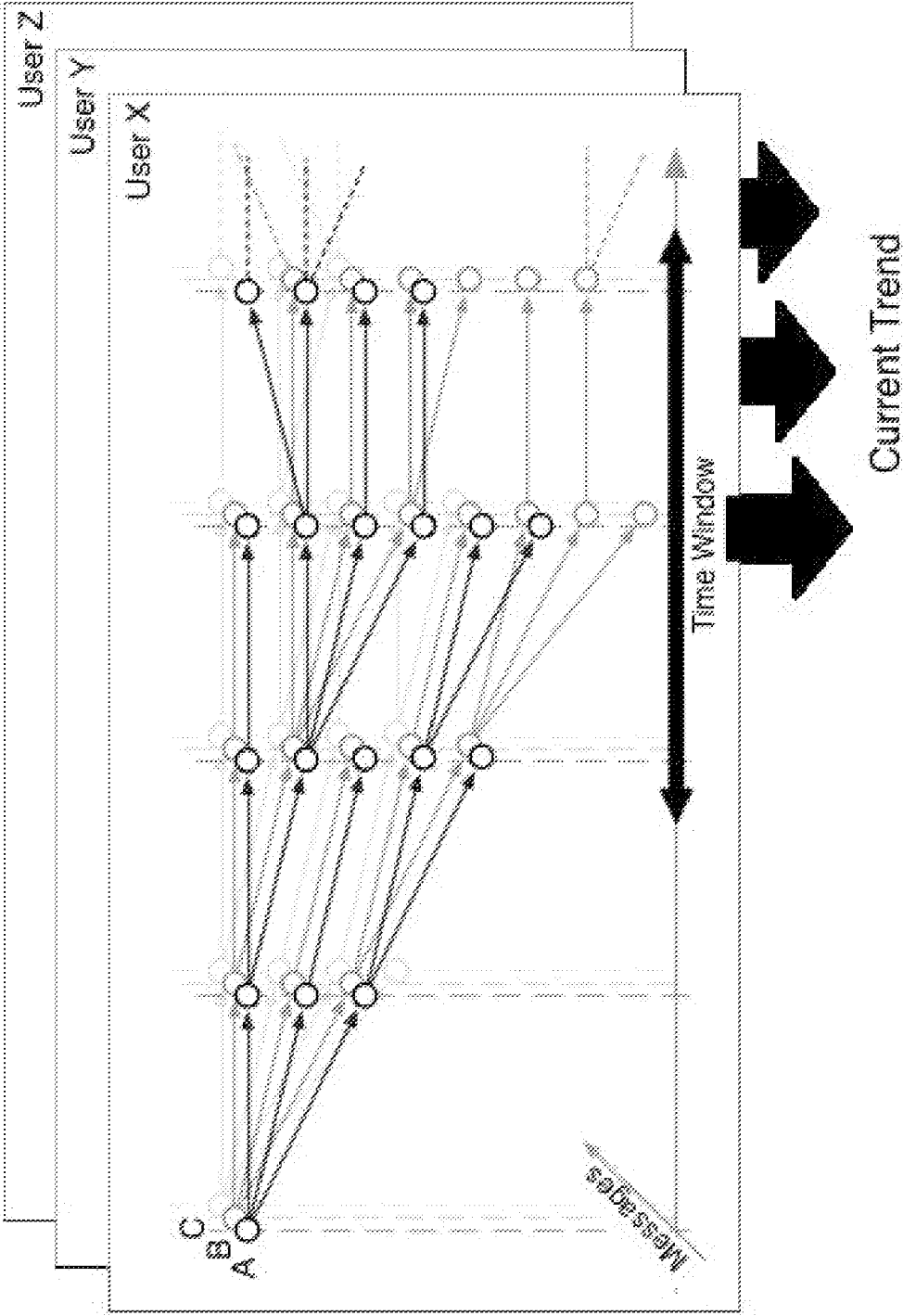


Fig. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2011/054797

A. CLASSIFICATION OF SUBJECT MATTER INV. G06Q10/00 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) G06Q		
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 practical, search terms used) EP0-Internal		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JAGAN SANKARANARAYANAN ET AL: "TwitterStand", PROCEEDINGS OF THE 17TH ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, GIS '09, 1 January 2009 (2009-01-01), page 42, XP55001896, New York, New York, USA DOI: 10.1145/1653771.1653781 ISBN: 978-1-60-558649-6 the whole document ----- -/--	1-19
☒ 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 document 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 July 2011		Date of mailing of the international search report 19/07/2011
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 Hauck, Rainer

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2011/054797

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	J. E. HIRSCH: "An index to quantify an individual's scientific research output", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, vol. 102, no. 46, 15 November 2005 (2005-11-15), pages 16569-16572, XP55001897, ISSN: 0027-8424, DOI: 10.1073/pnas.0507655102 cited in the application the whole document -----	7