



- (51) International Patent Classification: Not classified
- (21) International Application Number:
PCT/TR2015/050228
- (22) International Filing Date:
4 December 2015 (04.12.2015)
- (25) Filing Language: Turkish
- (26) Publication Language: English
- (30) Priority Data:
2014/15332 17 December 2014 (17.12.2014) TR
- (71) Applicant: **BOGAZICI UNIVERSITESI** [TR/TR];
Bogazici Universitesi Guney Kampus, Bebek, Istanbul
(TR).
- (72) Inventors: **BINGOL, Osman Haluk**; Bilgisayar
Muhendisligi Bolumu Bogazici Universitesi 2 Bebek,
34342 Istanbul (TR). **DOSLU, Metin**; Fetih Mahallesi
Libadiye Caddesi Duzgun Sokak Tahrali Sitesi Yildiz
Apartmani A Blok D:24, 34704 Atasehir/Istanbul (TR).
- (74) Agent: **ANKARA PATENT BUREAU LIMITED**; Be-
stekar Sokak No:10, Kavaklidere, 06680 Ankara (TR).

- (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, 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, ST, 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).

Published:

- without international search report and to be republished
upon receipt of that report (Rule 48.2(g))

(54) Title: CONTENT SENSITIVE DOCUMENT RANKING METHOD BY ANALYZING THE CITATION CONTEXTS

(57) Abstract: This invention relates to a method which provides, showing as well the relevant documents to the user even if the said documents that are closely related to the subject do not contain the keywords that are entered for search in their content, by analyzing citation contexts of every document in a data pool containing documents that are citing or linking to at least one document. In an alternative embodiment of this method, in the case when the documents cited by using the entered keywords are cited by other documents using other keywords, these said other keywords are considered as similar terms and the search is conducted as well by including these said similar terms.



WO 2016/099422 A2

SPECIFICATIONS
CONTENT SENSITIVE DOCUMENT RANKING METHOD BY
ANALYZING THE CITATION CONTEXTS

5

Field of the Invention

The present invention relates to a method, which prepares the way for finding as well the documents which are related to the subject but which do not include the entered keywords, by searching the entered keywords in the citation contexts of the documents.

Background of the Invention

15 The field of information retrieval is a field founded on the problems of selecting documents, that are related to a query made, from a document pool, and ranking the selected documents with respect to relevance. There are a number of techniques developed on this subject including term frequency - inverse document frequency (tf-idf). Basically, every document is addressed by some terms, meaning that, full-text indexing. Traditionally, every term corresponds to a dimension in a multi-dimensional vector space. Later, documents are represented as points in this space according to the terms they include. They are mapped to a point in this space by means of the terms included in the search. Afterwards, the terms that are "close" to the query are selected and this closeness is measured as distance in the said vector space.

Generally the documents do not have any relation between documents; they are independent from each other. On the other hand, some document types such as web pages or scientific abstracts, by nature, have connections to the others in the forms of hyperlinks or citations.

Google search engine, while using the relevance based on the document content, uses PageRank on the hyperlink network for estimating the rank of the selected document [2]. In the PageRank approach, every document is assigned an importance value called PageRank. PageRank of a document increases as the document gets more links from documents with higher PageRank.

Even though the rank value, when Google is in question, is independent from the query, HITS approach is dependent to the query [6]. For each query, a set of "hubs" and a set of "authorities" are defined. The hypothesis is that good hubs refers to good authorities which contain high quality information and vice versa.

It is claimed that not only the document itself, but also the documents citing it contains information about the document that is cited. The part of the citing document, which contains the citation is called "citation context". It is believed that the citation context contains important information about the document that is cited [5, 8, 9].

In the state of the art, patent document numbered US6457028 B1 discloses gathering the related documents from the documents that are linked to each other, by using the method of co-citation analysis. If a document A is giving links to the documents B and C, then B and C are thought to be relevant to each other. If B and C receive links together not only in A but also in multiple documents, the relevance of them are considered to be strengthened. In this approach, only the condition whether a link is present between the documents is utilized, however, context information that is used when citing is not taken into account.

In the patent document numbered WO2006/001906 A in the background of the invention, a text document is worked on and word groups are formed from the text. These groups are related by a relationship. By this, a network consisting of word groups as nodes and two groups are connected by an edge if they are related is obtained. The nodes of this network are ranked by known techniques such as

PageRank and HITS. By this, word groups are also ranked. This ranking is used in determining the keys that would explain the document, determining the important sentences. In this patent document, a single text document is worked on and the words in the said text is used to obtain a network. However, in the suggested
5 invention, the network is totally different from this. In the suggested invention, there is a plurality of documents which give reference to each other, and in addition to the state of the art operations realized by the words in the document such as finding keywords and abstracting the text, the reference context of the referring documents is used. By this, a network is formed, however, in the formed
10 network, the document itself is shown by a node. A reference given from one document to a second document is shown by an edge. Hence, the obtained network is a directional edge-labeled network. Additionally, the context of the referring document at the referred place is also added to the directional edge as a label of this connection.

15

In the state of the art, patent document numbered US20080071739 A1 focuses on the additional information about the relevant documents selected by the search engine. The search engine selects the documents suiting the terms of the query of the user. It, not only gives the title and the link of the document while showing
20 them to the user, but also tries to give brief information about the document in order to help the user. It, under normal circumstances, compiles this short information from the content of the document. In some cases, a text, to be compiled like this, might not be present inside the selected document, moreover in some cases; no text might be present in the content. In some cases, the search
25 engine might not find the content to return the short information from. In the search engine, additional text information, which might help in this subject, can be gathered from the documents referring to the selected document. This patent suggests a method in this subject. The terms that are present in the place of citation in the referring document to are compiled as explanatory information for
30 the selected document and are presented to the user. However in this document, as opposed to the suggested invention, the terms that are present in the referred

place, meaning that in the context of reference, are not used during the search engine selection. As a result, the documents, that do not contain the keywords that are used while searching, could not be selected by the search engine.

5 In the patent document numbered EP0637805 B1 in the background of the invention, finding the lexical meaning of a word passing in a text is studied. A word having multiple meanings, the same word receiving different affixes by the linguistic rules according to the place it is used are the obstructing factors. Already present techniques are used for stripping the affixes and reducing the
10 word to its principals. Once the principal word is found, in order to infer which of the multiple meaning of the said word is used in the text, the sentence in which the said word passes is also analyzed. By using the context information, it is tried to understand which of the different meanings of a word is. Additionally, multiple word connections with the context of the word are also utilized. If we use one of
15 the given examples, when the term passes like "under the table" gives a totally different meaning than when the term "table" passing in the text is searched by itself. In this patent, close approaches to the approach in the suggested invention are present. There are (i) using the context of the searched word, (ii) using not only the words but also the word groups. On the other hand, the subject here is to
20 infer in which meaning the word passing in the target text is used. However, in the suggested invention, there is no single texts and a word, the meaning of which is to be found in the said text. The content of a text in a group of referring texts, and hence, in order to be found when it is searched, the context in which the reference is given from the other referring texts is used.

25

Objects of the Invention

The object of the invention is to provide a method, which prepares the way for finding the documents which are related to the subject even if they do not contain
30 the entered keywords, by searching the entered keywords in the citation contexts of the documents.

Another object of the invention is to provide a method which enables finding the documents, that are closely related to the subject but can only be retrieved by similar words other than the entered keywords.

5

During the search, a methodology that is suitable for both relevance and significance is needed. An document defines the main contributions of the cited document and uses related keywords when citing that document. This, in turn, gives very valuable information about the relevance of the cited document to the subject in question. Documents which include the related terms and that are heavily cited mean important contributions to the subject in question, and thus, having more citations mean the cited document is more significant.

The citation contexts of a citing document might have numerous possible meanings, it might be off topic or it might bring criticism more than approval. It is hard to determine the content of the citation automatically. However, as a whole, if an article is cited by multiple articles containing the same terms, it receives a kind of collective approval in the field that the term represents. It can infer the cumulative understanding of the crowd related to the cited document, from the cumulative citation contexts of the citing articles.

The biggest contribution of the inventive method is that it can determine important documents about the target subject even though they do not contain a term representing the subject in question. In this method, using the citation contexts helps to find the articles in the following cases especially when it is not possible to locate them using the full text indexing methods:

- In the case that an article presents a new concept and then another person finds another concept over the said first concept, if important articles for the second concept are being searched, articles relating to the first concept are also wanted to be looked at. For example, "Hadoop" is derived from

the Google File System and the "MapReduce" articles of Google. In this case, if someone is looking for important articles about "Hadoop", even if they do not contain the term "Hadoop", the articles about "MapReduce" are also needed to be looked at.

5

- When they are closely related concepts, for example c_1 and c_2 , it is possible for a document to mention c_1 but not c_2 . In the approach of full text, it is not possible to reach documents about c_2 . For example, in complex networks, the concepts "power law" and "small world" are closely related. However some documents mentioning "small world" without mentioning "power law" could be found. In cases like this, when a person looks for important documents about "power law", he should also look at documents about "small world".

10

- 15 In this invention, the words "citation" and "link" are used interchangeably. While articles, laws, patents might be given as examples to the citations between the documents, links established from web pages to other web pages might be given as examples to links. All of the methods disclosed in this invention could be used, as the given examples, for any documents or web page that are citing or linking
- 20 each other.

Detailed Description of the Invention

- Methods in order to fulfill the objects of the present invention are illustrated in the attached figures, where:

25

Figure 1. Citation network of the exemplary pseudo-articles

Figure 2. Citation Networks

Figure 2a. Term labeled citation network $G(A, C)$.

- 30 **Figure 2b.** Term- $\alpha 1$ specific citation network $G_{\alpha 1}(A, C_{\alpha 1})$

Figure 2c. Term- $\alpha 4$ specific citation network $G_{\alpha 4}(A, C_{\alpha 4})$

Figure 2d. Like term citation network $G_{S_{\alpha l}}(A, C_{S_{\alpha l}})$ for set of like terms $S_{\alpha l} = \{\alpha_1, \alpha_4\}$

Figure 2e. Bipartite graph of documents versus terms

Figure 2f. Matrix F of terms versus documents

5 **Figure 3.** Intersecting terms

Figure 4. Steps of the context sensitive search method.

Figure 5. Steps of the method of searching with like terms.

Elements shown in the figures are numbered as follows:

10

100. Context sensitive search method

110. Separating the citation contexts

120. Identifying the meaningful keywords/word groups from the citation contexts

15 **130.** Forming the citation network, as a directed graph, from the citing article to the cited articles

140. Writing the keywords, citing and cited documents in a table T

150. Entering the keywords/word groups to start a search

160. Searching the keywords/word groups in the said table T

20 **170.** Taking the cited documents corresponding to the keywords in a pool of related documents

180. Ranking using a ranking algorithm

200. Method of searching with like terms

25 **210.** Entering the keywords/word groups

220. Inferring the terms which are similar to the entered keywords

230. Conducting a search in the said table T for the first keywords/word groups entered.

240. Conducting a search in the said table T for the inferred similar terms

30 **250.** Forming a relating document pool for all results in the table T corresponding to the entered keywords and similar terms

260. Ranking the documents taken from the document pool using a ranking algorithm

Citation context sensitive search method (100), which provides access to the
5 related documents by the chosen keywords, essentially comprises the steps of;

- extracting the citation contexts of each document, that cites at least one document, in a database containing the documents (110),
- identifying meaningful keywords/word groups from the citation context of each document (120),
- 10 - forming the citation network, an edge-labeled directed graph, where nodes are the documents. There is an edge from the citing article to the cited article and the edge is labeled by the inferred keywords (130),
- constructing a table T (see Table 1) which contains the citing document, the cited document, and the inferred keywords in the citation (140),
- 15 - entering the keywords/word groups to initiate a search (150),
- identifying citing and cited documents by searching the entered keywords in the said table T (160),
- forming a relating document pool by adding the cited documents identified in (160), (170),
- 20 - ranking the documents taken to the related document pool by using any ranking algorithm (180).

In the subject matter of citation context sensitive searching method (100), in order
25 to provide access to the related documents via selected keywords, firstly, a table T consisting of keywords used for citing, along with citing and cited documents based on the keywords that are already present in the citation context, needs to be constructed. Once the said table T is formed, all searches are conducted through this table T.

In order to form the said table T, first of all, the citation context of every document, that cite another document, is extracted (110), and then meaningful keywords/word groups (terms) are inferred from the citation context (120).

- 5 The best articles are cited by numerous articles with relevant terms in the citation context. This, in turn, shows that the cited document is relevant to the subject. For this, in the method subject to the invention, a simple method is followed for determining/defining the citation terms in the citation context. The terms used for explaining the cited article stand close to the citation point. As also shown in
- 10 Figure 1, citation context is the text around the citation marker that is present inside the window, the size of which is predefined. The size of this text can be defined by a specific sentence, word or character number around the citation point. As a result of this, as present in the previous studies, the method subject to the invention also takes the citation context as a window with a fixed size, for
- 15 example 400 characters in length, around the citation point.

- Once citation context is obtained, the next step is the identification of meaningful keywords/word groups (terms) from the citation context of each document (120). A word or a word group, which states something or some concept in a specific
- 20 field is referred to as “term”. Every bigram, that is present in citation context, is used as a definitive term for the cited article. For example, if three articles are cited in the same citation context, every bigram in this content is taken as definitive terms for the said three articles. The number of words, taken as a term, is one or two but it can be expanded for using n-grams of any n. However,
- 25 bigrams are preferred since they are the most used n-grams for explaining specific terms such as “scale free”, “map reduce”, and “preferential attachment”. While this situation helps inferring meaningful terms from the citation context, it eliminates problems such as synonymous words in the case of single words.

- 30 After inferring meaningful terms from the citation context of each document (120), a citation network is formed (130). A citation network is an edge-labeled

directed graph, where nodes are the documents. There is a directed edge from the citing article to the cited article. The edge is labeled by the inferred keywords. An edge in a citation network carries a lot more information than merely a binary relationship. The terms that the citing author used for explaining the cited document could be taken from the citation context.

In Figure 1, an example of a small citation network, formed by six pseudo-articles, is given. The citation contexts are underlined, cited articles indicated in brackets, such as “[2]”, together with the terms taken from the citation context are emphasized by bold font, such as “term-1”.

For example, the corresponding term labeled citation network of Figure 1 is given in Figure 2(a), where the set of vertices is $A = \{a_1, a_2, \dots, a_{|A|}\}$, and $T = \{\alpha_1, \alpha_2, \dots, \alpha_{|T|}\}$ is the set of terms. The directed edge from a_1 to a_3 is labeled with terms in set $T_{1,3} = \{\alpha_1, \alpha_4, \alpha_5\}$. The edge a_1, a_2 is not there since there is no such citation, that is $T_{1,2}$ is the empty set. In other words, edges are labeled with the terms used in the citation context of the citing document while citing. The conventions for A and T are given below.

$A = \{a_1, a_2, \dots, a_{|A|}\}$ is the set of all documents. Lower case letters of the Latin alphabet such as $i, j \in A$ are used to denote the members of A .

$T = \{\alpha_1, \alpha_2, \dots, \alpha_{|T|}\}$, is the set of all terms used in all documents in A . Letters of the Greek alphabet such as $\alpha, \beta \in T$ are used to denote the members of T .

25

The edge from i to j is labeled by the terms in T_{ij} . Set $T_{ij} \subset T$ is the set of all terms, which are in at least one citation context of article i citing article j . If no citation is made from article i to article j , then T_{ij} is the empty set. In the situations where article i cites article j , however there are no terms inside the citation context, then T_{ij} is again the empty set. It is also possible that the article i cites

30

article j more than once and a term might be presented in each of these citations. Having used at least once is enough for the term to be in T_{ij} .

More formally, the term labeled citation network $G(A,C)$, shown in Figure 2(a), is a directed graph by $C \subset A \times A$ where the edge is $(i, j) \in C$, if and only if article i contains at least one citation referring to article j . The edge $(i, j) \in C$ is labeled with all the terms inside T_{ij} .

In the exemplary term labeled citation network $G(A,C)$ given in Figure 2(a), the directed edges used for forming citation network between documents a_1, a_2, a_3, a_4, a_5 and a_6 are labeled with the terms $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ and α_5 . Document a_1 cites document a_3 . In the citation context it terms uses α_1, α_4 and α_5 . In the citation context for a_4 , there are terms α_1 and α_2 . Similarly, document a_2 uses terms α_1 for a_4 and α_1 and α_3 for a_6 .

After constructing the citation network (130), table T is constructed T (140). For every citing and cited document pair, there is a row. The terms used in the citation context are inserted into the corresponding row. These terms, though they change according to the content of the cited document, in one example of the invention are “scale free”, “preferential attachment”, “map reduce”. The directed edges used between the citing documents and cited documents are labeled with the terms used in the citation context. An exemplary table T formed for Figure 2(a) is shown below. In the exemplary table T shown below, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ and α_5 each represent a term.

Document a_6	α_2	Document a_1
Cited	Words used	Citing
Document a_4	α_1	Document a_2
documents	in citation context	documents
Document a_6	α_1, α_3	Document a_2
Document a_3	$\alpha_1, \alpha_4, \alpha_5$	Document a_1
Document a_5	α_4, α_5	Document a_3
Document a_4	α_1, α_2	Document a_1

Document a ₆	α_3	Document a ₄
-------------------------	------------	-------------------------

5

Table 1. The table T which is formed for the G(A,C) term labeled citation network given in Figure 2(a). Contains the citing and cited documents and the terms used in the citation context.

10

In the inventive citation context sensitive searching method (100), after forming the table T containing citing and cited documents and relating terms, a content sensitive search can be initiated.

15

In the said method, in order to initiate a search, first of all keywords/word groups/terms of the subject to be searched are entered (150). In the preferred embodiment of the invention, the entered terms are searched in a table T (160).

20 In another embodiment of the invention, for the terms entered for search (for example for α) a term- α specific citation network is formed and after that, citing documents, cited documents and entered α in this citation network are written in a table T and citing documents corresponding to this value α are determined (160). In an exemplary embodiment of the invention, an example of the term- α specific
25 citation network for α_1 and α_4 as entered terms, are shown in Figure 2(b) and Figure 2(c) respectively.

Suppose α_1 is the term/keyword for the search. Term α_1 is searched inside the table T. All the documents corresponding to α_1 in table T are selected and
30 considered as the related document pool (170). Thereby, not only the documents that contain α_1 , documents related to the subject but do not contain α_1 are also

selected. Hence access to all documents closely relating to the subject α_1 are provided.

After gathering all documents relation to α_1 in a document pool, the documents
5 taken from the document pool can be ranked by using any ranking algorithm
(180).

In one embodiment of the invention, for ranking the documents related to the
subject (180), a simple ranking module, which takes a bigram and gives a ranked
10 list of scientific articles in return is used.

So far search related to term is considered (100). Another method (200), which
provides access to the relevant documents by using similar terms, is explained
below. Here not only documents gathered in (100), but also documents related to
15 the terms that are similar to the entered terms are considered. So that documents,
which are closely related to the subject but which do not contain the keywords can
also be reached.

A searching method with similar terms (200), which enables finding the
20 documents, which are closely related to the subject, by using other keywords
along with the chosen keywords, reuses steps (110) through (140) and essentially
comprises the steps of;

- extracting the citation contexts of each document, that cites at least one
document, in a database containing the documents (110),
- 25 - identifying meaningful keywords/word groups from the citation context of
each document (120),
- forming the citation network, an edge-labeled directed graph, where nodes
are the documents. There is an edge from the citing article to the cited
article and the edge is labeled by the inferred keywords (130),
- 30 - constructing a table T (see Table 1) which contains the citing document, the
cited document, and the inferred keywords in the citation (140),

- entering the keywords/word groups to initiate a search (210),
- inferring the terms which are similar to the entered keywords (220),
- as in the case of (160), identifying citing and cited documents in the said table T by searching the entered keywords (230),
- 5 - as in the case of (160), identifying citing and cited documents in the said table T by searching the inferred similar terms (240),
- forming a relating document pool for all results in the table T corresponding to both the entered keywords and the similar terms (250),
- as in the case of (180), ranking the documents taken from the document
- 10 pool using any ranking algorithm (260).

One of the main approaches of the inventive searching method with similar terms (200) is using both the entered and the similar terms in the process of network

15 forming. This helps to expand the selected document set to include documents related to the similar terms.

In scientific publications, one term generally is not sufficient to explain a subject by itself and usage of only a single term is prone to noises because of the natural

20 usage of the language such as synonymous words. For every term, there is a set of articles that contain it. In Figure 3, as can be seen related to the terms "law of force", "scale free" and "preferential attachment", these article sets substantially coincide for some terms.

25 In the searching method with similar terms (200), table T is constructed by means of sequence of (110) through (140). After that, in order to initialize the searching process, keywords/word groups/terms are entered (210) and the terms that are similar to the entered keywords are inferred (220). In principle two terms are similar if they appear together in a considerable number of citation. Given a term,

30 inferring similar terms requires some tools.

In order to infer the terms which are similar to the entered keywords (220), first a term-article matrix is formed. Similar terms for the given term α , is the set of terms which have the article scope which substantially coincides with the article scope of the term α .

5

Term frequencies are related to the articles by a document matrix $F=[f_{\alpha j}]$ which has a size of $|T| \times |A|$. In this matrix, the entry $f_{\alpha j}$ is the count regarding how many articles use the term α in the citation context that cites article j . That is, $f_{\alpha j}$ is the in-degree of article j in term- α graph G_{α} . Therefore F is actually the adjacency matrix of the weighted bipartite graph between the article nodes and term nodes.

10

An example the bipartite graph given in Figure 2(e) is obtained from term labeled citation network as in Figure 2(a). The corresponding related term document matrix F is shown in Figure 2(f).

15

In the inventive searching method with similar terms (200), there are distinguishing terms which are used for especially defining smaller article sets are present as similar terms. Simple term frequency has a problem of assuming every one of the terms to have the same importance, however some terms have very little or no distinguishing power. For example, it is possible for almost the entire citation context of an article collection on the topic “cancer” to contain the term “cancer”. For this, the weights of the terms, which are present in numerous citation contexts, are lowered. In principle, the idea is reducing term frequency weight of a term by a factor that grows with its citation context frequency it appears. Term frequency-inverse document frequency (tf-idf) is a technique which is based on this idea. This method is widely used in information retrieval and text mining and it reflects how important a word is to a document in a collection. For this reason, in the inventive method (200), this technique is used for weighting the term frequencies.

20

25

30

The inverse document frequency for term α is defined by $g(\alpha)$.

$$g(\alpha) = \log \frac{|A|}{\sum_{j=1}^{|A|} \text{sgn}(f_{\alpha_j})}$$

where $\text{sgn}(x)$ is a signum function designed as

5

$$\text{sgn}(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

Afterwards, let us assume that $\mathbf{D} = [d_{\alpha\beta}]$ is a $|T| \times |T|$ diagonal matrix defined with below:

10

$$d_{\alpha\beta} = \begin{cases} g(\alpha), & \alpha = \beta, \\ 0, & \alpha \neq \beta. \end{cases}$$

We define the weighted term document matrix $\mathbf{N} = [n_{\alpha\beta}]$ of size $|T| \times |A|$ with $\mathbf{N} = \mathbf{D} \times \mathbf{F}$.

15

Afterwards, a relationship between the terms is established. α and β are assumed to be the α^{th} and β^{th} row vectors of $\mathbf{N}$ respectively, and the members α and β show the related weighted term frequencies of α and β for the articles inside the data set. In order to learn how much of the scopes of the articles of these terms coincide, comparison of the corresponding row vectors of α and β is realized. For this, in a preferred embodiment of the invention, as a measure towards the power linear correlation between two sample terms, Pearson Correlation Coefficient, which is widely used in the field of science, is used.

20

25 Afterwards, a Sample Pearson Correlation Matrix $\mathbf{P} = [p_{\alpha\beta}]$ of size $|T| \times |T|$ is defined and $\mathbf{P}_{\alpha\beta}$

$$p_{\alpha\beta} = \frac{\sum_{i=1}^{|A|} (\alpha_i - \bar{\alpha})(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^{|A|} (\alpha_i - \bar{\alpha})^2} \sqrt{\sum_{i=1}^{|A|} (\beta_i - \bar{\beta})^2}}$$

is the Sample Pearson Correlation between term α and β where α and β are the α^{th}
 5 β^{th} row vectors of N . The vector $\bar{\alpha}$ is the average of the entrance of the vector α .

The sample Pearson Correlation Coefficient is the measure of the linear correlation between two samples X and Y , and it can give a value between -1 and 1 (including -1 and 1). A value of 1 means that a linear equation defines the relationship between X and Y , and all data points are located on a line where Y
 10 increases with increasing X . A value of -1 means that all data points are located on a line where Y decreases with increasing X . This case is irrelevant for our data set, because, in order to take the value -1 for two terms α and β , they need to be complementary to each other. This is not possible for large article collections. The
 15 value 0 means that there is no linear correlation present between the samples.

For a given term α , the similar term set S_α is defined as $S_\alpha = \{\beta \in T \mid p_{\alpha\beta} > \delta\}$ for some value $0 < \delta < 1$. δ is the cross validation parameter and the value of δ changes between topics. Additionally, similarity point $p_{\alpha\alpha}$ for term α equals to 1.
 20 For this reason, since $\alpha \in S_\alpha$, S_α is not empty.

Weighted citation network which takes term α as basis and which is directed from the term set S_α is defined as follows:

25 The sub graphic $G_{S_\alpha}(A, C_{S_\alpha})$ of $G(A, C)$ is named as the citation network of the set of similar terms, where

$$(i) \quad C_{S_\alpha} = \bigcup_{\beta \in S_\alpha} C_\beta$$

- (ii) the weight of the edge $(i, j) \in C_{S\alpha}$ equals to the sum of weights of the edges combined $w_{ij} = \sum_{(i,j) \in T_{ij} \cap S_{\alpha}} P_{\alpha\beta}$.

For example from Figure 2(a), suppose $S_{\alpha 1} = \{\alpha_1, \alpha_4\}$ is the similarity set for a given term α_1 and δ . Then, the network of similar terms set $G_{S\alpha 1}(A, C_{S\alpha 1})$ for term α_1 is shown in Figure 2(d).

The keywords entered, in the inventive method (200), are searched in table T and related documents are identified (230).

10

In another embodiment of the invention, the citation network of terms similar to given term α is formed in (220). Therefore table T contains the documents related to similar terms to the entered keyword.

15 In another embodiment of the invention, after forming the citation network of similar terms set is formed for a given term α , terms (for example α_1 and α_4) similar to the entered keyword (α), the documents citing using these said similar terms in the citation context and cited documents are written in the table T. Thus, cited documents corresponding to the entered keyword (α) and the terms that are
20 similar to the entered keyword/keywords (α_1 and α_4), in other words documents closely related to the subject are determined.

Thus, documents closely related to the subject can be determined by means of another search in the table T for the terms similar to the entered (240). Thus, the
25 documents related to the entered keywords, and the documents related to similar terms, that is, closely related to the subject, are determined (250).

After the cited documents corresponding to the entered keyword and the similar terms are collected in a data pool (250),

The related documents, in the sense of both entered keywords and the similar terms, taken into the data pool. Then so selected documents are ranked by using any ranking algorithm (260).

CLAIMS

1. A citation context sensitive search method (100) which provides access to the related documents by the chosen keywords **characterized in that** it
5 essentially comprises the steps of;
 - extracting the citation contexts of each document, that cites at least one document, in a database containing the documents (110),
 - identifying meaningful keywords/word groups from the citation context of each document (120),
 - 10 - forming the citation network, an edge-labeled directed graph, where nodes are the documents; there is an edge from the citing article to the cited article and the edge is labeled by the inferred keywords (130),
 - constructing a table T (see Table 1) which contains the citing document, the cited document, and the inferred keywords in the citation (140),
 - 15 - entering the keywords/word groups to initiate a search (150),
 - identifying citing and cited documents by searching the entered keywords in the said table T (160),
 - forming a relating document pool by adding the cited documents identified in (160), (170),
 - 20 - ranking the documents taken to the related document pool by using any ranking algorithm (180).
2. A citation context sensitive search method (100) as in Claim 1 **characterized in that** in the step of “inferring meaningful keywords/word groups from the
25 citation context of each document which are taken into a window (120)” every term in the citation context is inferred.
3. A citation context sensitive search method (100) as in Claim 1 or Claim 2 **characterized in that** in the step of “identifying meaningful keywords/word
30 groups from the citation context of each document (120)” every term in the

citation context is used as a definitive term for the articles that are in the said citation context.

4. A citation context sensitive search method (100) as in any of the claims above
5 **characterized in that** in the step of “identifying meaningful keywords/word groups from the citation context of each document (120)” term can be single words, bigrams or n-grams.
5. A citation context sensitive search method (100) as in Claim 1 **characterized**
10 **in that** in the step of “forming the citation network, an edge-labeled directed graph, where nodes are the documents; there is an edge from the citing article to the cited article and the edge is labeled by the inferred keywords (130)” the edges are labeled by the terms used in the citation context of the citing document while citing.
- 15 6. A citation context sensitive search method (100) as in Claim 1 **characterized in that** in the step of “identifying citing and cited documents by searching the entered keywords in the said table T (160)”, the entered keywords are searched in a table T constructed by using term labeled citation network.
- 20 7. A citation context sensitive search method (100) as in Claim 1 or Claim 6 **characterized in that** in the step of “identifying citing and cited documents by searching the entered keywords in the said table T (160)”, for the keyword α entered for search, a term- α specific citation network is formed and after
25 that, citing documents, cited documents and entered α in this citation network are written in a table T and citing documents corresponding to this value α are determined.
- 30 8. A citation context sensitive search method (100) as in Claim 1 **characterized in that** in the step of “ranking the documents taken to the related document pool by using any ranking algorithm (180)”, a simple ranking module which takes a bigram and gives a ranked list of scientific articles in return is used.

9. A searching method with similar terms (200) which enables finding the documents which are closely related to the subject, by using other keywords along with the chosen keywords, in the case where the same document is referred to by multiple documents and by multiple keywords, **characterized in that** it essentially comprises the steps of;
- extracting the citation contexts of each document, that cites at least one document, in a database containing the documents (110),
 - 10 - identifying meaningful keywords/word groups from the citation context of each document (120),
 - forming the citation network, an edge-labeled directed graph, where nodes are the documents; there is an edge from the citing article to the cited article and the edge is labeled by the inferred keywords (130),
 - 15 - constructing a table T (see Table 1) which contains the citing document, the cited document, and the inferred keywords in the citation (140),
 - entering the keywords/word groups to initiate a search (210),
 - inferring the terms which are similar to the entered keywords (220),
 - as in the case of (160), identifying citing and cited documents in the said
 - 20 table T by searching the entered keywords (230),
 - as in the case of (160), identifying citing and cited documents in the said table T by searching the inferred similar terms (240),
 - forming a relating document pool for all results in the table T corresponding to both the entered keywords and the similar terms (250),
 - 25 - as in the case of (180), ranking the documents taken from the document pool using any ranking algorithm (260).
10. A searching method with similar terms (200) as in Claim 9 **characterized in that** in the step of “inferring the terms which are similar to the entered
- 30 keywords (220)”, the terms with which the other documents cite every document that cites another document, and the documents that are cited in the

citation context using the entered keywords, is investigated and these terms are taken as similar terms.

- 5 **11.** A searching method with similar terms (200) as in Claim 9 or Claim 10
characterized in that in the step of “inferring the terms which are similar to
the entered keywords (220)”, a term matrix $F = [f_{\alpha j}]$ is constructed where the
entry $f_{\alpha j}$ shows the count regarding how many articles use the term α in the
related citation context, in order to cite article j , and F is fundamentally taken
10 from a non-directed weighted bipartite graph between the article nodes and
term nodes.
- 15 **12.** A searching method with similar terms (200) as in any of the Claims 9 to 11
characterized in that in the step of “ inferring the terms which are similar to
the entered keywords (220)”, for finding the distinguishing terms which are
used for especially defining smaller article sets that are present as similar
terms, in order to lower the term weight of a term by a factor increasing with
the frequency of appearance of the term in citation contexts, term frequency-
inverse document frequency technique is used.
- 20 **13.** A searching method with similar terms (200) as in any of the Claims 9 to 12
characterized in that in the step of “inferring the terms which are similar to
the entered keywords (220)”, as a measure towards the power linear
correlation between two sample terms, Pearson Correlation Coefficient is
used.
- 25 **14.** A searching method with similar terms (200) as in Claims 9 characterized in
that the keywords that are entered in the step “as in the case of (160),
identifying citing and cited documents in the said table T by searching the
entered keywords (230)” are searched in a table T in which, the words inside
30 the citation context of every document in an information pool which contains
documents which are citing at least one document, every citing document and
every cited document are present.

15. A searching method with similar terms (200) as in Claims 9 **characterized in that** after forming the citation network of similar terms set for a given term α in the step “as in the case of (160), identifying citing and cited documents in the said table T by searching the entered keywords (230)”, the documents
5 citing using the said α in the citation context and cited documents are written in the table T, and citing documents corresponding to this value α are determined.
- 10 16. A searching method with similar terms (200) as in Claims 9 **characterized in that** the keywords that are entered in the step “ as in the case of (160), identifying citing and cited documents in the said table T by searching the inferred similar terms (240)” are searched in a table T in which, the words
inside the citation context of every document in an information pool which
15 contains documents which are citing at least one document, every citing document and every cited document are present.
17. A searching method with similar terms (200) as in Claims 9 **characterized in that** after forming the citation network of similar terms set for a given term α
20 in the step “ as in the case of (160), identifying citing and cited documents in the said table T by searching the inferred similar terms (240)”, writing the terms similar to this said α , the documents citing using these said similar terms in the citation context and cited documents in the table T and determining the documents corresponding to the entered term α and the terms similar to the
25 this said α .

FIGURE 1

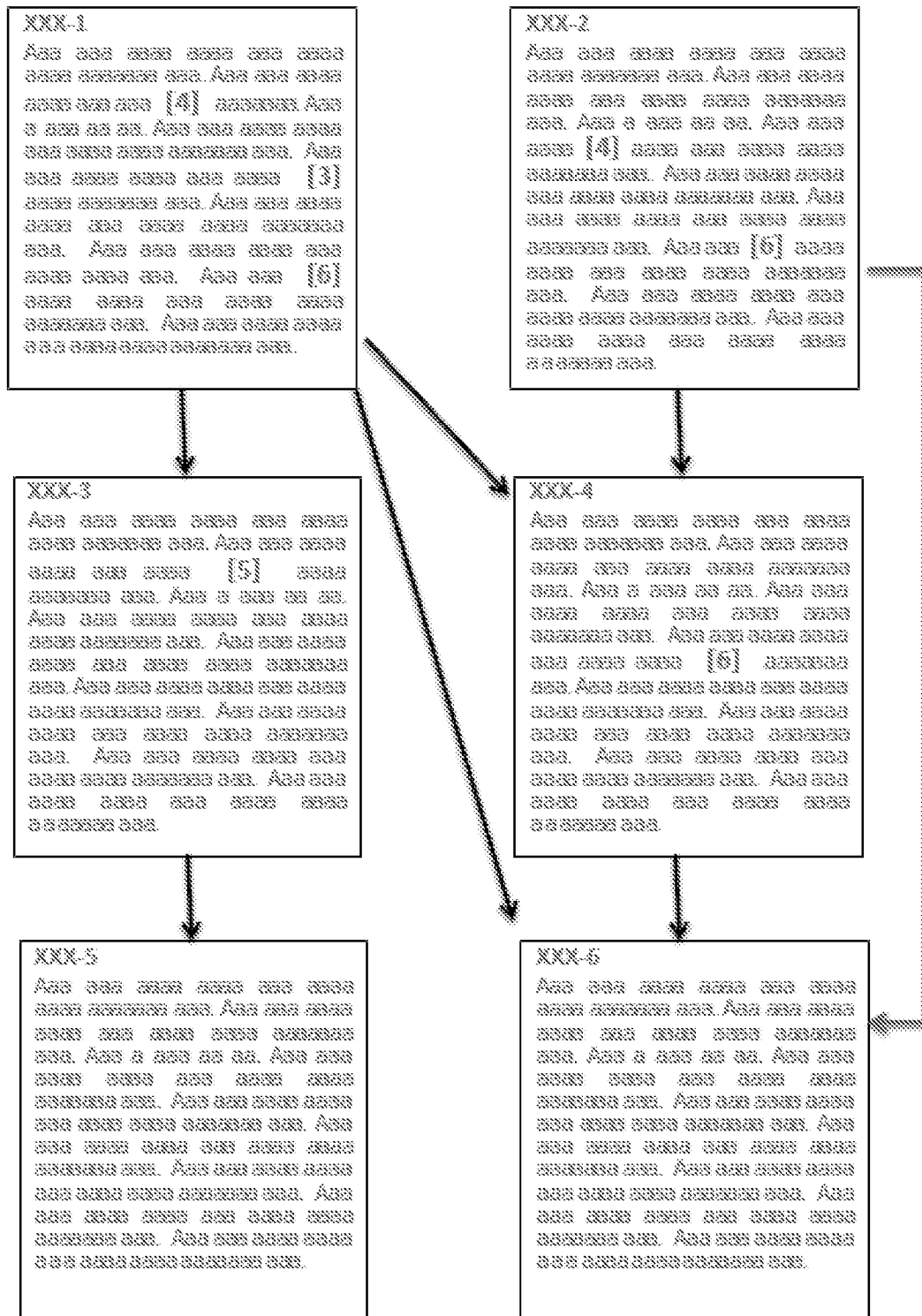


FIGURE 2

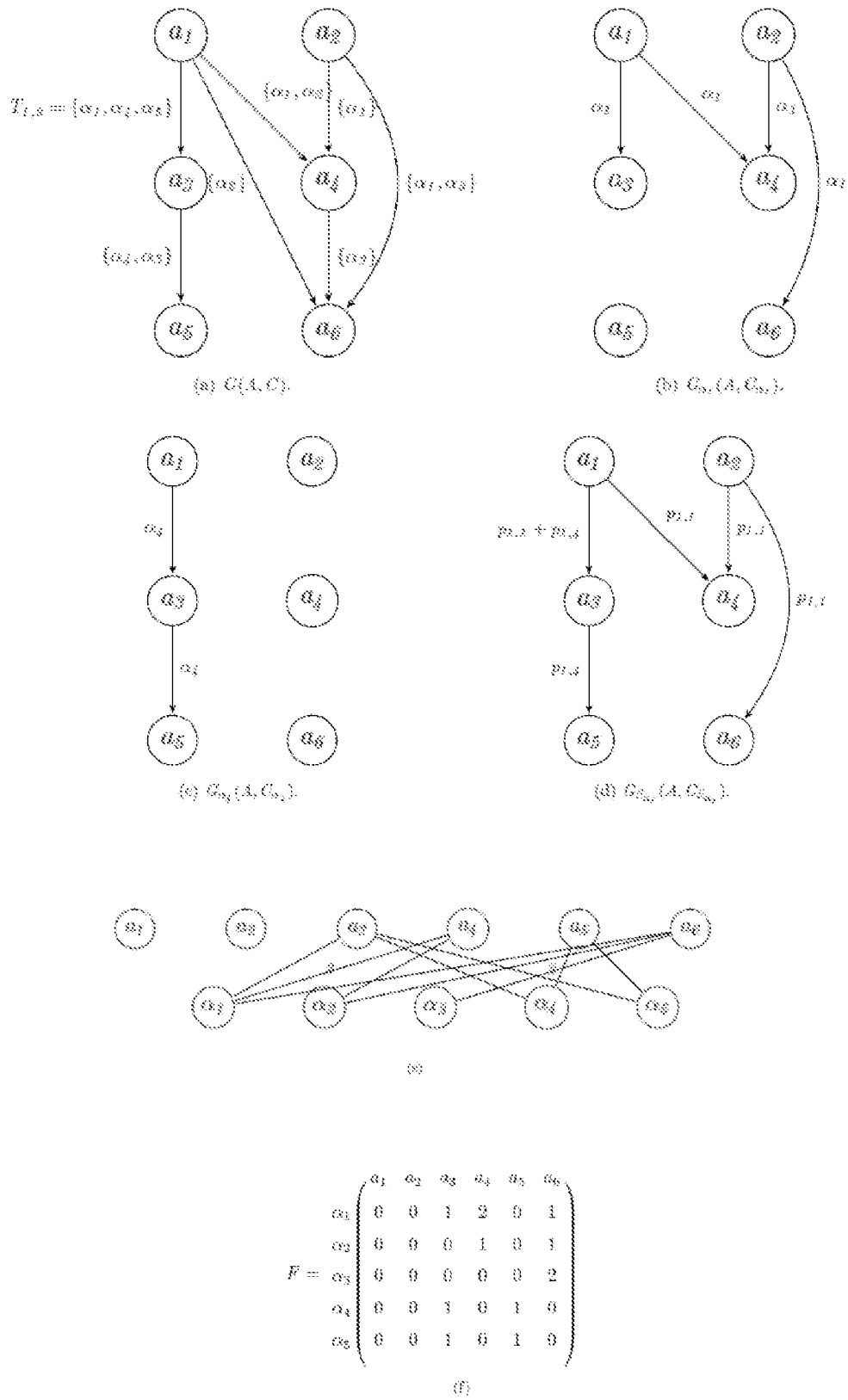


FIGURE 3

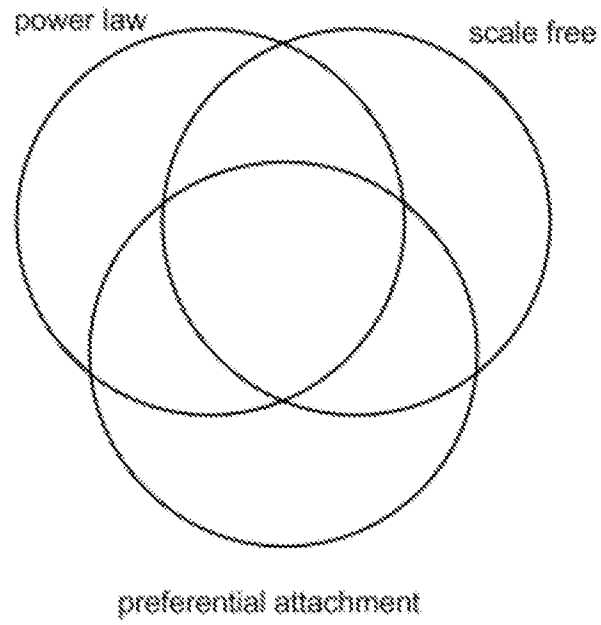


FIGURE 4

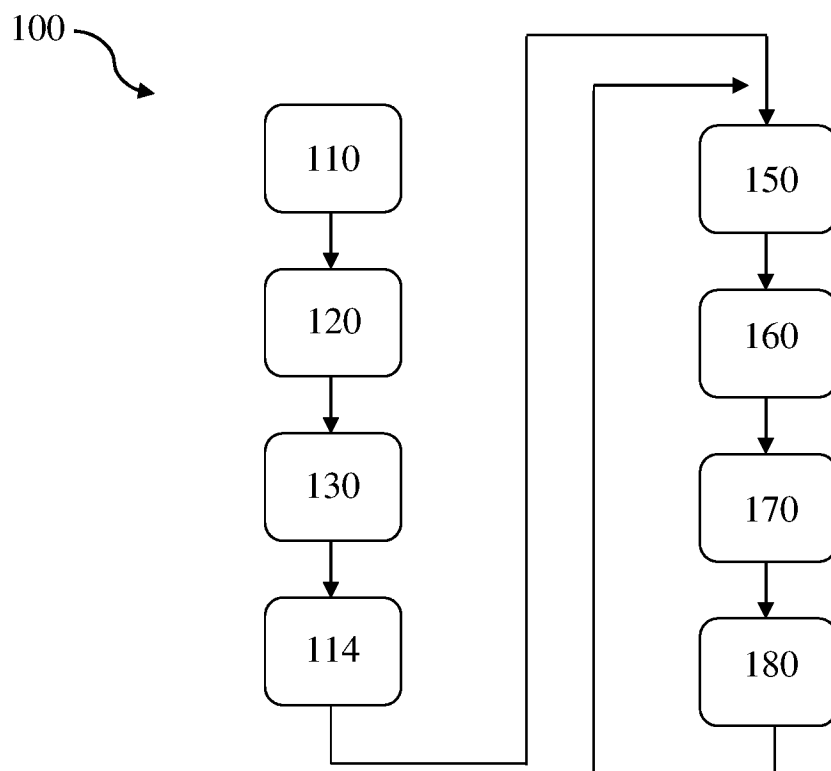


FIGURE 5

