On the information retrieval model by citation analysis

Hakim HARIK[1]; Madjid DAHMANE[2]

[1]Research Centre on Scientific and Technical Information (CERIST),
3 Rue des Frères-Aissou, Ben Aknoun, Algiers, Algeria
hhakim@mail.cerist.dz

[2]Research Centre on Scientific and Technical Information (CERIST),
3 Rue des Frères-Aissou, Ben Aknoun, Algiers, Algeria
mdahmane@wissal.dz

Abstract: In this paper, we propose to study the problem of the information retrieval. A representation model of the scientific production will be presented. Some properties will be released from the structure of the model presented. Using some techniques of the graph theory, we will propose a method of information retrieval containing the citation analysis.

Keywords: citation graph, information retrieval, pagerank, cocitation

1. Introduction

As greater volumes of documents became available on the Internet and data bases, users need more sophisticated tools to locate the information that is relevant to them; this makes urgent the need for effective Information Retrieval Systems.

Information Retrieval Systems is a branch of Computing Science that aims at storing and allowing fast access to a large amount of information. This information can be of any kind: textual, visual, or auditory van Rijsbergen (1979). The goal of an Information Retrieval System (IRS) is to retrieve information considered pertinent to a user's query formally expressed in the system's query language. The effectiveness of an IRS is measured through parameters which reflect the ability of the system to accomplish such goal.

In this paper, we propose to study the problem of the information retrieval by operations research techniques. A model of representation of the scientific production will be presented. Some properties will be released starting from the structure of the model presented in order to propose a method of information retrieval using the citation analysis.

2. Citation analysis

The citation is none other than the relation which binds a citing document and the cited document Garfield (1955), small (1978). Price (1970) specifies more this concept of citation: "If the article A has a bibliographical note using and describing the article B, then A contains a reference to B, and B receives a citation of A ".

The citation analysis is the process by which the impact or the "quality" of an article, an author or a revue is estimated depending on how many times they are mentioned in other works Garfield (1979), Cozzens (1981), Cronin (1981). This analysis requires the setting up of bibliographical data corpus, noted ψ, on which in our case consists the set of references or paper containing in database. This corpus is composed of a set of bibliographical reference. Each bibliographical reference includes a standard description of a document through a number of fields (title, authors, abstract,…).

The model which we use it for modelling structural of the citation is citation graph related to references:

a) Citation Graph related to references

A citation graph related to references $G_r=(V_r, U_r)$ is a directed graph such that:

- $V_r = \Psi$.
- $\forall\ i, j \in V_r$: $(i, j) \in U_r \Leftrightarrow i$ cites j.

The citation graph related to references $G_r=(V_r, U_r)$ is acyclic and through this model, it is possible to identify semantic relationships between an article and the documents cited in it Garfield (1955).

3. A model of information retrieval

Our approach for information retrieval is based on the notion of citation. It is summarized into two steps (figure 1):

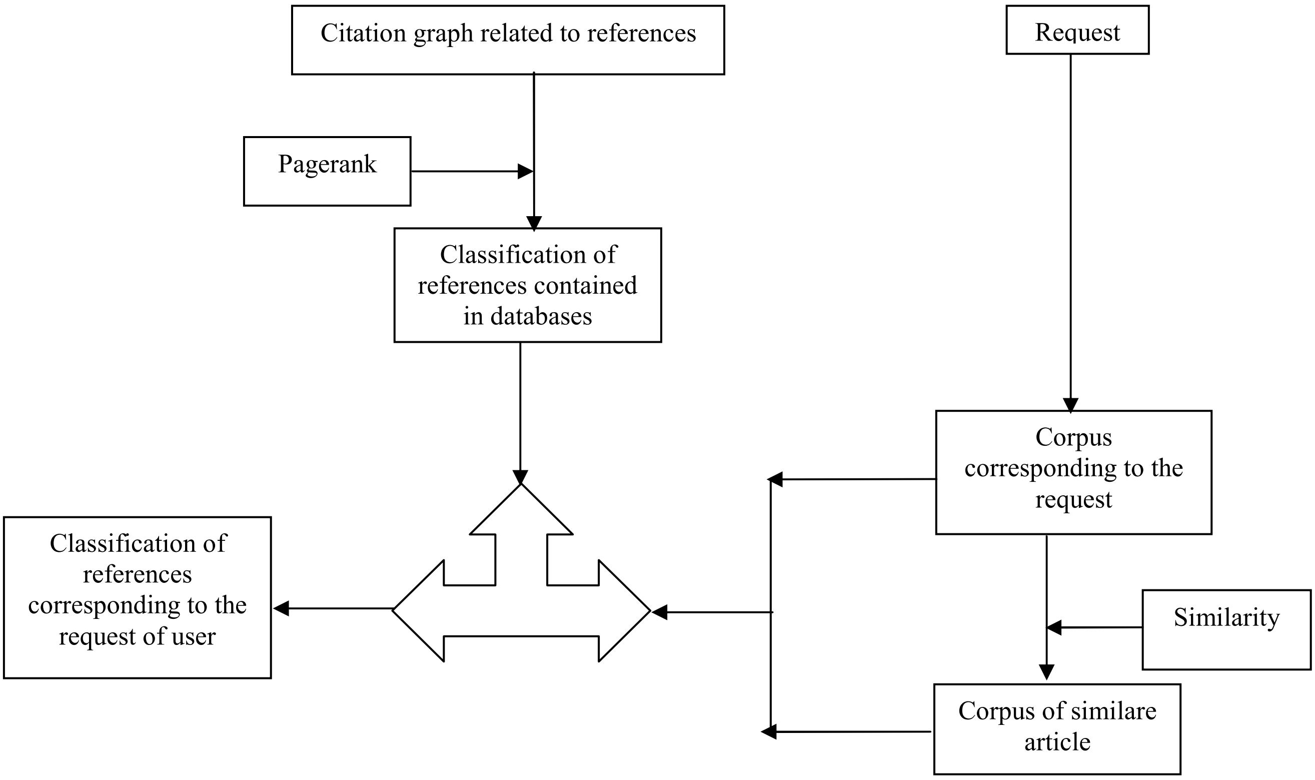
Citation graph related to references
Request
Pagerank
Classification of references contained in databases
Corpus corresponding to the request
Classification of references corresponding to the request of user
Similarity
Corpus of similare article

Figure 1

- **Collect:** The user starts an information retrieval through a request, the system will extract the corpus, noted L, which answers the request one comparing the terms of the request with the terms of the notice of each article in the data base, knowing that the notice is composed of the fields: titrate, author, words key, summary, reference, date. If the term appears in one of the fields, a system adds it in the corpus.

for each article of a corpus, a system seeks the similar articles in data bases using a similarity measure, and we will get corpus of similar article, noted S, of the original corpus L.

a) Similarity

A similarity measure is a function that associates a numeric value with a pair of sequences, with the idea that a higher value indicates greater similarity. In our case the similarity between two documents of data base will be calculated by the analysis of cocitation Marshakova (1973), Small(1973).

The analysis of the cocitation is a bibliometric technique used by the researchers of the information science to chart the intellectual structure of a research field, it consists of the enumeration of the documents which frequently appear in the list of the bibliographical references of the citing document, we propose an indicator, noted ICC (s, t), which measures the cocitation of the references s and t.

$$ICC(s, t)=| K_s \cap K_t |^2 \ / \ | K_s | \times | K_t |$$

Where, $| K_s |$ and $| K_t |$ are the number of citations of articles s and t respectively.

ICC (s, t) indicator is between 0 and 1.

$$ICC(s,t)=1 \ \Leftrightarrow \ \begin{cases} | K_s \cap K_t |= | K_s | \ \Rightarrow K_s \subset K_t \ \ldots\ldots\ldots\ldots\ldots\ldots..(1) \\ K_t \cap K_s|= | Kt | \ \Rightarrow K_t \subset K_s \ \ldots\ldots\ldots\ldots\ldots\ldots \ (2) \end{cases}$$

From (1) and (2) one finds $K_s= K_t$, thus the document s and t have a similarity in the contained bibliographical one.

Hence, the new corpus, noted C, will be the union of the previous corpus L, S. $C=L \cup S$.

- Ranking:

This part consists of ranking of the documents. After to have get a corpus C, it is interesting to rank them from the important one, thus related to the request of user. In our case, we rank all documents of data base then we rank all documents of corpus C from the initial ranking. In order to rank and measure the relative importance of document, we propose PageRank.

a) Pagerank

PageRank is the algorithm used by the Google search engine, originally formulated by Brin and Page (1998). It is a method for computing a ranking for every web page based on the graph of the web. The importance of a web page can be judged by the number of hyperlinks pointing to it from other web pages.

In other terms, the classification of the web is made according to their popularity, a kind of "democratic vote", a bond emitted by a page has towards a page B is compared to a "vote" of has for B. In more the one page receives "votes", in more this page is regarded as important by Google, exactly as the principle of the elections which we all know

As a citation graph related to references has the meaning of a graph of web. So, it's interesting to use pagerank algorithm adaptation for the problem of scientific papers ranking in citation graph related to references.

We define ranking, R by:

$$R(u)=(1-d)\ \Sigma_{v\in B(u)}\, R(v)/N(v) + d/n$$

Where: d is the dump factor which is a positive number such that $0 < d < 1$. Dump factor was proposed by PR inventors Page and Brin (1998) and widely used in different Page Rank computations. It helps to achieve two goals at once: 1) faster convergence using iterative computational methods; 2) problem becomes solvable for sure since all nodes have a possibility to be visited by a Random Surfer.

N*(v)* is the quantity of references for paper *v* and B(*u*) is variety of all articles which cite article *u*.

In the matrix from we can rewrite it as eigenvector problem:

$$R_{n+1}= (1\text{-}d).\ A.\ R_n+ d/n$$

Where A is a matrix defined as:

$$a_{ij} = \begin{cases} 1\,/N\,(i) & \text{if } i \text{ cites } j \\ 0 & \text{otherwise} \end{cases}$$

Hence the following algorithm allows to get ranking of data base:

```
Algorithm Pagerank
Data :    - Stochastic matrix A
          -
          - Dump factor $0 < d < 1$
          - A coefficient $\varepsilon$
Output: A ranking of references.
Begin
$R_0=Z$
Repeat
$R_{n+1}= (1-d). A. R_n+ d/n$
$\delta= ||R_{n+1}- R_n||$
Until $\delta\leq\varepsilon$.
END
```

4. Conclusions

A representation model of the scientific production using the citation for information retrieval is proposed. This model is based on the notion of similarity and ranking of documents by the notion of cocitation and adaptive pagerank algorithm. A formal evaluation is needed to validate our approach to help users for their information needs.

References

Cozzens E. S (1981). Taking the measure of science: A review of citation theories. *Newsletter of the International Society for the Sociology of Knowledge* 8 ,16

Cronin B (1981). The need for a theory of citation. *Journal of Documentation* 37, 16-24

Garfield E (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*,122, 108–111.

Garfield E (1979). *citation indexing: its theory and application in science, technology, and humanitie*. John Wiley, New York

Marshakova I.V., (1973). Document coupling system based on references taken from Science Citation Index. in Russia, Nauchno - Teknicheskaya Informatsiya, Ser.2 No.6,3.

Page and Brin; (1998). S. BRIN et L. PAGE. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, vol. 30, ṇ1–7, 107–117.

Price D. S (1970). Citation measures of hard science, soft science, technology, and non-science. In: Nelson, C. and Pollock, D, editors, *Communication among scientists and engineers.* Massachussett pp 3-22

Small H.G. (1973).Co-citation in the scientific literature. *Journal of the American Society for Information Science*. 24, 265-269.

Van Rijsbergen, C.J (1979). *Information Retrieval*. Butterworths, London, second edition.