On the Pagerank Algorithm for the Articles Ranking

Hacene AIT HADDADENE, Hakim HARIK and Said SALHI

***Abstract*— In this paper, we aim to study the ranking articles problem. A representation model based on a citation graph will be presented and an adaptation of a pagerank algorithm will be proposed for the problem of the articles ranking.**

***Index Terms*— citation graph, documents ranking, pagerank algorithm, similarity measure.**

I. Introduction

As greater volumes of documents has become available on the Internet and in data bases, users need more sophisticated tools to locate the information that is relevant to them; therefore it has become necessary to obtain an effective information retrieval systems.

Many of today's information retrieval systems use a traditional text processing to find all documents using the query terms, or related to the query terms by semantic meaning. Then, they rank the result using some ranking criterion. For the search engine, one popular way to create this ranking is to exploit the additional information inherent in the Web due to its hyperlinking structure. Thus, link analysis has become the means to ranking. One of the most successful and well-known link-based ranking systems is PageRank, the ranking system used by the Google search engine. Presently, for pages related to a query, an IR (information retrieval) score is combined with a PR (PageRank) score to determine an overall score, which is then used to rank the retrieved pages.

For the ranking in scientific domain, where the main metrics of articles contribution is the citations count [6], it's interesting to follow the approaches existing in search engine to rank web pages to the scientific papers. Chen et al [1] applied Page Rank idea for the scientific citations. In [11], authors applied "personalization" modification from above, where personalized vector was taken in proportion to the publishing journals weight. Then the validity of the rank was estimated by the cumulative gain function [11]. In [7], authors have developed an approach called "Focused Page Rank (FPR)" algorithm for the problem of scientific papers ranking which is based on the Focused Surfer model, where the probability to follow the reference in a paper is proportional to its citation count.

In this paper, we aim to study the articles ranking problem using an adaptation of the pagerank algorithm and the similarity measure. A model of representation of the scientific production will be presented. Some properties will be brought out- from the structure of the presented model- so that to propose an articles ranking method using the citation analysis.

H. AIT HADDADENE. LAID3 Laboratory, Operations Research Department, Faculty of Mathematics, USTHB University, BP32 El Alia, Bab Ezzouar 16111, Algiers, Algeria (e-mail: aithaddadenehacene@yahoo.fr)

H. HARIK. Research Centre on Scientific and Technical Information (CERIST), 5 Rue des frères Aissou, Ben Aknoun, Algiers, Algeria (e-mail: harik_hakim@yahoo.fr).

S. SALHI. Centre for Logistics & Heuristic Optimisation (CLHO), Kent Business School, University of Kent, Kent, UK (e-mail: S.Salhi@kent.ac.uk).

II. Citation analysis

The citation is none other than the relation which binds a citing document and a cited one [4], [10]. [9]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 review is estimated depending on how many times he (author) or it (article or review) is mentioned in other works [5], [2], [3]. This analysis requires the setting up of bibliographical data corpus, noted ψ, which represents (in this case) the set of references or papers to be ranked. Each of them includes a standard description of a document through a number of fields (title, authors, abstract, bibliography…etc.).

The model we have used is the 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, so it is possible, through this model, to identify semantic relationships between an article and the documents cited in it[4].

III. A method for articles ranking

Our approach for documents ranking is based on the notion of citation. It is summarized in two steps:

A. Similarity

A similarity measure is a function that associates a numeric value with a pair of sequences, with the principle that a higher value indicates greater similarity. In our case, the similarity between two documents of data base will be calculated by Jaccard index.

The Jaccard index- also known as the Jaccard similarity coefficient which we denote by S (initially called 'coined

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

coefficient' by Paul Jaccard)- is a statistic measure used for comparing the similarity and diversity of sample sets.

The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the sample sets union:

$$S(A, B)=| A\cap B | \ / | A \cup B | \quad (1)$$

Where:

- $| A\cap B |$ is a set of words which belongs to both articles A and B.
-$| A \cup B |$ is a set of all words of articles A and B.

B. Ranking

This part consists of the articles ranking according to their importance. We propose, as follows, an adaptation of a PageRank algorithm in order to rank and measure the relative importance of a document.

PageRank is the algorithm used by Google search engine, originally formulated by Brin and Page [8]. It is a method for computing a ranking for every web page based on the web graph. The importance of a web page can be judged by the number of hyperlinks pointing to it from other web pages.

In other words, the classification of the web is made according to their popularity, it assigns to pages authority scores that are higher for pages with many in-links from authoritative pages with relatively few out-links. It's a kind of "democratic vote", a bond emitted by a page A towards a page B is compared to a "vote" of A for B. The more the page receives "votes", the more this page is considered to be as important by Google, exactly as the elections principle which we all know.

The principle is to damp the Pagerank flow at each iteration and to redistribute the lost flow according to a given probability Z known as zap distribution [8].

Thus, for a web graph $G= (V, E)$, we define ranking R by:

$$R(u)= d \ .\Sigma_{v\in B(u)} R(v)/ N(v) + (1-d) \ . Z(u) \quad (2)$$

Where: $N(v)$ is number of outgoing links for web page v in a web graph G, $B(u)$ is a set of web pages pointed to a web page u, $Z(u)$ is a zap distribution for web page u and d is the dump factor ($0 < d < 1$) which was proposed by PR inventors Page and Brin [8] and widely used in different Page Rank computations. It helps to achieve two goals simultaneous: i) faster convergence using iterative computational methods; ii) the problem becomes solvable, for sure, since all nodes have a possibility to be visited by a Random Surfer.

In the matrix form, we can rewrite it as eigenvector problem:

$$R_{n+1}= d \ .A. \ R_n+ (1-d). \ Z \quad (3)$$

Where A is a matrix defined as:

$$a_{ij} = \begin{cases} 1/N(j) & \text{if } j \text{ points to } i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Hence the following algorithm allows to get a ranking of a web pages:

Pagerank algorithm

Data : - Stochastic matrix A
- Dump factor $0 < d < 1$
- A coefficient ε
- *Z*, zap distribution

Output: A ranking of web pages *V*.

Begin

$R_0=(R_0(v_1), R_0(v_2),\ldots, R_0(v_N)=Z$

Repeat

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

Until $||R_{n+1}- R_n||\leq\varepsilon$.

END

As a citation graph related to references is built the same way as a web graph. Then it's interesting to use an adaptation of pagerank algorithm for the scientific papers ranking, using a citation graph related to references.

In PageRank, the rank score of a page, *p*, is equally divided among its outgoing links. It gives the same score to its outgoing links nodes. It doesn't focus on the importance of each outgoing-link according to a page *p* and this is what we have added in our contribution to articles ranking case.

Then, in the following step, we propose a method for articles ranking based on similarities between articles and pagerank algorithm in a citation graph related to references. The principal idea of this algorithm is the probability to follow the reference in a paper which is proportional to its similarity. That is to say, the more an article is similar to a citing article the more score it receives from this article as far as they have the same content comparing to the others.

Thus, for the citation graph related to references $G_r=(V_r, U_r)$, we define ranking R by:

$$R_{n+1}(q) = d \times \left(\sum_{p\neq q}^{p\in B(q)} \left(R_{n+1}(p) \times \frac{S(p,q)}{S(p)} \right) \right) + (1-d) \times Z(q) \quad (5)$$

Where: $S(p) = \sum_{s\in L(p)} S(p,s)$

$L(p)$ is a set of an outgoing links for article *p* in the citation graph related to references G_r and $B(q)$ is a variety of all the articles citing the article *q*.

We can rewrite an adaptive pagerank algorithm in the matrix form:

Adaptive Pagerank algorithm

Data : - Stochastic matrix A'
- Dump factor $0 < d < 1$
- A coefficient ε
- *Z*, zap distribution

Output: A ranking of articles of Ψ.

Begin

$R_0=(R_0(p_1), R_0(p_2),\ldots, R_0(p_{Nr})=Z$

Repeat

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

Until $||R_{n+1}- R_n||\leq\varepsilon$.

END

Where A' is a matrix defined as:

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$a'_{ij} = \begin{cases} \dfrac{S(p_i, p_j)}{S(p_j)} & \text{if } p_j \text{ cites } p_i \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The previous algorithm allows us to make articles ranking using both an adapation of pagerang algorithm and the content similarity notion based on Jaccard index. Thus, the score assigned by each citing article to its cited articles is propostionnal to their common similarity.

Preliminary results will be given to show the superiority of this modified selection rule.

IV. Conclusion

A representation model of the scientific production, using the citation for articles ranking, is proposed. This model is based on the notion of similarity between articles and adaptive pagerank algorithm. A formal evaluation is needed to validate our approach.

References

[1] P. Chen, H. Xie, S. Maslov, S. Redner, "Finding scientific gems with Google's PageRank algorithm," Journal of Informetrics, 1, 2007, pp. 8-15.

[2] E. S. Cozzens, "Taking the measure of science: A review of citation theories," Newsletter of the International Society for the Sociology of Knowledge vol. 8, 1981, pp16-21.

[3] B. Cronin , "The need for a theory of citation," Journal of Documentation 37, 1981, pp. 16-24.

[4] E. Garfield, "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas," Science,122, 1955, pp.108–111.

[5] E. Garfield , *Citation indexing: its theory and application in science, technology, and humanitie*, Wiley, In Information sciences series, 1979, pp. 235-239.

[6] W. Glänzel, *Bibliometrics as a research field, A course on theory and application of bibliometric indicators*, Magyar Tudományos Akadémia, 2003, pp.1-115.

[7] M. Krapivin, M. Marchese,"Focused Page Rank in Scientific Papers Ranking," Lecture Notes in Computer Science; Vol. 5362, 2008, pp. 144 – 153.

[8] Page and Brin,. S. BRIN, L. PAGE, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems, vol. 30, 1998, pp.107–117.

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

[10] H.G. Small, "Co-citation in the scientific literature," Journal of the American Society for Information Science. 24, 1973, pp.265-269.

[11] Y. Sun, C. L. Giles, *Popularity Weighted Ranking for Academic Digital Libraries*, In 29th ECIR, 2007, pp. 605-612.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)