A Survey of Link Analysis Ranking

Antonis Sidiropoulos
Aristotle University of Thessaloniki, Greece

Dimitrios Katsaros
Aristotle University of Thessaloniki, Greece
University of Thessaly, Volos, Greece

Yannis Manolopoulos
Aristotle University of Thessaloniki, Greece

INTRODUCTION

During the past decade, the World Wide Web became the most popular network in the World. WWW grows with a very fast speed, thus the information that can be found through it is huge. In the early 90s, the first search engines for the WWW appeared. The user could give some keywords and the system returned a number of URLs (uniform resource locators) that contained the keywords. The order of the URLs in the return list was initially based on the number of the keyword occurrences in each URL. Some more sophisticated systems were taking into account the importance and the frequency of the keywords.

As WWW was growing, a simple keyword search could match hundreds of thousands of pages. A human can only check the first twenty or even some more of the URLs that the search engine returns. Consequently, the ordering of the search results became very important. The most important URLs that are related with the search keywords should be ranked first.

The link analysis ranking (LAR) is an objective way to sort search results. There are many advantages of the LAR over older methods. First of all the ranking is feasible without getting any feedback from the users. It is also not necessary to store the content of the URLs, but only the links. Another advantage is that it is difficult for the site developers to cheat by repeating keywords in the documents and moreover it may be pre-computed for all URLs. There are even more benefits using LAR to sort the search results that make it the best method used so far.

Existing LAR Algorithms

In this section, we will present the representative algorithms that perform Link Analysis Ranking. All these algorithms compute a score for each URL and usually the result is presented as a score vector. All these algorithms are computed iteratively. The initial score vector usually consists only of zeros or ones. In every computation step the previous score vector is used and the next score vector is computed. The computation repeats until the score vector converges to a constant value or until we reach a maximum number of steps. In some of the following algorithms, a normalization step is necessary after each iteration, otherwise the score vector will converge to infinity.

Throughout this section, we use the symbols of Table 1 to present all the algorithms in a unifying way.

Prestige

In 1949 (Seeley, 1949), an algorithm called *status* or *prestige* was applied in the scientific domain of social networks. It introduced the notion of *vertex* score based on the social network link analysis. Web can be considered as a social network, so the *prestige* algorithm can be applied over the Web-graph as Chakrabarti (2003) analyzes in his book.

The computation is based on the Web graph adjacency matrix A. An element $A[i,j]$ of this matrix contains the value 1, if page i links to page j. Starting with a prestige vector $\vec{p} = (1,...,1)^T$ we can compute the next Prestige vector $\vec{p}'$ as:

$$\vec{p}' = A^T * \vec{p} \qquad (1)$$

Copyright © 2008, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

Table 1. Notations

A	The adjacency matrix for the Web graph
N	The number of nodes (URLs) in the Web graph
I_x	The set of URLs that link to x
$\|I_x\|$	The number of URLs that link to x
O_x	The set of URLs that are pointed by x
$\|O_x\|$	The number of URLs that are pointed by x
d	Damping factor (set to 0.85 for PageRank)
b	Citation importance (usually set to 1)
a	Exponential Factor (>1, usually set to epsilon)

This assignment can be iterative until we reach a fix-vector. After each iteration, a normalization step must be applied. The normalization that is commonly used is done by summing all the vector elements to 1, $\| \vec{p}' \|_1 = 1$. The overall process is called power iteration (Golub & Loan, 1989) and the vector that p converges is called the *Principal eigenvector* of A^T.

According to our notation, the Prestige score for a URL x is the sum of the y URL-scores that link to x:

$$P'_x = \sum_{\forall y \in I_x} P_y \tag{2}$$

where P_x' is the prestige score for node x.

PageRank

PageRank was developed by Brin and Page (1998) at Stanford University. Nowadays it is used by the Google search Engine as the heart of the ranking system. Google has become the most popular search engine mainly due to the good rank behavior of PageRank. Originally, the `PageRank` score, PR, has been defined by Brin et al. (1998) as:

(3)

$$PR(A) = (1-d) + d\left(\frac{PR(t1)}{C(t1)} + \frac{PR(t2)}{C(t2)} + \ldots + \frac{PR(tn)}{C(tn)}\right)$$

Where $t1,\ldots,tn$ are pages linking to page A, C is the number of outgoing links from a page (out-degree) and d is a damping factor, usually set to 0.85.

PageRank looks like prestige, but it has the notion of *random walk*. Consider a Web surfer that surfs through the following links. Being in URL *I,* which has $C(i)$ links, the probability of moving to URL j that is pointed by i is $1/C(i)$[1] .

Then the probability of moving to another page that is pointed by j is $1/C(j)$, etc. If there are many cycles or the graph is disconnected, then the surfer will be trapped in a graph area. In order to avoid this entrapment, we instruct him or her not to follow these links forever, but he or she should jump to a random URL with a probability of 1-d. So, after following some links, he or she jumps to a uniformly selected random URL. PageRank computes the probability of the previous surfer to reach each URL.

Using the symbols of Table 1, the PageRank score for a node x (PR_x) is equivalent to:

$$PR_x = (1-d) + d \sum_{\forall y \in I_x} \frac{PR_y}{|O_y|} \tag{4}$$

Using vector symbols PageRank becomes:

$$\vec{PR}' = (1-d) * \vec{p} + d * L^T * \vec{PR} \tag{5}$$

With $\vec{p} = \left[\frac{1}{N}\right]_{N\times 1}$ and L is a matrix derived from A by normalizing all row-sums to one:

$$L[i,j] = \frac{A[i,j]}{|O_i|} \tag{6}$$

The damping factor d is used to guarantee the formula convergence and it is usually set to 0.85.

PageRank is precomputed for the entire Web-graph, so every page x has a PR value which denotes the probability of a Web surfer to reach page x by following

4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/survey-link-analysis-ranking/17794

Related Content

Smart Applications in Tourism
Cemal Inceand Gülmira Samatova (2020). *Handbook of Research on Smart Technology Applications in the Tourism Industry (pp. 345-370).*
www.irma-international.org/chapter/smart-applications-in-tourism/248563

Ratchet Head Pedagogy: A Narrative Autobiographical Inquiry about How We Learned to Customize and Tune Italian Motorcycles through Asynchronous Online Discussion
Ann-Louise Davidsonand Sylvain Durocher (2014). *Educational, Psychological, and Behavioral Considerations in Niche Online Communities (pp. 192-205).*
www.irma-international.org/chapter/ratchet-head-pedagogy/99302

Teaching and Learning Abstract Concepts by Means of Social Virtual Worlds
David Grioland Zoraida Callejas (2017). *International Journal of Virtual and Augmented Reality (pp. 29-42).*
www.irma-international.org/article/teaching-and-learning-abstract-concepts-by-means-of-social-virtual-worlds/169933

INSIDE: Using a Cubic Multisensory Controller for Interaction With a Mixed Reality Environment
Ioannis Gianniosand Dimitrios G. Margounakis (2021). *International Journal of Virtual and Augmented Reality (pp. 40-56).*
www.irma-international.org/article/inside/298985

Unravelling the Web: Adolescents and Internet Addiction
Laura Widyantoand Mark Griffiths (2011). *Virtual Communities: Concepts, Methodologies, Tools and Applications (pp. 2433-2453).*
www.irma-international.org/chapter/unravelling-web-adolescents-internet-addiction/48812