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:

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

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:

$$P'_x = \sum_{\forall y \in I_x} P_y \qquad (2)$$

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:

$$PR(A) = (1-d) + d\left(\frac{PR(t1)}{C(t1)} + \frac{PR(t2)}{C(t2)} + \ldots + \frac{PR(tn)}{C(tn)}\right) \qquad (3)$$

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:

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

Using vector symbols PageRank becomes:

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

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:

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

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

forward links. While a user enters the search keywords, it is performed a Boolean keyword match and a set of pages is returned in descending order of PR. This is an advantage but at the same time, it is also a disadvantage. The good side is that the user searches are very fast since the PR values are precomputed. The weak part is that the PR value is not query relevant. This is the reason that Google uses additional heuristics for the query result ordering.

During the last decade, there were a lot of attempts in the computer science literature to speed up the PageRank computation by using approximation methods. Kamvar, Haveliwala, Manning, and Golub (2003b) tried to predict the PageRank final score by using Aitken Extrapolation and Quadratic Extrapolation. They also (Kamvar, Haveliwala, Manning, & Golub, 2003a) presented a method that exploits the block structure of the Web. Lu et al. (2004) presented the PowerRank that exploits more Web attributes. This is gained by building several graphs in different granularity (graph of pages, hosts, domains etc.). Also Lu et al. (2004) introduced the notion of SOLB (same out-link behavior) for acceleration. Another method for acceleration is described by Arasu, Novak, Tomkins, and Tomlin (2002) that uses Gauss-Seidel method and SOR (successive over-relaxation) over the PageRank computation.

Other variations of PageRank are the PopRank (Nie, Zhang, Wen, & Ma, 2005) and *topic-sensitive* PageRank by Haveliwala (2003). This adds query related score to PageRank.

On the other hand, there is a lot of research about the PageRank behavior under certain graph characteristics such as the community structure on the Web (Bianchini, Gori, & Scarselli, 2003).

SCEASRank

SCEASRank (scientific collection evaluator by advanced scoring) is a generalized version of PageRank that can be used for both Web-graphs and citation-graphs. The generalized formula of SCEASRank is:

$$SR_x = (1-d) + d \sum_{\forall y \in I_x} \frac{SR_y + b}{|O_y|} a^{-1} \qquad (7)$$

where d is a damping factor, b is the factor of link/citation importance and a is an exponential factor that is usually set to ε (epsilon). b is set to one when the algorithm is applied on citation-graphs (Sidiropoulos & Manolopoulos, 2005, 2006). For the Web-graph case, b is preferred to have a value of 0. The exponential factor a adds to the random walk scenario the notion of the *path length memory*. As the path length of the followed links by the surfer increases, the possibility of following a link is divided by a. This means that if the surfer had followed a path of 5 links, then the possibility d to follow a link is divided by a^5. This assumption speedups the computation of the score matrix and the computation speed is much better than the PageRank one.

HITS

While Stanford was developing the PageRank, the IBM Almaden research center was defining HITS (Kleinberg, Kumar, Raghavan, Rajagopalan, & Tomkins, 1999). `HITS` (hypertext induced topic search) has been proposed to rank Web pages that are retrieved while searching through a browser. The notion behind `HITS` is the discrimination between hubs and authorities. Hubs are pages with good links, whereas authorities are pages with good content. Any node can be a hub or authority. Thus, `HITS` computes two vectors of scores. Originally the scores for hubs and authorities were defined by Kleinberg (1999) as:

$$\vec{a} = A^T * \vec{h}$$
$$\vec{h} = A * \vec{a} \qquad (8)$$

where A is the adjacency matrix of the citation graph. $A_{i,j}$=1, if page i links to page j and it is zero otherwise. $\vec{a}$ is a vector where its i-th element stands for the authority score of page i, while vector $\vec{h}$ contains the scores of the hub nodes. By using the terminology of Table 1, `HITS Authority (HA)` and `HITS Hub(HH)` scores for a URL x can be computed as:

$$HA_x = \sum_{\forall y \in I_x} HH_y$$
$$HH_x = \sum_{\forall y \in O_x} HA_y \qquad (9)$$

Formally, `HITS` is computed over a graph subset. This subset is prepared severally for every user query. The set consists of all these pages that contain the user keywords, plus the pages that point to them, as well as all the pages that are pointed by them. After the

subset computation, the HITS algorithm is performed by ignoring the rest of the Web-graph.

In practice, HITS ranking has some weak points. As Borodin, Rosenthal, Roberts, and Tsaparas (2005) proved, a hub is penalized when it points to "poor" authorities. As a result, the "poor" authorities become "poorer" during the computation.

Like the PageRank case, a lot of variations of the HITS algorithm exist. Two of them are the *randomized* HITS and the *subspace* HITS (Ng, Zheng, & Jordan, 2001). Finally, a very interesting work is presented by Farahat, LoFaro, Miller, Rae, and Ward (2006). In this work a path length history is imported into the adjacency matrix *A*.

Salsa

Stochastic approach for link structure analysis (SALSA) proposed by Lempel and Moran (2001) is a variation of HITS as it uses the notion of hubs and authorities, but the score is computed by a random walk (Lempel & Moran, 2000).

Having a collection of nodes *C* and an initial graph *G*=(*C*,*E*), Lempel builds a new graph $\tilde{G} = (V_h, V_a, E)$ where V_h is the set of nodes that have outgoing links, V_a is the set of nodes that have incoming links and *E* is the set of edges. Each URL $s \in C$ is represented by two nodes $\tilde{G}$ of , s_h and s_a. Each WWW-link s®r is represented by an undirected edge connecting s_h and r_a. Then the hub-matrix $\tilde{H}$ and the authority-matrix $\tilde{A}$ are defined as:

$$\tilde{h}_{i,j} = \sum_{\{k|(i_h,k_a),(j_h,k_a)\in\tilde{G}} \frac{1}{\deg(i_h)} \frac{1}{\deg(k_a)}$$

$$\tilde{a}_{i,j} = \sum_{\{k|(k_h,i_a),(k_h,j_a)\in\tilde{G}} \frac{1}{\deg(i_a)} \frac{1}{\deg(k_h)} \quad (10)$$

Where $deg(i_a)$ is the degree of node *i* as an authority and $deg(i_h)$ is the degree of node *i* as a hub. The formulation of SALSA score by using our notation is simply:

$$SA_x = \sum_{\forall y \in I_x} \frac{SH_y}{|O_y|}$$

$$SH_x = \sum_{\forall y \in O_x} \frac{SA_y}{|I_y|} \quad (11)$$

FUTURE WORK AND CONCLUSION

In this article, we presented the most representative link analysis ranking algorithms. The literature around this family of algorithms is very rich. The literature can be separated into two families: the research that is trying to improve their quality by examining their weak points and the research that is trying to accelerate the speed of computation either by using mathematical methods or by exploiting the Web-graph characteristics. For the future, we expect more work in *personalizing* PageRank and also in distributed and parallel computation of the above algorithms. It is also possible the appearance of an even more sophisticated algorithm that will be defined specifically for ranking the user search results. For example, an extension to the previous algorithms could be the use of a weighted adjacency matrix. Each matrix element *A*[*i*,*j*] could be zero when there is no link from URL *i* to *j* and it could be any real number greater than zero when the link exists. This could be considered as a weight of link i®j and could be computed based on various factors, such as the position of the link in the referring page, the keywords that are associated with the link, the font size of the link, etc.

REFERENCES

Arasu, A., Novak, J., Tomkins, A., & Tomlin, J. (2002). Pagerank computation and the structure of the Web: Experiments and algorithms. In *Proceedings of the 11th International Conference on World Wide Web—Posters.* Honolulu, Hawaii, USA: ACM Press.

Bianchini, M., Gori, M., & Scarselli F. (2003). Pagerank and Web communities. In *Proceedings of IEEE/WIC International Conference on Web Intelligence*: *Web Intelligence* (pp. 365-371). Halifax, Canada: IEEE Computer Society.

Borodin, A., Rosenthal, J. S., Roberts, G. O., & Tsaparas, P. (2005). Link analysis ranking: Algorithms, theory and experiments. *ACM Transactions on Internet Technologies*, *5*(1), 231-297.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings 7th WWW Conference on Computer Networks* (Vol. 30, pp. 107-117). Brisbane, Australia: Elsevier.

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext data* (pp. 205-206). Morgan Kaufmann Publishers.

Farahat, A., LoFaro, T., Miller, J., Rae, G., & Ward, L. (2006). Authority ranking from hits, pagerank, and salsa: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, *27*(4), 1181-1201.

Golub, G. & van Loan, C. (1989). Matrix computations. *John Hopkins University Press*.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on knowledge and data Engineering*, *15*(4), 784-796.

Kamvar, S., Haveliwala, T., Manning, C., & Golub, G. (2003a). *Exploiting the block structure of the Web for computing pagerank.* (Tech. Rep.), Stanford University.

Kamvar, S., Haveliwala, T., Manning, C., & Golub, G. (2003b). Extrapolation methods for accelerating Pagerank computations. In *Proceedings of the 12th International World Wide Web Conference* (pp. 261-270). Budapest, Hungary: ACM Press.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604-632.

Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). The Web as a graph: Measurements, models, and methods. In *Proceedings of Computing and Combinatorics, 5th Annual International Conference*, COCOON: Vol. 1627. Lecture Notes in Computer Science (pp. 1-17). Tokyo, Japan: Springer.

Lempel, R. & Moran, S. (2001). SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, *19*(2), 131-160.

Lempel, R., & Moran, S. (2000). The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, *33*(1-6), 387-401.

Lu, Y., Liu, X., Li, H., Zhang, B., Xi, W., Zhen, C., Yan, S., & Ma, W. Y. (2004). Efficient PageRank with same out-link groups. In *Proceedings of the 1st Asia Information Retrieval Symposium*, AIRS: Vol. 3411. Lecture Notes in Computer Science (pp. 141-152). Beijing, China: Springer.

Lu, Y., Zhang, B., Xi, W., & Zhen, C. (2004). The powerrank Web link analysis algorithm. In *Proceedings of the 13th International Conference on World Wide Web--Alternate Track Papers & Posters* (pp. 203-211). New York, NY, USA: ACM Press.

Ng, A. Y., Zheng, A. X., & Jordan, M. I. (2001). Stable algorithms for link analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 258-266). New Orleans, Louisiana, USA: ACM Press.

Nie, Z., Zhang, Y., Wen, J. R., & Ma, W. Y. (2005). Object-level ranking: Bringing order to Web objects. In *Proceedings of the 14th International Conference on World Wide Web*. (pp. 567-574). Chiba, Japan: ACM Press.

Seeley, J. (1949). The net of reciprocal influence: A problem in treating sociometric data. *Canadian Journal of Psychology*, *3*, 234-240.

Sidiropoulos, A., & Manolopoulos, Y. (2006). Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal for Systems and Software*, *79*(12), 1679-1700.

Sidiropoulos, A., & Manolopoulos, Y. (2005). A citation-based system to assist prize awarding. *ACM SIGMOD Record*, *34*(4), 54-60.

KEY TERMS

Adjacency Matrix: A matrix that corresponts to a directed graph *G=(V,E)*. Each element of an adjacency matrix *A[i,j]* has the value of 1 if there is a link from node *i* to node *j* and zero otherwise.

Authority: A Web-page with "good" content. Following Kleinbergs definition, a Web page is an authority if there "good" HUBS pointing to it.

Citation Graph: A dirrected and usually unweighted graph *G=(V,E)* that corresponts to a set of publications. Each vertex of the graph corresponts to a publication. Each edge $i \rightarrow j$ of *G* means that publication *i* cites publication *j*.

HTTP: hypertext transfer protocol. A network protocol used to fetch and/or transmit files over the network. This is the most commonly used protocol in the Web.

HTTPS: The hypertext tranfer protocol secure. The HTTP over a secure layer, usally the SSL (secure sockets layer).

Hub: A Web-page is a HUB if it has links to authorities.

Social Network: A social network is a set of people or groups of people with some pattern of contacts or interactions between them. Nowadays, the meaning of the term is very broad covering also other types of networks, like technological (Web, Internet, power grid), biological (gene, metabolic, food networks), whose nodes exhibit certain interactions between them.

URL: Uniform resource locator. A string representing the Web location of an object. It consists of tree parts. (a) the protocol--usually http or https, (b) the server, and (c) the full path of the file containnnign the object.

Web Crawler: Also known as "spider," "robot," "Webbot," or "bot." It is a program that fetches Web pages to a server in order to be indexed probably by another program. It usally "crawls" the Web by following links, but it is also an option to find the URLs that should be fetched in a database.

Web graph: A dirrected and usually unweighted graph that corresponts to a part of the Web. Each vertex of the graph corresponts to a Web page. Each edge of the graph corresponts to a Web link.

ENDNOTES

[1] Assuming that there is only one link to each URL