A Significant Improvement to Clever Algorithm in Hyperlinked Environment

Minhua Wang
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260
and
College of Business and Information Systems
Dakota State University
Madison, SD 57042
mw1@cse.buffalo.edu

ABSTRACT

This article addresses a problem with the existing approach by Clever Algorithm on studies of Web hyperlink structure: It does not take into consideration of user behavior when following multiple consecutive hyperlinks. We present an improved algorithm and show that any connected hyperlink graph has a unique hyper-weight (authority weight and hub weight) distribution. Our formulation has connections to the probability following each hyperlink, allows any norm in normalization, and does not require technical assumptions.

Keywords

World Wide Web; Information Retrieval; Hyperlink; Matrix Theory.

1. INTRODUCTION

Search services on the World Wide Web are the information retrieval systems that most people are familiar with. There is an additional source of information that an information retrieval system on the World Wide Web can harness: namely, the opinions of people who create hyperlinks. Hyperlinks encode a considerable amount of latent human judgment, and this type of judgment can be used to locate the information on the Web.

In a hyperlinked environment, a simple approach to finding quality pages is to assume that if page A has a hyperlink to page B, then the author of page A thinks that page B contains valuable information. Using this basic idea, Dr. Jon M Kleinberg ([K98]) developed a connectivity analysis algorithm for hyperlinked environments. Given an initial set of results from a search service, the algorithm extracts a subgraph (a hyperlink graph) from the Web containing the result set and its neighboring pages. This is used as a basis for an iterative computation that estimates the value of each page as a source of relevant hyperlinks and as a source of useful content. The algorithm was later implemented by IBM Almaden Research Center in Clever, a prototype search engine, and was named Clever Algorithm in a featured article of Scientific American ([M99]). Clever Algorithm is also named as HITS Algorithm in some research papers ([KK99]).

Clever Algorithm computes two scores for each page: an authority score and a hub score. Pages that have high authority scores are expected to have relevant content, whereas pages with high hub scores are expected to contain hyperlinks to relevant content. The computation of authority and hub scores is done as follows: Let G = (V, E) be the hyperlink graph. For every node p_i in V, let $a[i]$ be its authority score and $h[i]$ its hub score. Initialize $a[i]$ and $h[i]$ to *1* for all nodes in V. Then, while the vectors *a* and *h* have not converged:

{

For all nodes in V, $a[i] := \sum_{(p_j, p_i) \in E} h[j]$;

For all nodes in V, $h[i] := \sum_{(p_i, p_j) \in E} a[j]$;

Normalize the *a* and *h* vectors in L^2 norm.

}

Kleinberg proved that the a and h vectors will eventually converge subject to a certain technical assumption. The pages are then ranked by authority and hub scores respectively.

Since the publication of Clever Algorithm, it has become the primary algorithm to compute authority and hub scores on the Web. There are several studies on the improvement of the generation of the hypergraph (e.g. [BH98]). But the algorithm procedure itself was not questioned.

We now reveal a problem with Clever Algorithm through a simple example as follows: Let G be the 4-node graph

$$p_1 \leftarrow p_2 \rightarrow p_3 \rightarrow p_4,$$

i.e. there are three hyperlinks: from p_2 to p_1, from p_2 to p_3, and from p_3 to p_4. Applying Clever Algorithm to G, both the authority score for p_4 and the hub score for p_3 would converge to 0. This is an abnormal scoring, since the authority value of p_4 and the hub value of p_3 should be conferred by the hyperlink from p_3 to p_4. Another observation is that, if we remove node p_4 from G, by applying Clever Algorithm, p_1, p_2, and p_3 in the resulting graph:

$$p_1 \leftarrow p_2 \rightarrow p_3$$

will be assigned exactly the same authority scores and hub scores as in G. Thus Clever Algorithm is actually "nullifying" the node p_4 in G.

The basic reason why Clever Algorithm presents this kind of problem is as follows: It is based on the authority-hub reinforcement through single directed hyperlinks only (by adjacency matrix) and has no consideration of following multiple consecutive hyperlinks (a normal behavior when users navigate the Web), and thus no score updating is done through multiple consecutive hyperlinks.

Due to the problem with Clever Algorithm, new algorithm needs to be designed. We state our new design in the next section.

2. New Algorithm

We stick with the idea of authority-hub interaction and start with inventing the new matrix to be used in the updating stages of Clever Algorithm instead of the adjacency matrix. To take consideration of multiple consecutive hyperlinks, from any page p_i to another page p_j, we need to determine all multiple consecutive hyperlinks from p_i to p_j, which are exactly all directed paths on the hyperlink graph from p_i to p_j. Our approach is trying to figure out the chance (probability) of following each directed path and make contribution to the score updating.

Definition 2.1: Let G = (V, E) be a connected hyperlink graph, with V = $\{p_1, p_2, \ldots, p_n\}$ where $n > 1$, and let P denote the *hyperlink probability matrix* of the hyperlink graph G; the (i, j)th entry of P is equal to the (positive) probability of following the directed hyperlink from p_i to p_j on page p_i if such hyperlink exists, and is equal to 0 otherwise.

Remark: Since there is a (positive) probability of not following any directed hyperlink ([LB01]), the sum of entries in each row of P is strictly less than 1. Further discussions on how to find P are given in [LB01] and [BL00].

We now generate a matrix H such that the (i, j)th entry of H is the probability of following all multiple-hyperlinks (directed paths) from p_i to p_j.

Definition 2.2: The *multiple-hyperlink probability* matrix

$$H = \sum_{m=1}^{\infty} P^m = P\,(I - P)^{-1}.$$

Our improved algorithm to find authority and hub scores are as follows: Let G = (V, E) be a connected hyperlink graph with V = $\{p_1, p_2, \ldots, p_n\}$ where $n > 1$. Find the hyperlink probability matrix P and compute the multiple-hyperlink probability matrix H. For every node p_i in V, let $a[i]$ be its authority score and $h[i]$ its hub score. Initialize $a[i]$ and $h[i]$ to 1 for all nodes in V. Then, while the vectors a and h have not converged:

{

For all nodes in V, $a[i] := \sum_{j=1}^{n} H_{ji} h[j]$;

For all nodes in V, $h[i] := \sum_{j=1}^{n} H_{ij} a[j]$;

Normalize the *a* and *h* vectors.

}

In the above algorithm, we allow the norm to be in any form while L^1 norm is preferred.

3. Main Theorem

Our next work is trying to establish the existence and uniqueness result of our newly designed authority and hub scoring system.

Definition 3.1: A *hyper-weight distribution* on G is defined as an ordered pair (a^*, h^*) where a^* is a nonnegative unit column vector of length *n* (called authority weight) and h^* is another nonnegative unit column vector of length *n* (called hub weight), such that $a^* = H^T h^* / \| H^T h^* \|$ and $h^* = Ha^* / \| Ha^* \|$.

Remark: $\| \bullet \|$ is the norm. The (a^*, h^*) pair is actually the desired "equilibrium" values for authority scores and hub scores. The two equalities are the basic means by which authorities and hubs reinforce another.

If (a^*, h^*) is a hyper-weight distribution, then a^* is a unit eigenvector corresponding to a positive eigenvalue of the matrix $H^T H$. Fortunately, $H^T H$ has certain "positiveness" for connected hyperlink graph. Using Matrix Theory techniques, we can prove that the largest eigenvalue of the matrix $H^T H$ is simple and one of its eigenvectors is the a^* we are looking for.

Main Theorem: Any connected hyperlink graph has a unique hyper-weight distribution. The convergence of our improved algorithm is guaranteed with $a \to a^*$ and $h \to h^*$.

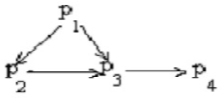
4. Examples and Conclusions

We compare our new improved algorithm to Clever Algorithm in several Web experimental examples and find our new algorithm is giving convincing results. We just present two simple examples here. The second example is actually extracted from a real Web search result.

Example 1: The previous demonstrated 4-node graph $p_1 \leftarrow p_2 \rightarrow p_3 \rightarrow p_4$.

	authority score	hub score
Clever Algorithm	$(1/2, 0, 1/2, 0)^T$	$(0, 1, 0, 0)^T$
Our improved algorithm	$(1/4, 0, 1/4, 1/2)^T$	$(0, 1/2, 1/2, 0)^T$

Example 2: Another 4-node graph



	authority score	hub score
Clever Algorithm	$(0, 0.3820, 0.6180, 0)^T$	$(0.6180, 0.3820, 0, 0)^T$
Our improved algorithm	$(0, 0.1706, 0.4737, 0.3558)^T$	$(0.4317, 0.3676, 0.2007, 0)^T$

The overhead of our new algorithm is the probability matrices. This makes the improved algorithm with higher computation complexity. Further research could be conducted on how to figure out the probability matrices in real hyperlinked environments and how to speed up the algorithm execution.

5. ACKNOWLEDGEMENTS

We sincerely thank Dr. Xin He, Dr. Jinyi Cai, and anonymous referees for their comments on a draft of this article.

6. REFERENCES

[BB98] K. Bharat, and A. Broder. *A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines*. Proc. 7th International World Wide Web Conference, 1998.

[BH98] K. Bharat, and M. Henzinger. *Improved Algorithms for Topic Distillation in Hyperlinked Environments*. Proc. 21st SIGIR, 1998.

[BL00] J. Borges, and M. Levene. *Data Mining of User Navigation Patterns*, in Web Usage Analysis and User Profiling, pp. 92-111. Published by Springer-Verlag as Lecture Notes in Computer Science, Vol. 1836, 2000.

[BP98] S. Brin, and L. Page. *The Anatomy of a Large-Scale Hypertexual Web Search Engine*. Proc. 7th International World Wide Web Conference, 1998.

[CD97] J. Carriere, and R. Kazman. *WebQuery: Searching and Visualizing the Web Through Connectivity*. Proc. 6th International World Wide Web Conference, 1997.

[HJ90] R. Horn, and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[K98] J. Kleinberg. *Authoritative Sources in a Hyperlinked Environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998, extended version in Journal of the ACM 46(1999), also appears as IBM Research Report RJ 10076, May 1997.

[KK99] J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. *The Web as a Graph: Measurements, Models and Methods*. Invited survey at the International Conference on Combinatorics and Computing, 1999.

[KR99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. *Trawling the Web for Cyber Communities*. Proc. 8th International World Wide Web Conference,1999.

[LB01] M. Levene, J. Borges, and G. Loizou. *Zipf's Law for Web Surfers*. Knowledge and Information Systems an International Journal, 3, 2001.

[M99] Members of the Clever Project (S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins). *Hypersearching the Web*. Feature article, Scientific American, June 1999.