Selective Hypertext Induced Topic Search

Amit C. Awekar
NC State University
Raleigh, NC 27695, USA
acawekar@ncsu.edu

Pabitra Mitra
Indian Institute of Technology
Kharagpur, India - 721302
pabitra@cse.iitkgp.ernet.in

Jaewoo Kang
NC State University
Raleigh, NC 27695, USA
kang@csc.ncsu.edu

ABSTRACT

We address the problem of answering broad-topic queries on the World Wide Web. We present a link based analysis algorithm SelHITS, which is an improvement over Kleinberg's HITS [2] algorithm. We introduce the concept of virtual links to exploit the latent information in the hyperlinked environment. We propose a novel approach to calculate hub and authority values. We also present a selective expansion method which avoids topic drift and provides results consistent with only one interpretation of the query, even if the query is ambiguous. Initial experimental evaluation and user feedback show that our algorithm indeed distills the most important and relevant pages for broad-topic queries. We also infer that there exists a uniform notion of quality of search results within users.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*

General Terms

Algorithms

Keywords

Link analysis, Searching, Topic Distillation

1. INTRODUCTION

Searching information on the WWW is now a common practice. Traditional information retrieval techniques are not sufficient for searching information on the WWW. Kleinberg's HITS algorithm [2] is a well known algorithm for answering broad-topic queries in hyperlinked environment. We improve it by selectively choosing the candidate set of pages for ranking and novel approach to ranking the pages. Section 2 briefly describes the original HITS algorithm. Section 3 introduces our SelHITS algorithm. Section 4 summarizes the initial experimental results and conclude the paper.

2. KLEINBERG'S HITS ALGORITHM

HITS algorithm [2] is briefly outlined in Figure 1. It associates two scores with each page. *Authority* which is a measure of authoritative information contained in the page and *Hub* which is a measure of links to good authorities. Using some existing search system, HITS algorithm gets a set of relevant pages for user query. This is referred to as the root set. Then all the pages from one link neighborhood of the root set are added to the root set. This is referred to as the base set. Consider a page P_i. Let Par_i be set of pages which have hyperlink to P_i and are present in the base set. Let Chi_i be set of pages that P_i has hyperlink to and are present in the base set. Let hub value for P_i be denoted as H_i and authority value as A_i. Then $A_i = \sum_{l \in Par_i} H_l$ and $H_i = \sum_{l \in Chi_i} A_l$. If E is the adjacency matrix for the base set, then authority vector V_a is calculated as $V_a = E^T E V_a$. The vector V_a converges to the principal eigenvector of $E^T E$. The matrix $E^T E$ is real, symmetric and non-negative. Hence its principal eigenvector is real and non-negative. Hub vector V_h is calculated as $V_h = E V_a$. Lagville et al. [3] present a comprehensive survey of various modifications proposed for HITS and other eigenvector based algorithms for web information retrieval.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

3. SELHITS ALGORITHM

SleHITS algorithm is briefly outlined in Figure 2. It is motivated by following observations. First, considering just page to page connectivity ignores other latent information in the WWW. For example, if multiple pages from same host are present in the root set then it indicates that there exists a community of pages on that host and it is relevant to the query topic. This context of location of pages is ignored by HITS algorithm. To exploit such latent information in the hyperlinked environment, we propose a novel approach to calculate hub and authority values along with the concept of virtual links. Many pages added to the root set in the expansion step of HITS algorithm are irrelevant or noisy. They substantially degrade the quality of results. We should be selective in expanding the root set. For ambiguous queries we should be able to distill pure topic i.e. giving results consistent with only one interpretation of the query. Now we describe proposed improvements to address these problems.

3.1 Novel Approach to Ranking

We consider two types of links. *Actual links* are the hyperlinks actually present between different pages. *Virtual links* are the pseudo links that we hypothetically insert. If a page P_i has actual link to page P_j in the root set then we insert pseudo links from P_i to all other pages in the root set which reside on the same host as that of P_j. Considering virtual links helps exploit the context of location of a page.

Figure 1: HITS Algorithm

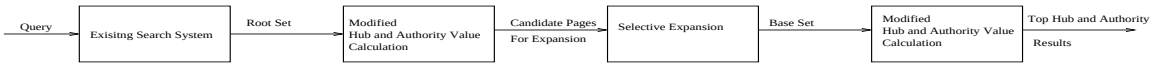

Figure 2: SelHITS Algorithm

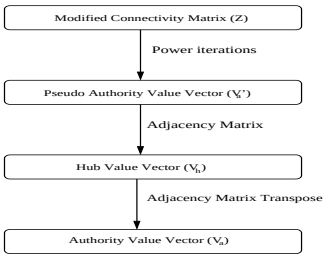

Figure 3: Calculating Hub and Authority Values

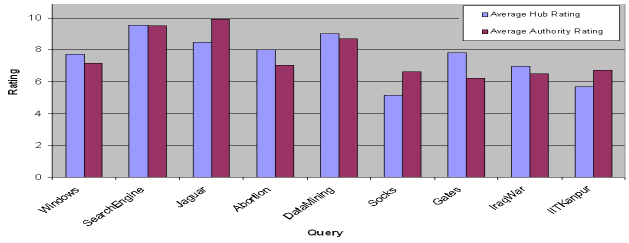

Figure 4: Average Ratings for Hub and Authorities

For a given root set, consider adjacency matrix E. $E[i,j] = 1$ if P_i has actual link to P_j. We define modified adjacency matrix Z such that $Z[i,j] = 1$ if P_i has actual or virtual link to P_j. Then hub and authority vectors for the root set are calculated as shown in Figure 3. Same can be represented in equation form as $V_a^{'} = Z^T Z V_a^{'}$. First, calculating pseudo authority values assigns same authority value to all pages on one host as we are considering virtual links. Then actual hub values are calculated using these pseudo authority values as $V_h = EV_a^{'}$. This calculation of hub vector boosts hub values for those pages which have hyperlink to hosts which contain more pages in the root set. Finally actual authority values are calculated as $V_a = E^T V_h$. This is required as initial authority values of some pages are unnecessarily boosted because of virtual links. We calculate hub and authority values twice, once on the root set and once on the base set, as illustrated in Figure 2.

3.2 Selective Expansion

After calculating hub and authority values on the root set, we select top 20 hubs and top 20 authorities for expansion. Good hubs point to good authorities [2], so we add outlinks of top 20 hubs to the root set. Good authorities are pointed by good hubs [2], so we add inlinks of top 20 authorities to the root set. As compared to HITS algorithm, this selective expansion procedure drastically reduces size of the base set and avoids topic drift as irrelevant pages are not added to the root set.

If query is ambiguous then we have corresponding disjoint communities in the link structure [2]. Top hubs and authorities are from the same community as hub and authority values are mutually reinforcing. Selective expansion procedure adds new pages from single community, which further boosts hub and authority values for that community. Hence we are able to give results consistent with only one interpretation of the query. If user is interested in other interpretation of the query then we can simply remove current community from candidate pages and again run our algorithm.

4. EXPERIMENTS AND CONCLUSION

We tested our algorithm for 9 sample queries. We observed significant topic drift with HITS algorithm for these queries. For each query we generated top 20 hubs and top 20 authorities by SelHITS algorithm. Each query results were evaluated by 3 different users on the scale of 0 to 10. Figure 4 shows average scores for each query. Within each query, scores for hubs and authority are similar. Also, we noticed that scores of different users for a given query were also similar. So we can infer that there exists a notion of quality of search results within users. For ambiguous queries, we got results consistent with only one interpretation of the query. For example, for query "gates" all results were related to "Bill Gates". More details about experimental results are available at [1].

We briefly discussed the SelHITS algorithm which is an improvement over Kleinberg's HITS algorithm for answering broad-topic queries. Novel approach to calculate hub, authority values and selective expansion of the root set are the main contributions of our work. Brief experimental evaluation suggests that algorithm performs satisfactorily. We are currently working on extensively testing the proposed algorithm for variety of queries and applying the same algorithm for focused crawling problem.

5. REFERENCES

[1] More details about experiments on SelHITS algorithm. `http://www4.ncsu.edu/~acawekar/`SelHITSResults/.

[2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5–7):604–632, 1999.

[3] A. Langville and C. Meyer. A survey of eigenvector methods of web information retrieval. *SIAM Rev.*, 47(1):135–161, 2005.