

Finding Related Hubs and Authorities

Paul-Alexandru Chirita
Learning Lab Lower Saxony
Deutscher Pavillon Expo Plaza 1
30539 Hannover, Germany
+49(0)511.762.9731
chirita@learninglab.de

Daniel Olmedilla
Learning Lab Lower Saxony
Deutscher Pavillon Expo Plaza 1
30539 Hannover, Germany
+49(0)511.762.9741
olmedilla@learninglab.de

Wolfgang Nejdl
Learning Lab Lower Saxony
Deutscher Pavillon Expo Plaza 1
30539 Hannover, Germany
+49(0)511.762.9714
nejdl@learninglab.de

Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Storage and Retrieval—
Information Search and Retrieval

General Terms

Algorithms, Measurement, Performance

Keywords

Hubrank, hubfinder, PageRank, web search

ABSTRACT

In this paper a new algorithm for finding hubs (and authorities) related to a specific Web page is presented. Various aspects have been investigated in this direction, like extensions of the HITS algorithm or modifications to the Pagerank algorithm. Our focus is mainly oriented on computing hub scores, but the formulas for authority scores are almost always analogous.

1. INTRODUCTION

The analysis of the link structure of the World Wide Web has been an extensively investigated issue in the past few years. The reason is probably the extraordinary impact produced by the publication of the Google pagerank algorithm [6]. Also, the size and diversity of all Web pages from the Internet (now reaching about 3 billion pages) obliges researchers to seek for faster and more personalized search engine algorithms. As is it described in [1] there are three different types of queries according to their intent: *navigational* (to get a particular site), *informational* (to retrieve information assumed to be present on one or more web pages) and *transactional* (to perform some web-mediated activity).

We are focusing on *informational queries* where a user is interested in finding pages related to specific issue. Current web search engines retrieve several important pages but they are not always the right ones. Our approach intends to give the user the possibility to find hubs related to the pages he is interested in, increasing the accuracy of results and therefore satisfaction of the user. In section 2 previous work in this direction is described. Section 3 presents our new algorithm for ranking pages and section 4 presents the algorithm for finding hubs (and authorities) related to a specific Web page. Afterwards, in section 5 we show some initial results, and finally, in section 6 some conclusions are drawn and our further work is described.

Copyright is held by the author/owner(s).
WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM 1-58113-680-3/03/0005.xxx.

2. PREVIOUS WORK

In this section we describe the most currently used ranking algorithms for computing page and/or hubs and authorities scores.

2.1 Pagerank

Pagerank [6] can be imagined as a “vote” from a set of pages in the Web to a specific page. One link to that page represents one vote. Briefly, if we consider the Web as a graph and A is its adjacency matrix, then we can compute the ranks of Web pages in the x vector using the following formula:

$$x^{n+1} \leftarrow c \cdot A \cdot x^n + (1 - c) \cdot e$$

where c is a constant (called the dumping factor, usually set to 0.85) and e is a vector with all components equal to 1 at each iteration.

2.2 HITS, SALSA and Randomized HITS

The idea behind the HITS algorithm [3] is to compute two scores for each page in a community, that is a hub score and an authority score. The algorithm starts with a set R of pages with high pagerank (as if it were computed by a search engine). Then, this set is firstly extended using a method which we will later call the Kleinberg Extension. For each page p in the set R , two sets are constructed: one with pages that p points to and the other with d or less pages pointing to p . These two sets are then added to the base set R . The extension procedure may be repeated several times. After the targeted set of pages is generated, the hubs and authorities scores are computed. Two weights are generated: an authority weight and a hub weight.

SALSA [4] is a stochastic approach similar to the HITS algorithm but based on a weighted in-degree analysis and computationally more efficient.

Randomized HITS [5] seems very much like a combination between Pagerank and HITS. It computes two vectors, one for hub scores and one for authority scores, like HITS. On the other hand, it is based on a fixed point iteration which is very similar to the one of Pagerank.

3. HUBRANK

HubRank is a new algorithm for computing hubs and authorities scores. Our new approach is a slight modification of the Google pagerank algorithm [6]. It may also be compared to [5] where the authors combine the Google pagerank with the HITS algorithm into a new fixed point iteration for computing hub and authority scores. The idea behind our algorithm is that pages with a bigger out-degree should have a bigger hub rank and similarly pages with bigger in-degree should have a bigger authority rank. To achieve this, the personalization vector is modified to consider

the out-degree/in-degree of the pages. More intuitively, the random surfer will always prefer pages with a big out-degree when it gets bored. This way, the global importance of the pages is also playing an important role in defining general scores, as the random surfer will follow the out-going links with a higher probability than the random ones. However, the c constant has to be smaller than 0.85, that is the random surfer has to prefer the random pages more often. The algorithm for computing hub scores is depicted in the following lines.

Let N be the total number of pages in the Web graph
 Let SO be the sum of all the out-degrees of the pages
 Let O_i be the out-degree of page i
 Let the components of the personalization vector e be:

$$e_i = O_i \cdot \frac{N}{SO}$$

Apply the Google pagerank using e as the personalization vector:

$$x^{n+1} \leftarrow c \cdot A \cdot x^n + (1 - c) \cdot e$$

Return x

Analogously, the authority scores can be computed setting the components of the personalization vector to $e_i = I_i \cdot \frac{N}{SI}$, where I_i is the in-degree of page i and SI is the sum of all the in-degrees of the pages of the Web.

4. HUBFINDER

Now that we have the hub (authority) scores for the pages, we must define an algorithm for finding other hubs, somehow related to an initial base set. Before defining such an algorithm, let us see what we have had in the beginning. All the previous algorithms (HITS [3], SALSA [4], Randomized HITS [5]) are focusing on applying the Kleinberg extension several times and then computing hub and authority scores for the resulting pages. When searching for *related* hubs (authorities), one will probably need to build a very large set of pages before having found enough representative pages (good hubs/authorities). Remember that in our case *related* means accessible via the link structure of the Web (following either in-going, or out-going links). Our idea has the following skeleton algorithm as a basis:

Let Γ be the Base Starting Set of pages whose related hubs/authorities we are looking for
 $\Gamma =$ Apply the Kleinberg Extension on Γ once
 $\Gamma' = \Gamma$
 For $i = 1$ to σ do:

$\Gamma'' =$ Apply the Kleinberg Extension on Γ'
 once
 Trim Γ'' to contain only *interesting* pages,
 which are *not* contained in Γ
 $\Gamma = \Gamma + \Gamma''$
 $\Gamma' = \Gamma''$

End For
 Trim Γ to contain only interesting pages
 Return Γ

The only particular aspect is which are the *interesting pages*? It depends on the approach we take. For example, the naïve approach considers more interesting the pages with a high global Google pagerank. One may also calculate the Google pagerank for the

small graph created after applying the Kleinberg extension, but this would provide poorer results. A page with high global pagerank is definitely more important (globally) than a page with high local pagerank and small global pagerank. Also, even when thinking that a local page might be closer to user's preferences than the global one, we may argue that if the global one is accessible from the user's starting set, then it is considered at least partially important by the user.

5. INITIAL RESULTS

Pagerank considers important pages regarding to sites recommendations (links). Hubs and authorities take into consideration mutual reinforcement between pages based on in-going and out-going links. In our experiments HubRank has shown to take into consideration both aspects at the same time.

It is also worth mentioning that in an experiment of the HubFinder algorithm with 6 steps (6 Kleinberg extensions) and 3 pages in the base starting set, the three algorithms presented above generated the following results for a Web crawl of 10,000 pages:

Table 1: Number of pages at each iteration of HubFinder

Step	HubFinder				HITS
	Discovered	New	New and Interesting	Total	Total
1	125	125	125	128	128
2	180	52	43	171	180
3	143	78	39	210	254
4	536	481	159	369	937
5	1913	1718	415	784	2755
6	2297	1721	342	1126	3414

Finally, let us say that while our algorithm (HubFinder) needed 5.6 seconds to run this test, the approach based on the [[3]] needed 10.11 seconds.

6. CONCLUSIONS AND FURTHER WORK

We have presented a new algorithm which combines the main properties of Pagerank and HITS thus improving accuracy in user topic-queries as well as a new and faster algorithm to compute hubs related to specific pages.

We intend to apply some properties of SimRank [2] in order to further extend the resulting set of hubs from HubFinder. Also we must perform tests on much larger datasets.

7. REFERENCES

- [1] A. Broder. A taxonomy of web search. Technical report, IBM Research, 2002.
- [2] G. Jeh and J. Widom. Simrank: A measure of structural-context similarity, 2002.
- [3] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [4] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):387–401, 2000.
- [5] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Stable algorithms for link analysis. In *Proc. 24th Annual Intl. ACM SIGIR Conference*. ACM, 2001.
- [6] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.