Searching the Most Authoritative & Obscure Sources from the Web

Bharat Bhushan[1] , Narender Kumar [2]

[1]Department of Computer Science & Applications, Guru Nanak Khalsa College, Yamuna Nagar (Haryana), India
[2] # 478/9 Laxman Colony, Thanesar City, Kurukshetra (Haryana), India

Summary
This paper discusses the ways to rank the website using authoritative and obscurity parameters. Website ranking algorithm is leveraging parameters with primary focus on Authoritativeness and relevancy. Additionally, obscurity parameters like Subscriber Counts from Google Reader, Authority Count from Technorati are analyzed to rank a source which is less visible or not too popular or related to a non-famous Zone. This tool viz. source rank tool will be an asset to help the industries working under data extraction using web to rank the sources as per their business needs e.g. a source with low page rank (PR) can be ranked as high if a person who writes the article is an expert in his domain.

Keywords:
Website, Search Engine, Source Rank tool, Blog, SEO, PR

1. INTRODUCTION

Searching for targeted information from the web can be challenging due to the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to develop the searching tool in order to quickly and efficiently find, extract, filter, and evaluate the desired information and resources from high quality data available on the web.

Brin and Page [1], and Klienberg [2] emphasised that the structure of Internet hyperlinks is an effective indicator of the relevance and importance of a web document. Ask.com and Teoma use a hyperlink analysis algorithm based on Klienberg's HITS (Hypertext induced topic selection algorithm). Similar to what is found in research papers, the link to (or mention of) another document on the web should be an indication of its importance to web users. Indirectly, hyperlinks are a kind of review or screening of a webpage by a web user.

Google's success and the concept of page rank [1] is based on the underlying assumption that the link to another page is a PageRank and a website's linking structure are still considered the most important measure of a website's ability to rank on any given search query "on topic" link and that the link is an unbiased link to a website

Matt Cutts [3] addressed the concept of the effectiveness of "Page Rank sculpting" by using the no follow attribute. The idea behind PageRank sculpting is that one can minimize the passing of PageRank to other pages within a site by using the no follow attribute on pages under a single URL.

Zoltán Garcia-Molina of Stanford University and Jan Pedersen of Yahoo [4] released a paper in 2004 on the "TrustRank" method. TrustRank is an algorithm that can be used to help automatically identify "trusted," human-reviewed webpages (in other words, a small set of human-selected seed pages). They claim this method can filter out a significant amount of web spam by beginning with a known set of "trusted" authority webpages. For example, the open directory project [5] contains a set of such documents.

Gyongyi et al. [6] discussed Link Spam Detection Based on Mass Estimation and defined "link spamming" as "an attempt to mislead search engines and trigger an artificially high link-based ranking of a specific targeted webpage." They claim that "spam mass" is a measure of the impact of link spamming on a page's ranking.

Mohhammad A Tayebi et. al. [7] discussed that blogs have become one of most important parts of web but there are not so efficient search engines for them. One reason is differences between regular web pages and blog pages and inefficiency of conventional web pages ranking algorithms for blogs ranking. There are some works in this field but have not considered yet. They developed blogs ranking algorithm called B2Rank based on users' behavioral features these features.

Lan Nie Baoning Wu Brian D. Davison [8] proposed a novel cautious surfer to incorporate trust into the process of calculating authority for web pages. They evaluate a total of sixty queries over two large, real-world datasets to demonstrate that incorporating trust can improve Page Rank's performance.

It is the search engine developer's job to generate a set of highly relevant documents for any search query, using the available parameters on the web. The task is challenging because the available parameters usable by the algorithm

Manuscript received August 5, 2010
Manuscript revised August 20, 2010

are not necessarily the same as the ones web users see when deciding if a webpage is relevant to their search. A good ranking algorithm would require either more variables or rely on factors a webpage author cannot control directly.Using more variables in a ranking algorithm naturally makes the manipulation of its search results more difficult.

Today most of the websites use SEO techniques to get the good PR. The biggest challenge that knowledge industry is facing today is to segregate the sources under authority/relevancy and obscurity. The problem is solved by building a sophisticated algorithm using Dmoz, Wiki, Google Authentic view, Del.icio.us, Google PR (Page Rank),Back Links (Yahoo),Traffic Rank (Alexa),Index Pages (Yahoo),Google News Index Pages, Google News Pick Status, Technorati Blog Reactions.

Algorithm also uses some of the obscurity parameters for blog source rank like Subscriber Counts from Google Reader, Posts on Blog from Google Reader, Authority Count from Technorati..

Proposed Model

Search engines like GOOGLE, ALTAVISTA etc. available on the net to retrieve the information. The purpose of these search engines is to search and retrieve the information from various websites as per the search specifications. Here the information is retrieved in a general form, may or may not suites a particular person. The search engines use a crawler to get the information from various sources and make it available to its users. But in this proposed model specific demand oriented information is retrieved a vast range of web sources from around the world like: Local and world news In-depth industry and subscription sources Blogs Company websites Governmental and regulatory agencies

Algorithm The algorithm works on 12 parameters. The 11 parameters it scraps from the web, which calculates the Rank of the source i.e. High, Mid or Low. The 12th parameter is user priority which is top of all the parameter. This can not over ride the Rank of the source calculated through 11 parameters.

Sources Ranking works on 12 factors.

1. GOOGLE PR (PAGE RANK)
2. BACK LINKS (YAHOO)
3. TRAFFIC RANK (ALEXA)
4. INDEX PAGES (YAHOO)
5. GOOGLE NEWS INDEX PAGES
6. GOOGLE NEWS PICK STATUS
7. DMOZ
8. GOOGLE AUTHENTIC VIEW
9. TECHNORATI RETURNS
10. USER PRIORITY
11. DEL.ICIO.US
12. WIKI

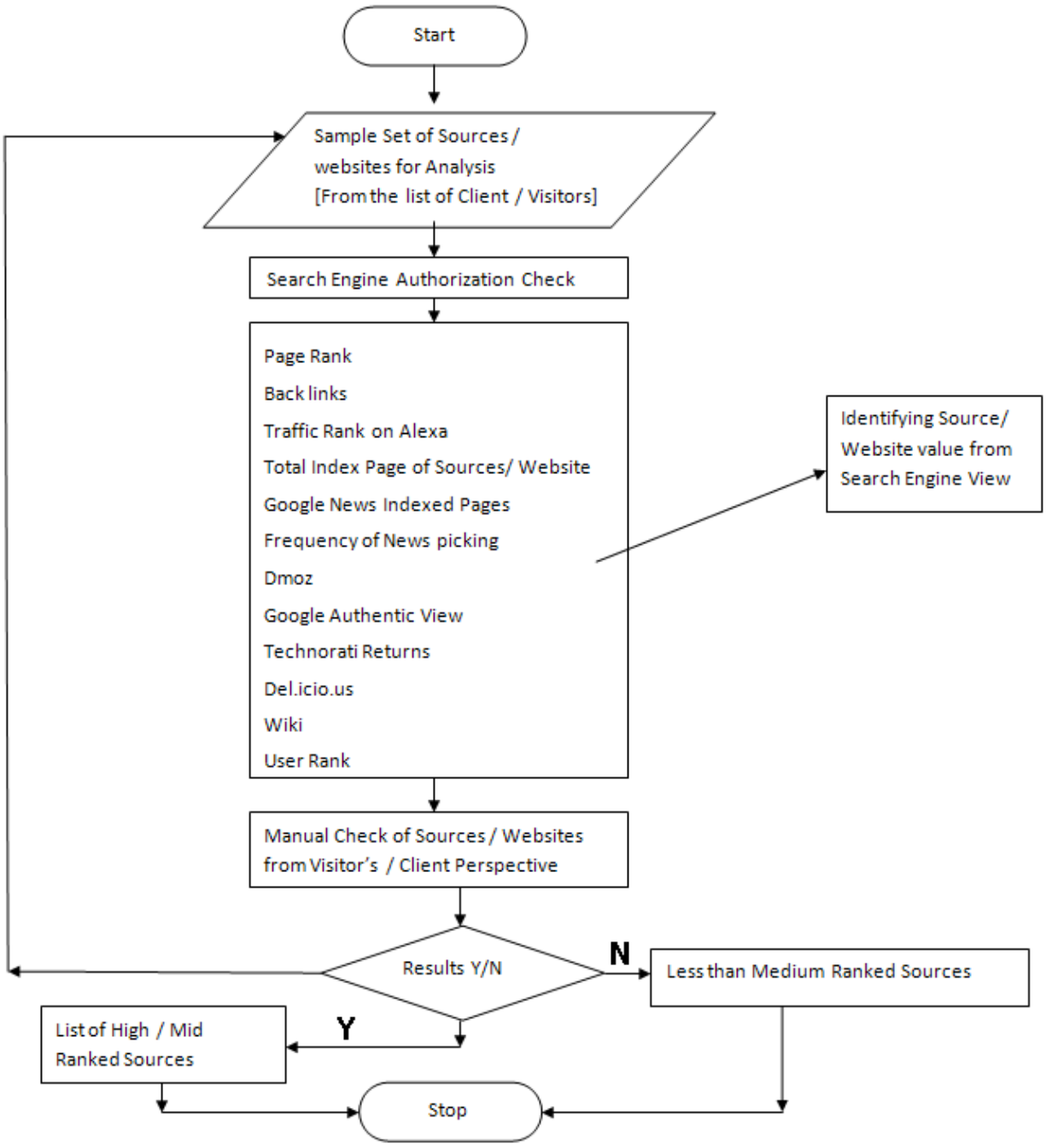
Source Rank Factors

Sr.No	Factor	High	Mid	Low
		Threshold Values		
1	Google PR (Page Rank)	7, 8, 9, 10	4, 5, 6	0, 1 ,2, 3
2	Back Links (Yahoo)	501+	51-500	0-50
3	Traffic Rank (Alexa)	1-100000	100001-500000	500001+
4	Index Pages (Yahoo)	5001+	101-5000	0-100
5	Google News Index Pages	1001+	51-1000	0-50
6	Google News Pick Status	Hourly (1)	Weekly (1)Monthly	No
7	Dmoz	Y (1)	N (0)	N (0)
8	Google Authentic View	Y (1)	N (0)	N (0)
9	Technorati Returns	1000+	51-1000	0-50
10	User Priority	Y (1)	N (0)	N (0)
11	Del.icio.us	Y (1)	N (0)	N (0)
12	Wiki	Y (1)	N (0)	N (0)

The threshold value chart for all the 11 parameters is shown above. The values scarped from the web is then compared with the ranges specified and assign the parameter with High, Mid or Low**.**

"Website / SOURCE RESEARCH" is a process to get refined and authenticated results from different search engines.

Process Flow



Basic Query Process Model

Whenever any query is executed on any search engine they act as

Step 1: Search Engine's recognize the region of the query i.e. from which region or country that query is coming.
Step 2: After regionalizing the query search engines detects the available or free datacenter for that query.
Step 3: Index server or index datacenter is database contains all the index pages of the web which crawled and stored by search engines.
Step 4: Document server is temporary server contains those entire web pages have that search pattern which we passed as a Search Query above.
Step 5: Algorithm applies on the document (web pages) exists in the web server.
Step6: Results returned on the basis of Algorithm developed by any search engine

The current algorithm is leveraging parameters with primary focus on Authoritativeness and relevancy: Stack 1 (Wiki, Dmoz, Google, Del), Stack 2 (Google PR, Alexa, Yahoo index etc)

- Initiate Ranking for ALL covered Sources
- Ability to slice and dice Sources by sector, customer, region etc to know the top ranked Sources to derive 'defaults'
- Prioritize Maintenance work based on the Rank

Leverage Source Rank in Editor's Work Flow given there is direct correlation of the Rank with Sources from which documents are marked for digests.

Enhance the current algorithm to cover a combination of authoritativeness and obscurity

- All of the above.
- Scrape the surface to uncover 'combination' Sources to provide differentiated coverage.
- Explore if this model can be a candidate for patenting/white paper etc.

RESULTS:

INPUT:

Source / Website is an input for the tool under process and source research document contains some searching guide lines so that one can research new source through search engines in an effective way

OUTPUT:

Sno	Homepage	Source Rank
37002	http://www.financial-gauges.com/	Low
37003	http://www.floorfacts.com/flooring-blog/	Low
37004	http://www.foamex.com/	Mid
37005	http://www.forrester.com/	High
37006	http://www.gartner.com/	High
37007	http://www.geekstreak.com/	Mid
37008	http://www.guideline.com/	Low
37009	http://www.homebuildingshow.co.uk/	Low
37010	http://www.imsa.edu/	High
37011	http://www.infectioncontroltoday.com/	Mid
37012	http://www.maketraveltrip.com/	Low
37013	http://www.mattfarina.com/	Low
37014	http://www.mckinsey.com/	High
37015	http://www.mimeetings.com/	Low
37016	http://www.molagers.org/	Low
37017	http://www.nahbrc.org/	Mid
37018	http://www.naturalproductsmarketplace.com/	Mid
37019	http://www.nepc.com/	Mid
37020	http://www.nexans.com/	Mid
37021	http://www.nrtinc.com/	Low
37022	http://www.otis.com/	Mid
37023	http://www.panhandleenergy.com/	Low
37024	http://www.phoneplusmag.com/	Mid

37025	http://www.progress-energy.com/	Mid
37026	http://www.proton.com/	Mid
37027	http://www.pzena.com/	Low
37028	http://www.reichhold.com/	Mid
37029	http://www.rhondda-cynon-taff.gov.uk/	Mid
37030	http://www.richardwinters.com/	Low
37031	http://www.rrb.gov/	Mid
37032	http://www.sbsbroadcasting.com/	Low
37033	http://www.schneider-electric.com/	High
37034	http://www.shba.com/	Mid

Discussion and Conclusion

Source Rank is a method devised by our Research to measure the importance of a web page (Source) According to certain parameters followed by different search engines and other authorized resources on the web. The Source Rank web services are devised to compute the Rank of the sources. The primary aim is to provide comfort to the sales team where they can trust few sources (High Rank Sources) and few they know are not that good (Low Rank).Another aim is to support the efficiency of work flow by reducing the number of documents by filtering on the basis of Source Rank. Source Rank Web service will determine the Rank of source on the basis of their popularity and relevancy on the web.

This will help the industries working under data extraction using web to rank the sources as per there business needs.. A source with low PR can be ranked as high if a person who writes the article is an expert in his domain.

References

[1] S. Brin, "The Anatomy of a Large-Scale Hypertextual Web Search Engine" (1998).

[2] Jon M. Kleinberg , Authoritative Sources in a Hyperlinked Environment, Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, and as IBM Research Report RJ 10076, 1998.

[3] Matt Cutts blog, http://www.mattcutts.com/blog/seeing-nofollow-links/

[4] Gyöngyi, Zoltán; Hector Garcia-Molina, Jan Pedersen "Combating Webspam with rustRank" Stanford University.

[5] www.dmoz.org. Human edited directory of the web.

[6] Gyongyi, Berkhin, Garcia-Molina, and Pedersen,"Link Spam Detection Based on Mass Estimation" technical report , Stanford University, Oct. 31, 2005.

[7] Mohhammad A Tayebi. S. Mehdi Hashemi, Ali Mohades, "B2Rank: An Algorithm for ranking Blogs Based on Behavioral Features",Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence , IEEE Computer Society Washington, DC, USA.

[8] Lan Nie Baoning Wu Brian D. Davison, A Cautious Surfer for PageRank, Department of Computer Science & Engineering Lehigh University, Bethlehem, USA, 2007.

[9] http://Wikipedia.org/

[10] http://google.co.in/

[11] http://del.icio.us/

[12] http://yahoo.com/

[13] Alexa, http://www.alexa.com/

[14] http://www.dmoz.org

[15] http://www.technorati.com/

Bharat Bhushan received the M Sc (Physics), from Panjab Univ. Chandigarh and M.Sc. (Comp. Sc.), MCA degrees from Guru Jambeshwar University. Respectively. Presently working as Head, Department of Computer Science and Applications, Guru Nanak Khalsa College, Yamuna Nagar (affiliated to Kurukshetra University, Kurukshetra-Haryana, India) and senior most teacher of computer science in Haryana since 1984. He is a member of Board of Studies of Computer Science, Kurukshetra University and member of Advisory Board of educational programme (EDUSAT) launched by Govt. of Haryana to impart online education. His research interest includes Software engineering, Digital electronics, networking and Simulation Experiments.

Narender Kumar received B.Sc. (Computer Science & Application) degree from Guru Nanak Khalsa College, Yamunanagar Affiliated to Kurukshetra University and MCA Degree from Ch. Devil Lal Post Graduate Regional Center of Kurukshetra University. He has Two Year Teaching Experience & presently working as a senior research analyst.