Study on Page Ranking Algorithms for Search Engine Optimization

Tin Tin Yu
University of Computer Studies, Mandalay
baybayyu@gmail.com

Tin Tin San
University of Computer Studies, Mandalay
tintinsan.tintinsan@gmail.com

Abstract

As the size of the World Wide Web increase and search engines play an important part for the retrieving information. Search Engines provide the gateway for users trying to explore information from web pages. The purpose of web search engines is to return web page lists that are relevant to the user query. The problem with web search relevance ranking is estimating the result list and its relevance with the user query. Since Search Engine Optimization (SEO) has become an important part of search engine marketing, the evolution of page rank algorithms have been focused on information retrieval research fields. In this paper, the top two algorithms namely, HITS and Page Rank algorithms have been reviewed and compared in this paper.

Keyword: World Wide Web, Search Engine, Search Engine Optimization, Page ranking Algorithms.

1. Introduction

World Wide Web provides us with necessary data digitally available as hypertext. Data may include web pages, images, information and text. For this reason, this hypertext pool is dynamically changing and it is more difficult to find useful information. The economic importance of web will enhance the academic interest.

Today search engine and tools are plagued by the four major problems: (1) the abundance problem, that is, the phenomenon of hundreds of irrelevant documents being returned in response to a search query, (2) limited coverage of the web, (3) a limited query interface that is based on syntactic based on keyword-oriented search and (4) limited customization to individual users[4].

Clustering and Classification have been useful and active areas of machine learning research that promise to help us cope with the problem of information overload on the internet.Clustering techniques become one of the alternative solutions for the problem of search engine [5]. Clustering provided an organized way to manage our search engine. Cosine similarity is also widely used in text or documents clustering. Ranking of the documents is done using their similarity values. The top ranked documents are regarded as more relevant to the query.The issues of improving search engines have been solved by employing classification. Classification deals with such type of problem that the retrieved results from traditional search engines are topic-independent[7].

2. Related Work

There has been considerable research on ranking results. We collect related work in this section to provide an overview of web search engine and ranking algorithms that are the most relevant or most closely related the content presented in this paper.

Search research on the web has a short and concise history. The World Wide Web Worm (WWWW) [7] was one of the first web search engines. It was subsequently followed by several other academic search engines, many of which are now public companies. Compared to the growth of the Web and the importance of search engines there are precious few documents about recent search engines [6]. According to Michael Mauldin (chief scientist, Lycos Inc) [3], "the various services (including Lycos) closely guard the details of these databases". However, there has been a fair amount of work on specific features of search engines. Especially well represented is work which can get results by post-processing the results of existing commercial search engines, or produce small scale "individualized" search engines. Finally, there has been a lot of research on information retrieval systems, especially on well controlled collections.

Information Retrieval (IR) Systems are the predecessors of Web and search engines. These systems were designed to retrieve documents in curate digital collections such as library abstracts, corporate documents, news, etc. Traditionally, IR relevance ranking algorithms were designed to obtain high recall on medium-sized document collections using long detailed queries. Furthermore, textual documents in these collections had little or no structure or hyperlinks. Web search engines incorporated many of the principles and algorithms of Information Retrieval Systems, but had to adapt and extend them to fit their needs. Early Web Search engines such as Lycos and AltaVista concentrated on the scalability issues of running web search engines using traditional relevance ranking algorithms. Newer search engines, such as Google, exploited web-specific relevance features such as hyperlinks to obtain significant gains in quality. These measures were partly motivated by research in citation analysis carried out in the bibliometrics field [11].

Page Rank is one of the methods Google uses to determine a page's relevance or importance. It is only one part of the story when it comes to the Google listing[12].

3. Web Search Engine

We use Search Engines to search for information across the Internet. Internet being an ever-expanding ocean of data, their importance grew with every passing day. The diversity of the information itself made it necessary to have a tool to cut down on the time spent in searching[2].

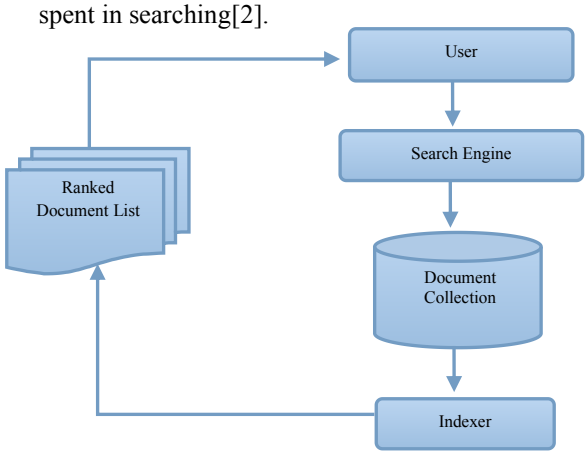

Figure 1: Architecture of Search Engine

The figure1 shows the architecture of search engine.When user types something in the search

engine box the search engine processes the user request by matching the user query with the results stored in the database. The results are stored in the database in the form of web pages. These web pages are ranked on the basis of content of the web page, relevant keywords used in the web pages, the frequency of occurring of keywords in the web page [15]. If the title, description, content of the web page is more relevant and important then web pages are listed at the top. The title or description of the web sites should appear to user as the useful link because the users normally visit or attempt to click on the web pages that are given at the top. The web pages are ranked on basis of the numbers assigned to these web pages. The web pages are stored in the database and retrieved with help of search engine[3]. These are the technical aspects of SEO. Page ranking algorithms are used and revise to produce the more optimized accurate and relevant resulted list of user query. In the rest of paper, we discuss the popular page ranking algorithms.

4. Page Ranking Algorithms

The World Wide Web contains an enormous amount of information, but it can be exceedingly difficult for users to locate resources that are both high in quality and relevant to their information needs. Issues that have to be dealt with are the detection of relevant information, involving the searching and indexing of the Web content, the creation of some meta knowledge out of the information which is available on the Web, as well as the addressing of the individual users' needs and interests, by personalizing the provided information and services. In this paper we discuss mainly two algorithms and other related ranking algorithms [1].

Search engines use two different kinds of ranking factors: query dependent factors and query independent Factors. Query-dependent are all ranking factors that are specific to a given query, while query-independent factors are attached to the documents, regardless of a given query. Query-dependent factors used by search engines are measures such as word documents frequency, the position of the query terms within the document or the inverted document frequency, which are all measures that are used in traditional Information Retrieval. Some of the query independent factors are Link popularity, click popularity and up to dateless etc. Ranking algorithms based on link popularity, falls under link based ranking algorithm category [10].

Two popular web page ranking algorithms are HITS and Page Rank. HITS emphasizes mutual reinforcement between authority and hub web pages, while Page Rank emphasizes hyperlink weight normalization and web surfing based on random walk models.

4.1. Hypertext Induced Topics Search

Hypertext Induced Topics Search (HITS) is developed by Jon Kleinberg. It uses hubs and authorities to define a recursive relationship between web pages. The algorithm performs a series of iterations, each consisting of two basic steps:

Step1- Authority Update: Update each node's Authority score to be equal to the sum of the Hub Scores of each node that points to it. That is, a node is given a high authority score by being linked to by pages that are recognized as Hubs for information.

Step2- Hub Update: Update each node's Hub Score to be equal to the sum of the Authority Scores of each node that it points to. That is, a node is given a high hub score by linking to

nodes that are considered to be authorities on the subject.

As in figure 2, a hub links to the many authorities and an authority is linked to by many hubs. The role of hub is to advertise the authoritative pages. They contain useful links towards the authoritative pages[13]. In other words, hubs point the search engine in the "right direction". In real life, when you buy a car, you are more inclined to purchase it from a certain dealer that your friend recommends. Suppose in a car show room, the authority would be the car dealer, and the hub would be your friend. You trust your friend, therefore you trust what your friend recommends.

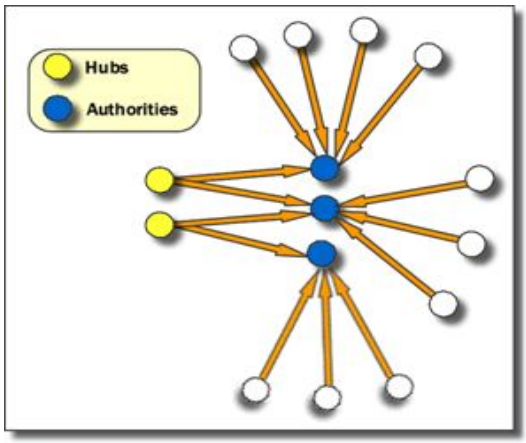

Figure 2: Hubs and Authorities

HITS has the ability of ranking page according to the query topic. Because of this fact, HITS can give the result list that is more relevant with the user query, in other words, more relevant authority and hub pages. This type of ranking may also be combined with the information retrieval based ranking techniques.

HITS algorithm has some limitation. First of all, it does not have the anti-spam capability. It is quite easy to influence HITS by adding out-links from one's own page to point to many good authorities. This boosts the hub score of the page[1]. Because hub and authority scores are interdependent, it in turn also increases the authority score of the page. Another problem of HITS is topic drift. In expanding the root set, it can easily collect many pages (including authority pages and hub pages) which have nothing to do the search topic because out-links of a page may not point to pages that are relevant to the topic and in-links to pages in the root set may be irrelevant as well because people put hyperlinks for all kinds of reasons, including spamming. The query time evaluation is also a major drawback. Getting the root set, expanding it and then performing Eigenvector computation are all time consuming operation[16].

4.2. Page Rank Algorithm

Page Rank is an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, Page Rank is an excellent way to prioritize the results of web keyword searches. For most popular subjects, a simple text matching search that is restricted to web page titles performs admirably when Page Rank prioritizes the results. For the type of full text searches in the main Google system, Page Rank also helps a great deal.

The rank scores can be expressed as percentage or logarithmic. Google uses a logarithmic scale to express the rank scores of sites. From the point of view of link analysis,the more link that the site is pointed , the more rank score it may have. On the other hand, every site has their own rank score according to their authority.Thus, the rank score can not be considered depend only on the number of links.It needs to consider total rank score of links the that point to the site. In figure 3, Page

C has a higher rank score than page E, even though there are more links to E than C. The only one link that point to C has high rank score[14].

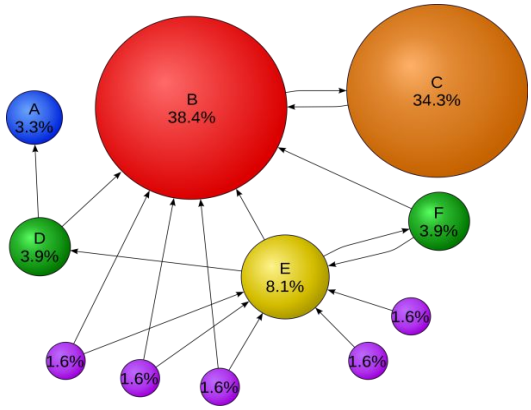

Figure 3: Mathematical page rank for a simple network

The main advantage of Page Rank is its ability to fight spam. Recognizing and fighting spam is an important issue in Web search. A page is important if the pages pointing to it are important. Since it is not easy for Web page owner to add in-links into his/her page from other important pages, it is thus not easy to influence Page Rank. As Page Rank is a query independent algorithm i.e. it pre-computed the rank score. So it takes very less time .Both these two advantages contributed on the basis of the the popularity of a page. For calculating rank value of a page, it consider the entire web graph, rather than a small subset, it is less susceptible to localized link spam[17].

The main disadvantage of Page Rank is that it favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing site.In Page Rank, the web pages are ranked according to the number of clicks made on that particular web page but this may lead to illegal ranking of web pages . In other word, whenever a query is given the pages that are satisfying the query are presented according to the rank of the page. The top most one will be given highest priority. The highest priority is because the number of clicks on that particular web page are more without concerned with the content that is present in that particular web page. For this purpose the ranking should be given according to the content present in the web page rather than the number of clicks made on that particular web page. Because if a wrong page is presented to end user then he will browse the page which will increase the click count of the traced page which is wrong. This leaves the web page with highest priority. For this purpose it is better to rank pages according to the content in the web page. This leads to the combination of text mining with web mining.

4.2.1. Extension of Page Rank Algorithms

HITS and Page Rank, are commonly used in various search engine. Several algorithms have been developed to improve the performance of these methods. Page Rank algorithm is more popular than the HITS because of its features especially for spam fighting and it can reduce time consuming. Page Rank algorithm is used by the famous search engine such as Google. In this section, we describe the two extension method of Page Rank algorithm. In Weight Page Rank (WPR)method,it takes into account the importance of both the in-links and the out-links of the pages and distributes rank scores based on the popularity of the pages. Simulation studies using the website of Saint Thomas University show that WPR is able to identify a larger number of relevant pages to a given query compared to standard Page Rank.Some papers show the results of WPR performs better than the conventional Page Rank algorithm in terms

of returning larger number of relevant pages to a given query[8].

Another extension of Page Rank algorithm is Time Page Rank (TPR) algorithm. The motivation behind TPR is depend on the user interest. Most users are interested to the latest information. This method emphasizes to modify the weak point of favoring the older page in Page Rank algorithm. Since some older page exists in Web for a long time, their rank score might be higher than the new pages which are high quality and which can give latest information that users want. TPR uses the time function f(t) (0<f(t)<1)based on the probability, where t is the difference value between the current time and the last updated time for a page. It defined two value f(t) and 1-f(t), the first f(t) value is to follow the actual line of the page and the second 1-f(t) is to jump to a random page without a link. If the page is old and it was not updated for a long time, the value of 1-f(t) should be large and otherwise it should be small[9].

4.3.Comparison of HITS and Page Rank Algorithm

Both page rank and HITS algorithm are different link analysis algorithms that employ different models to calculate web page rank. Each ranking algorithm provides a definite rank to resultant web pages. Between the two algorithm, Page Rank is more popular than the HITS algorithm because of its features especially time saving and spam fighting. On the other hand, the result of HITS can recommend for relevancy of user query because it is query dependent algorithm. On the basis of this study, we summarized the feature of HITS and Page Rank algorithm as in Table 1.

Table 1. Comparison of HITS and Page Rank Algorithm

Features	Algorithms	
	HITS	Page Rank
Major Strength	Query dependent	Fighting spam
Major Weakness	Time consuming	Query independent
Technique	Web Structure Mining and Web Content Mining	Web Structure Mining
Methodology	Scores are computed on the hub and authorities	Scores are computed on the number of clicks
Input Parameter	Content, back link, forward links	Only back links
Domain Area	Twitter, IBM	Google
Relevance of user query	More	Less
Complexity	<O (log N)	O (log N)
Dependability	Hubs and Authorities	Stability time of web page
Fight Spam	No	Yes
Computation Time	Online	Offline
Proposed By	Jon Kleinberg	Sergey Brin and Larry Page

5. Conclusion

This paper discussed about the studying on the two popular ranking algorithms namely HITS and Page Rank for search engine optimization, advantages and disadvantages of these two algorithm, the extension of Page Rank algorithm and finally we made comparison of features of these two algorithms. The difference between rankings produced by difference algorithms reflects slightly different but useful emphasis. Basically, in degree and out degree are fundamental important in web ranking.

After the study analysis of ranking algorithms, we can conclude that an efficient web page ranking algorithm should be able to solve the challenges and limitation of existing application efficiently and it should compact with global standards of web technology in terms of accuracy, relevancy and response time of the results.The further enhancements and researches to the ranking algorithms should be focused continuously in order to get the more optimized ranking results.

As for the future work, we have plan to study the recognizing and fighting spam techniques. If we could add fighting spam feature to the HITS ranking algorithm by implementing as a hybrid algorithm, the more relevant ranked results with the user query would be given.

References

[1]Bing Lu, "Web Data Mining", Exploring Hyperlinks, Contents, and Usage Data, Department of Computer Science, University of Illinois, Chicago, liub@cs.uic.edu

[2]I.va Gregurec, Petna Grd "Search engine optimization website analysis of selected faculties in Croatia". Central European conference on information and intelligent systems, Sept 19-21, 2012 PP-213.

[3]Mauldin, Michael L. Lycos "Design Choices in an Internet Search Service", IEEE Expert Interview http://www.computer.org/pubs/expert/1997/trends/x1008/mauldin.html

[4]McBryan 94, Oliver A. McBryan. GENVL and WWWW: "Tools for Taming the Web. First International Conference on the World Wide Web", CERN, Geneva (Switzerland), May 25-26-27, 1994.

[5]Minky Jindal and Nisha Kharb ,"k-mean s clustering technique on search engine data set using data mining tool", CSE/IT Department, ITM University, Sector-23A, Gurgaon, India.

http://www.cs.colorado.edu/home/mcbryan/ypapers/www94.ps

[6]Pinkerton 94, Brian Pinkerton, "Finding What People Want: Experiences with the Web Crawler"., The Second International WWW Conference Chicago, USA, October 17-20, 1994.

[7]Rajesh Singh and Reamer (Rajasthan), "An approach for search engine optimization using classification - a data mining technique ", Scholar in Bhagvant University, S.K. Gupt, A.P.,CSED, BIET, Jhansi (U.P.)

[8]Wenpu Xing , Ali Ghorbani, "Weighted Page Rank Algorithm", Faculty of Computer Science, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada

[9]X.Li, B.Liu and P.S. Yu. Time Sensitive Ranking with Application to Publication Search. Forthcoming paper.2006

[10]http://searchengineland.com/guide/seo/types-of-search-engine-ranking-factors

[11]http://www.wordstream.com/articles/internet - search-engines-history

[12]http://www.db.stanford.edu/%7Ebackrub/google.html

[13]http://soltisconsulting1.files.wordpress.com/2013/08/sentinel-visualizer-3.jpg

[14]http://en.wikipedia.org/wiki/File:PageRanks-Example.svg

[15]http://www.aukbc.org/research-areas/nlp/projects/sengine.html

[16]http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

[17]http://en.wikipedia.org/wiki/PageRank