Journal of Physical Science and Innovation
ISSN: 2277-0119
Volume 9, No. 2, 2017

WEB PAGE CONTENT BASED DOCUMENT RANKING FOR SEARCH ENGINE OPTIMIZATION

[1]Agbator Lawrence & [2]Akhetuamen Sylvester
[1]Department of Computer Science, Edo State Institute of Technology and Management, Usen, Edo State
[2] Department of Computer Science, Federal Polytechnic Auchi
Email: mail4agbator@gmail.com, divinelaw1@yahoo.com

ABSTRACT

Search engine optimization is the process of enhancing the efficient ranking of a website or a web page in a search engine's normal search results. In information retrieval from the web by search engine, the more highly ranked a page or site is on the search results page, which will in turn make it to appear more frequently on the search results list, the more visitors it will receive from the search engine's users. Search engine optimization may target different kinds of search, including image search, local search, video search, academic search, news search and industry-specific vertical search engines. As an Internet marketing strategy, Search engine optimization considers how search engines work, what people search for, the actual search terms or keywords typed into search engines and which search engines are preferred by their targeted audience. Authors publish in a wide variety of formats, which includes deliberately misleading search platforms and hence increasing the chance of retrieving irrelevant web pages and this action has led to the degradation of search result. This paper presents a content-based document ranking method to counter this phenomenon and help to improve the educational relevance of information retrieved using search engines by concerned users.

Keywords: Walled-Search, Electronic-Publications, Web, Search-Engine, Retrieval-Process, Ranking, Algorithm, Random-Surfer

INTRODUCTION

Search engines is where we start just about all endeavors on the World Wide Web these days, it is important that we stop and think about the tools we are using and how they can affect our productivity and even teaching and learning. Document retrieval on the World Wide Web (WWW), the world's largest collection of documents, is a challenging and important task. The scale of the www is immense, consisting of at least 20 billion publicly visible web pages distributed on millions of server worldwide. (Maurice 2010). Many years ago man had has to depend only on the four walls of a library complex to carry out search for information for various reasons and most specifically for academic reasons. Those certainly were the days looking up something by keyword in the library's card catalog and hoping the book you want isn't already being read by someone else. We really were limited in our quests for new knowledge to what was carefully cataloged by librarians. Then, in the mid 90's, several different digital curations started taking place. Online directories of links that were handmade could be browsed by sometimes thousands of topics. You can still visit the Yahoo! Directory, one of the oldest and which was a gold mine for anyone paying for a link back in the day. These were great because in the most trusted directories, you could be assured that the site was legitimate, spam free, and relevant to the topic you were looking for. But with millions of new pages of content being created each and every day, the directory system just couldn't keep up. With the birth of the search engine, we can find any and all relevant content with whatever search term we want.

However, at least four major problems are fundamentally associated with today's search engine namely;

1. We typically only ever make it to the top few listed sites in search results, so how much more content are we missing out on?
2. There are tons of less than ethical websites out there that use lots of techniques to artificially inflate their search engine rankings (Web Spamming).

3. Certainly when working with students, most search engines can serve up results that definitely aren't appropriate.
4. Plus, do you trust everything you find on search engines

The Web organises information by employing a hypertext paradigm. Users can explore information by selecting hypertext links to other information. As the web continues its explosive growth, the need for searching tools to access the web is increasing,

THE WALLED SEARCH

Something that is likely to increase is the use of "walled search" environments. This lets you filter out search results based on certain types (ie. videos, images, etc.) and can also limit search to only a pre-determined set of sites. Besides the built in filters in Google and Bing, for instance you might find these education specific tools useful:

- ❖ SweetSearch.com – A safe search engine for students
- ❖ AppleEngine.com - A search engine for teachers to find free resources
- ❖ WatchKnow.org - An organized (and searchable) directory of hand-picked educational videos

When responding to queries, the goal of an information retrieval system ranging from web search, to desktop search, to call center support is to return the results that maximize user utility. So, how can a retrieval system learn to provide results that maximize utility? The conventional approach is to optimize a proxy measure that is hoped to correlate with utility. A wide range of measures has been proposed to this effect

THE WEB AND ELECTRONIC PUBLICATIONS

When we consider the search engines on the web today, we conclude that they continue to use indexes which are very similar to those used by the librarians a century ago. What has changed then? Three dramatic and fundamental changes have occurred due to the advances in modern computer technology and the boom of web. First, it became a lot cheaper to have access to various source of information (baeza-yates and Ribeiro-Note, 1999). This allows reaching a wider audience than ever possible before. Second, advances in all kinds of digital communication provided greater access to networks. This implies that the information sources are available even if distantly located and that the access can be quickly (frequently, in a few seconds) third, the freedom to post whatever information someone judges useful has greatly contributed to the popularity of the web (Baeza-yates and Ribeiro- 1999). For the first time in history, many people have free access to a large publishing medium. Fundamentally, low cost, greater access publishing freedom has allowed people to use the web and larger digital libraries, which techniques will allow retrieval of higher quality? Secondly, with the ever- increasing demand for access, quick response is becoming more and more a pressing factor. Thus, which techniques will yield faster indexes and smaller query response time? Thirdly, better understanding of the user behaviour affect the design and deployment of new information retrieval

INFORMATION RETRIEVAL (IR) SYSTEM

Information retrieval (IR) is the process of representing, storing, organizing and accessing information items. The representation should provide the user with easy access to information of interest, (Baeza-Yates and RIbeiro-neto, 1999) for example given an information need by a user, how we characterize a simple query that will ensure that information retrieval system retrievals exactly the relevant document, how will the semantic relationship between the query and information required be represented in a model? This is the problem of characterization of user information need.

SEARCH ENGINE USER BEHAVIOR

Findings on search engine user behavior indicate that users are not willing to spend much time and cognitive resources on formulating search queries (Machill, Neuberger, Schweiger, & Wirth, 2004), a fact that results in short, unspecific queries (e.g., Höchstötter & Koch, 2008; Jansen & Spink, 2006) .

While this surely applies to Web searching, similar behavior can also be found in other contexts, such as in scientific searching (Rowlands et al., 2008) or library searches (Hennies & Dressler, 2006). Search queries tend to be very short and do not show variations over a longitudinal period. Nearly half of the search queries in Web searching still only contain one term. On average, a query contains between 1.6 and 3.3 terms, depending on the query language. Höchstötter and Koch (2008) gave an overview of different studies measuring users' querying behavior (including query length and complexity). Most searching persons evaluate the results listings very quickly before clicking on one or two recommended Web pages (Hotchkiss, Garrison, & Jensen, 2004; Spink & Jansen, 2004). Users consider only some of the results provided, mainly those results presented at the top of the ranked lists (Granka, Joachims, & Gay, 2004; Pan et al., 2007), and even more prefer the results presented in the "visible area" of the results screens, that is to say, the results visible without scrolling down the page (Höchstötter & Lewandowski, 2009). Lastly, results selection is determined by presenting some results in a different manner than the other results that is, emphasizing certain results by means of the use of color, frames, or size

THE RETRIEVAL PROCESS

To describe the retrieval process, we use simple and generic software architecture as shown in Figure 1 below. First of all, before the retrieval process can even be initiated, it is necessary to define the database. This is usually done by the manager of the database, which specifies the following :(a) the documents to be used, (b) the operations to be performed on text, and (c) the text model (i.e. , the text structure and what element can be retrieved). The text operations transform the original documents and generate a logical view of them.(Baeza-Yates and RIbeiro-neto, 1999) once the logical view of document is defined, the database manager (using the DB manager module) builds an index of the text. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the inverted file as indicated in figure below. The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times, given that a document database is indexed, the retrieval process can be initiated. The user first specifies a user need, which is then parsed and transformed by the same text operations applied to the text, and then query operations might be applied, before the actual query, which provides a system representation for the user need to be generated. The query is then processed to obtain the retrieved documents. Fast query processing is made possible by index structure previously built.(Baeza-yatae and RIbeiro-neto, 1999). The retrieved documents are ranked according to a likelihood of relevance. The user then examines the set of ranked documents in the search for useful information. At this point, he might pinpoint a subset of the document seen as definitely of interest and initiate a user feedback cycle. In such a cycle, the system uses the documents selected by the user to change the query formulation. Hopefully, this modified query is a better representation.

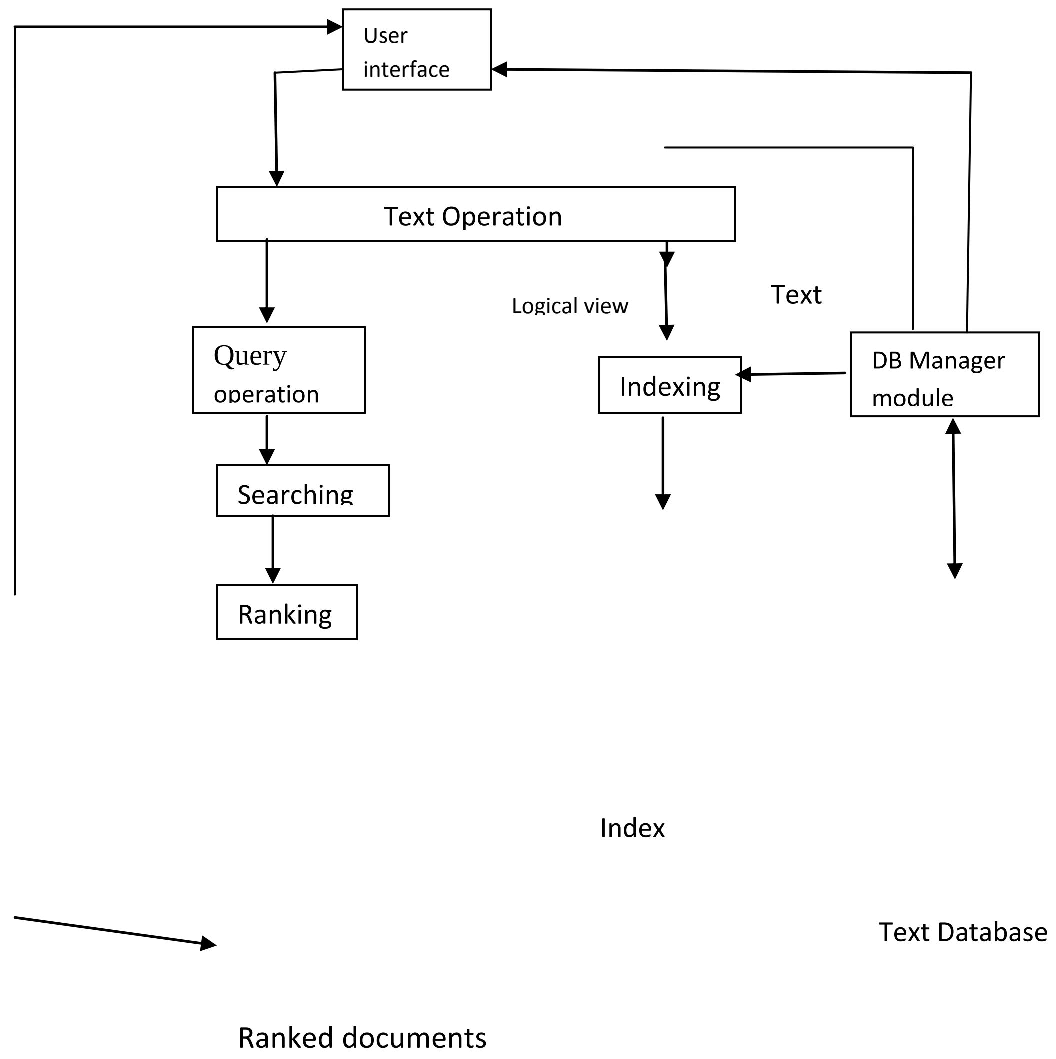

Figure 1 Logical view of information retrieval cycle

ANALYSIS OF THE RANKING ALGORTHM OF EXISTING SYSTEM

Existing systems uses the link graph of the web by creating a map of the hyperlinks in web documents. This link graphs is used to establish a popularity measure of web pages which translate into importance or high rank of the page.(Lawrence and sergy,1998). A popular and infact our case study hear is the page rank developed by Sergey brin and Lawrence page of Google.

Pagerank Algorithm

The original page rank algorithm was described by Lawrence page and Sergey Brin (1998) in several publications. it is given by

PR(A)=(1-d)+d(PR(T1)/C(T1)+...+PR(Tn)/c(Tn))

Where PR(A) is the pages rank of pages pages A,
PR(Ti) is the page rank of pages Ti which link to pages A
C(Ti) is the number of outbound link on a page Ti and
d is a damping factor which can be set between 0 and 1

So, first of all, we see that pagerank does not rank web site as a whole, but is determined for each page individually. Further, the page rank of page A is recursively defined by the pageranks of those pages which link to page A..(Lawrence and Sergey,1998) the pagerank of pages Ti which link to pages A does not influence the pagerank of page A uniformly, within the pagerank algorithm, the pagerank of page T is always weighted by the number of outbound links C(T) on page T. This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T..(lawrenec and segey,1998). The weighted pagerank of pages Ti is then added up, the outcome of this is that an additional inbound link for page A will always increase page A's pagerank. Finally, the sum of weighted pageranks of all pages Ti is multiplied with a damping factor d which can be set between 0 and 1 . Thereby, the extent of page rank benefit for a page by another linking to it is reduced.

The Random Surfer Model

In their publications, Lawrence and Sergey (1998) gave a very simple intuitive justification for the pagerank algorithm with no regard towards content, the random surfer visit a web page with a certain probability which derives from the pages's pagerank. The probability that the random surfer clicks on one link is solely given by the number of links on that page. This is why one page's pagerank is not completely passed on to a page it links to, but is divided by the number of links on the page. So, the probability for the random surfer reaching one page is sum of probabilities for the random surfer following links to this page. Now, this probability is reduced by the damping factor d , the justification within the random surfer model therefore, is that the surfer does not click on an infinite number of links, but gets bored sometimes and jumps to another at random.(Lawrence and surgey 1998).

Our Proposed Ranking System

Our proposed ranking method comes from both the ideal in the work done by halter (1975), of specialty and non-specialty words which also is the underlying principle behind the ability of some word to be more relevant in a document than other, also the work done by Gianni Amati and van Rijisbergen in 2002, on probabilistic models of information retrieval based on measuring the divergence from randomness. This shows that words which bring little information are randomly distributed on the whole set of documents. The piosson distribution model used by both of them showed that the smaller this probability is , the less its token are distributed in conformity with the model of randomness and higher the informative content of term. Hence, determining the informative content of a term can be seen as an inverse test of randomness of term within a document with respect to the term distribution in the entire document collection.

Thirdly, studies from the distribution of words in large documents, has helped to ascertain the discriminative power of tokens. Based on these discoveries, we have been able to come out with an underlying principle for our content based ranking method, and that is co-occurrence of words in document collection. For a multi-word search, the situation is more complicated, now multiple hit lists must be scanned through at once so that hits occurring close together in a document are weighted higher than hits occurring far apart. The hit lists are matched up so that nearby hits is matched together. For every matched set of hits, proximity is computed; the proximity is based on how far apart the hits are in the document (or anchor) but is classified into 10 different value "bins" ranging from a phrase match to "not even close". (Lawrence and sergey 1998). Counts are computed not only for every type of hit but for every type and proximity. Every type and proximity pair has a type of the count-weights and the type-prox-weight. The counts are converted into counts-weights and we take the dot product of the count-weights and the type-prox-weights to compute an IR score. All of these numbers are matrices and can all be displayed with the search results using a special debug mode. These displays have been very helpful in developing the ranking system.

SUMMARY AND CONCLUSION

In modern information retrieval, attention is gradually shifting from the paradigm of using web technology to determine the behavior of software dedicated to the retrieval of sensitive contents in the web, to statistical evaluation of the information content of document copus. The paradigm shift is necessitated by the increase in the actions of publishers with sinister motive on the web through any of the search platforms. The web is unprecedented in many ways: unprecedented in scale, unprecedented in the almost-complete lack of coordination in its creation, and unprecedented in the diversity of backgrounds and motives of its participants. Each of these contributes to making web search different and generally far harder than searching "traditional" documents. The analysis of hyperlinks and the graph structure of the web have been instrumental in the development of web search. Such link analysis is one of many factors considered by web search engines in computing a composite score for a web page on any given query. Link analysis for web search has intellectual antecedents in the field of citation analysis, aspect of which overlap with an area known as biblometrics. Link analysis on the web treats hyperlinks from a web page to another as a conferral of authority. Clearly, not every citation or hyperlink implies such authority conferral; for this reason, simply measuring the quality of a web page by the number of in-inks (citations from other pages) is not robust enough. For instance, one may contrive to set up multiple web pages pointing to a target web page, with the internet of artificially boosting the latter's tally of in-links. The phenomenon is referred to as link spam. This is what web spammers in recent times have used to degrade the quality of search results from search platforms using the link structure to determine a relevant document to a user query. However, the link structures of the web still posses the strength to guide a crawler towards effective indexing of the web. In this paper we have tried to combine the lexical strength of the words in the interpretation of what is wholly contained in a document to a user query.

RECOMMENDATIONS

Search engines remain the entrance door to the World Wide Web. Therefore in view of the ease at which web spammer can mislead present search engines based on the too much dependence of their ranking algorithm on the web link structure, we hereby strongly recommend:

1. That one, present search engines ranking algorithm be built around web page content for textual pages, in other to guarantee retrieval of relevant information from the www which will in turn determine the quality of learning made possible if such retrieved materials are consulted for academic purpose.

2. Academic retrieval platform should be developed to integrate the content-based retrieval algorithms that the commercial search platforms may not be willing to adopt for economic reasons, profit maximisation and speed trade off.

REFERENCES

Amati, G and Van Rijsbergen, C. j., (2002) probabilistic models of information retrieval based on measuring divergence from randomness. ACM Transactions on Information systems, pages 357_389.

Baeza-Yates, R. and Ribeiro-Neto, B., (1999) Modern information Retrieval. USA. Addition Wesley.

Christine L.B, Anne J.G, Gregory H.L, Richard M, David G, Rich G, and patricia M. (2000). Evaluating Digital Libraries for Teaching and Learning in Undergraduate Education: A Case study of the Alexandra Digital Earth ProtoType-ADEPT. Retrieved 14th of April, 2010 from http://findarticles.com/p/articles/mi387/is_2_49/ai_72274394/pg_3?tag =content;col1.

Ethan T.(2007) Nine people, places and things that their names://blogs.static.mentafloss.com/BLOGS/archives/22707.HTML. retrieved 20th October. 2016.

Harte, P.S.(1975)A probabilistic approach to automatic keyboard indexing. Parti. On Distribution of specialty Words in a Technical Literature. Wiley Periodicals, inc., AWiley company

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 478-479.

Höchstötter, N., & Koch, M. (2008). Standard parameters for searching behaviour in search engines and their empirical evaluation. Journal of Information Science,34(1), 45-65.

Höchstötter, N., & Lewandowski, D. (2009). What Users See – Structures in Search Engine Results Pages. Information Sciences, 179(12), 1796-1812.

Hennies, M., & Dressler, J. (2006). Clients Information Seeking Behaviour: An OPAC Transaction Log Analysis, CLICK06, ALIA 2006 Biennial Conference.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, 42(1), 248-263.

Machill, M., Neuberger, C., Schweiger, W., & Wirth, W. (2004). Navigating the Internet: A Study of German-Language Search Engines. European Journal of Communication,19(3), 321- 347.

Rowlands, I., Nicholas, D., Williams, P., Huntington, P., Field house, M., Gunter, B. (2008).The Google generation: The information behaviour of the researcher of the future. Aslib Proceedings: New Information Perspectives, 60(4), 290-310.

Lewandowski. D. (2012) A Framework for Evaluating the Retrieval Effectiveness of Search Engines. Jouis, Christophe: Next Generation Search Engine: Advanced Models for Information Retrieval. Hershey, PA: IGI Global. http://www.igi-global.com/book/next- generation-search-engines/59723. Retrieved 20th July 2013.

Lawrence P. and Sergy B.,(1998) the anatomy of a large scale hyper textual web search engine. USA. Stanford press.

Maurice D. K. (2010). The size of the world wide web. Retrieved 14 April 2010 from http//www.worldwidewebsie.com/.

Micheal K, (1996). "Chiming in on yahoo's roar.' mEDIAWEEK,6(3):9-12.

Plachouras V., Ounis .i, and amati g., (2005) The static absorbig model for the web. *Journal of web Engineering, 4(2): 165,186.*

Reference to this paper should be made as follows: Agbator Lawrence, Akhetuamen Sylvester (2017) Web Page Content Based Document Ranking for Search Engine Optimization. J. of Physical Science and Innovation, Vol. 9, No. 2, Pp. 71-84