EFFICIENT RANKING ALGORITHM FOR ENHANCED ACCESS TO DIGITAL LIBRARIES ON THE WEB.

[1]Agbator Lawrence (BSC, Computer Science)
Mobil: 08051156364, Email: divinelaw1@yahoo.com
And
[2]Aghahowa A. (B.A, English)
And
[3]John Sule Imokhai (B.A, MLM)
[1] Department of Computer Science
[2]Department of General Studies,
[3]Library Department, Edo State Institute of Technology and Management, Usen, Edo State, Nigeria

ABSTRACT
Retrieving information in the World Wide Web, the world's largest collection of documents is a challenging and important task. The scale of the WWW is immense, consisting of at least twenty billion publicly visible web pages distributed on millions of servers world-wide. There is no enforcement on adherence to formal protocols to publish in the web. 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 results. This paper presents a content-based document ranking method to counter this phenomenon and seeks to increase the relevance of search result to a user query.

INTRODUCTION
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 20.09 billion publicly visible web pages distributed on millions of servers world-wide. ((Maurice 2010). Web authors follow few formal protocols, they often remain anonymous and publish in a wide variety of formats.

The Web organizes 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. Yahoo! is one of the big names in Web directories. A pair of Stanford graduate students founded Yahoo! in 1995.(Ethan, 2009). Recently, a host of new search and directory sites now offer a wide range of Web-searching services (Michael 1996). Examples include Google, mamma, Alta Vista, InfoSeek, Open Text and Excite.

INFORMATION RETRIEVAL (IR) IN THE LIBRARY
Libraries were among the first institutions to adopt IR systems for retrieving information. Usually, systems to be used in libraries were initially developed by academic institutions and later by commercial vendors. In the first generation, such systems consisted basically of an automation of previous technologies (such as card catalogs) and basically allowed searches based on author name and title. In the second generation, increased search functionality was added which allowed searching by subject headings, by keywords, and some more complex query facilities. In third generation, which is currently being deployed, the focus is on improved graphical interfaces, electronic forms, hypertext features, and open system architectures.(Baeza-Yates and Ribeiro-Neto, 1999)

THE WEB AND DIGITAL LIBRARIES
If 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 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 sources of information (Baeza-Yates and Ribeiro-Neto, 1999). This allows reaching a wider audience than ever possible before. Second, the advances in all kinds of digital communication provided greater access to networks. This implies that the information source is 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-Neto, 1999).

For the first time in history, many people have free access to a large publishing medium.

Fundamentally, low cost, greater access and publishing freedom have allowed people to use the web (and modern digital libraries) as a highly interactive medium. Such interactivity allows people to exchange messages, photos, documents, software, and videos and to chat in a convenient and low cost fashion. Further, people can do it at the time of their preference (for instance, you can buy a book late at night), which further improves the convenience of the service. Thus, high interactivity is the fundamental and current shift in the communication paradigm (Baeza-Yates and Ribeiro-Neto, 1999).

The three main questions that needs to be addressed are: First, despite the high interactivity, people still find it difficult (if not impossible) to retrieve information relevant to their information needs. Thus, in the dynamic world of web and large digital libraries, which techniques will allow retrieval of higher quality? Second, 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? Third, the quality of the retrieval task is greatly affected by the user interaction with the system. Thus, how will a better understanding of the user behaviour affect the design and deployment of new information retrieval strategies?

EDUCATIONAL APPLICATIONS OF DIGITAL LIBRARIES

Educational applications of digital libraries range from primary school through graduate school and across all disciplines. Faculty and librarians alike are concerned about ways to implement digital libraries in education. The Council on Library and Information Resources in 1999 held a meeting "to consider changes in the process of scholarship and instruction that will result from the use of digital technology and to make recommendations to ensure that libraries continue to serve the research needs of scholars." Among their recommendations was that institutions of higher education should "place more emphasis on training and support for faculty use of information and instructional technologies."(Christine, Anne, Gregory, Richard, David, Rich and Patricia, 2000).

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 characterise a simple query that will ensure that information retrieval system retrieves exactly the relevant document, how will the semantic relationship between the query and information required be represented on a model? This is the problem of characterization of the user information need.

DATA RETRIEVAL AND INFORMATION RETRIEVAL

Data retrieval, in the context of an information retrieval system, consist mainly of determining which documents of a collection contain the keywords in the user query which most frequently, is not enough to satisfy the user information need. The user of an IR system is concerned more with retrieving information about a subject than with retrieving data, which satisfy a given query (Baeza-Yates and Ribeiro-Neto, 1999). A data retrieval language aims at retrieving all objects, which clearly defined conditions such as those in a regular expression or in a relational algebra expression. Thus for a data retrieval system, a single erroneous object means total failure. For an information retrieval system, however, the retrieved object might be inaccurate and small errors are likely to go unnoticed. The main reason for this difference is that information retrieval usually deals with natural language text, which is always not well structured and could be semantically ambiguous. On the other hand, a data retrieval system such as a relational database deals with data that has a well-defined structure and semantics (Baeza-Yates and Ribeiro-Neto, 1999).

Data retrieval, while providing a solution to the user of a database system, does not solve the problem of retrieving information about a subject or topic. For an information retrieval system to be effective in its attempt to satisfy the user information needs the IR system must somehow interpret the contents of the information items (documents) in a collection and rank them to the degree of relevance to the user query. This interpretation of document content involves extracting syntactic and semantic

information from the document text and using this information to match the user information need. The difficulty is not only knowing how to extract this information but also knowing how to use it to decide relevance.(Amati and Rijsbergen, 2002). Thus the notion of relevance is at the centre of information retrieval. In fact, the primary goal of an information retrieval system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible.(Baeza-Yates and Ribeiro-Neto, 1999).

THE RETRIEVAL PROCESS

To describe the retrieval process, we use simple and generic software architecture as shown in figure 2 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 the text, and (c) the text model (i.e., the text structure and what elements 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 the documents 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 the 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. 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 the index structure previously built.(Baeza-Yates 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.

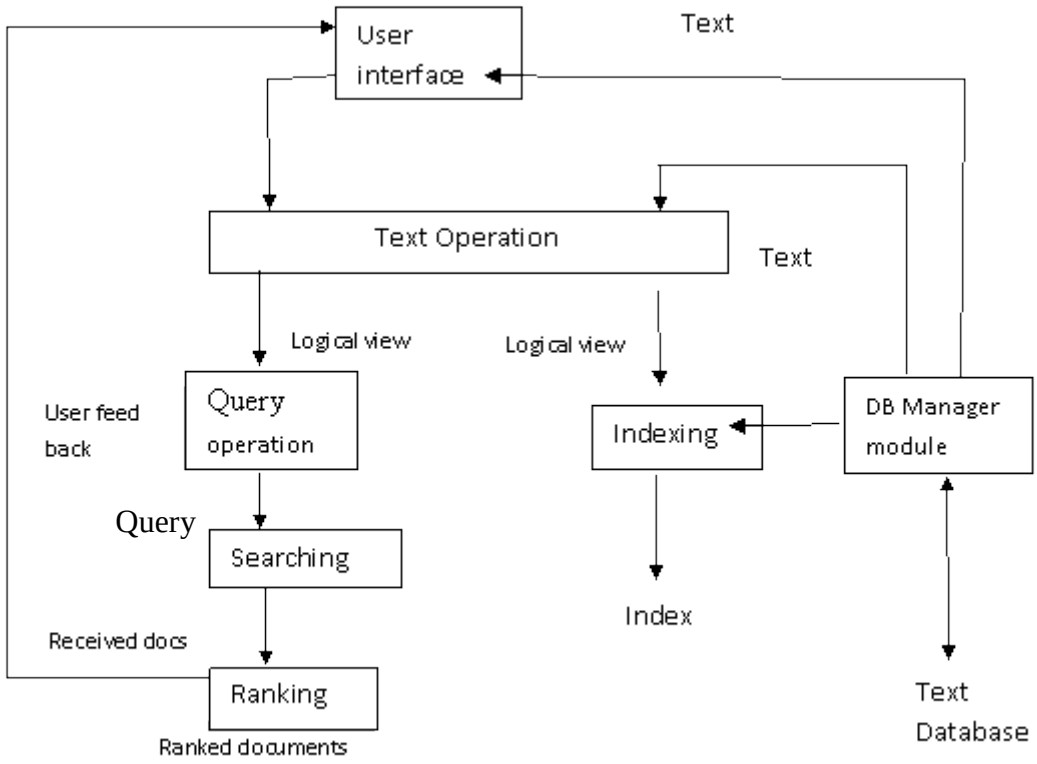

Figure 2 Logical view of information retrieval cycle

ANALYSIS OF THE RANKING ALGORITHM OF EXISTING SYSTEM

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

PAGERANK ALGORITHM

The original PageRank 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 PageRank of page A,
PR(Ti) is the PageRank of pages Ti which link to page A,
C(Ti) is the number of outbound links on 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 sites as a whole, but is determined for each page individually. Further, the PageRank 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 page A does not influence the PageRank of page A uniformly. Within the PageRank algorithm, the PageRank of a 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. .(Lawrence and sergey, 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 the 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 PageRank benefit for a page by another page 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. They consider PageRank as a model of user behaviour, where a surfer clicks on links at random with no regard towards content. The random surfer visits a web page with a certain probability which derives from the page'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 the 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 page at random. .(Lawrence and sergey, 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 is the underlying principle behind the ability of some words to be more relevant in a document than others. Also the work done by Gianni Amati and Van Rijsbergen 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 Poisson distribution model used by both of them showed that the smaller this probability is, the less its tokens are distributed in conformity with the model of randomness and the higher the informative content of the term. Hence, determining the informative content of a term can be seen as an inverse test of randomness of the 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 important 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 hits from the multiple 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-prox-weight. The counts are converted into count-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 and matrices 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 behaviour of software dedicated to the retrieval of sensitive contents in the web, to statistical evaluation of the information content of document copus. This paradigm shift is necessitated by the increase in the actions of publishers with sinister motive on the web. The actions of these publishers have led to a decline in the quality of information retrieved from 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, aspects of which overlap with an area known as bibliometrics. 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-links (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 intent of artificially boosting the latter’s tally of in-links. This 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 platform 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 words in the interpretation of what is wholly contained in a document with the strength of the web link structure to improve the quality of search results through the increase relevance of retrieved documents 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 spammers 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 sustain retrieval of relevant information from digital libraries which in turn determine the quality of learning made possible when such retrieved materials are consulted.
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 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. Addison 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 Alexandria Digital Earth ProtoType – ADEPT. Retrieved 14th of April, 2010 from http://findarticles.com /p/articles/mi m1387/is_2_49/ai_72274394/pg_3/?tag=content;col1.

Ethan T.(2007) Nine people, placesand things that their names:http://blogs.static.mentafloss com/blogs/archives/22707.html

Harter, P. S.(1975)A probabilistic approach to automatic keyword indexing. Part I. On Distribution of Specialty Words in a Technical Literature. Wiley Periodicals, Inc., A Wiley Company

Lawrence P. and Sergey B.,(1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine.. USA. Stanford press.

Maurice D. K. (2010). The size of the World Wide Web. Retrieved 14 April 2010 from http://www. worldwidewebsize.com/.

Michael K, (1996). "Chiming in on Yahoo's roar," Mediaweek, 6(3):9-12.

Plachouras V., Ounis .I, and Amati G.,(2005) The Static Absorbing Model for the Web. Journal of Web Engineering, 4(2):165,186.