EVALUATION OF FULL TEXT SEARCH RETRIEVAL SYSTEM

Kehinde Aruleba

8 PUBLICATIONS **8** CITATIONS

Peter Oriogun
Lead City University

50 PUBLICATIONS **140** CITATIONS

nigeria computer society
www.ncs.org.ng

12th International Conference

EVALUATION OF FULL TEXT SEARCH RETRIEVAL SYSTEM

* K.D. Aruleba[1], R.D. Aremu[2], P.K. Oriogun[3], K.K. Agbele[4], A.O. Agho[5]

[1,3,4,5]Elizade University, Department of Mathematics and Computer Science,
Ilara-Mokin, Ondo State, Nigeria

[2]University of Ilorin, Faculty of Communication and Information Science, Department of Computer Science, Ilorin, Kwara State, Nigeria

[1] kehinde.aruleba@elizadeuniversity.edu.ng, [2]draremu2006@gmail.com,
[3]peter.oriogun@elizadeuniversity.edu.ng, [4]kehinde.agbele@elizadeuniversity.edu.ng,
[5]adrian.agho@elizadeuniversity.edu.ng,

ABSTRACT

With a number of search engines on the web and each with different indexing and ranking methods and different coverage, finding the one that gives the best results for a query becomes a bit challenging. The main problem however, that existing Search engines have to deal with is how to avoid irrelevant information and to retrieve the relevant ones. This current work presents a new approach for retrieving relevant information on the Web, by adopting breadth-First search algorithm. The implementation result of the retrieval system was analysed using recall and precision model for three departments at Elizade University. By learning from users' behaviour, the approach can return very high quality search results, with a strongly reduced computing load.

Keywords: Full-Text Retrieval System, Evaluation Approaches, Ir, Search Engines, Elizade University

1. INTRODUCTION

Search engines are designed to help users to quickly find useful information on the web (Takakuwa, 2000). With a number of search engines on the web and each with different indexing/ranking methods and different coverage, finding the one that gives the best results for a query becomes a bit challenging. Previous studies shows that the performance of search engines depends on the performance measures used and the application domains. The performance of search engines can be evaluated using various measures such as precision, coverage, response time, recall and interface (Dong and Su, 1997). In this paper, we focus on recall and precision of search engines. The quality of searching for the right information accurately would be the precision value of the search engine. For example if we have Precision = 6 / 10 it implies that out of the 10 retrieved documents only 6 are relevant. Recall is the ability of a retrieval system to obtain all or most of the specifically relevant documents in the collection. For example Recall = 6 / 20 because there are 6 specifically relevant document out of the 20 documents retrieved.

2. LITERATURE REVIEW

Recent efforts to create digital libraries have grown exponentially. A survey of the literature (Kreitz, 1996; Kreitz and Orgden, 1990) on digital libraries and initiatives offers definitions of digital library and challenges as well. This article focuses on electronic library resources within three Departments at Elizade University.

The importance of Information Retrieval (IR) keeps growing as the amount of digital information keeps expanding at an ever-increasing rate. Stored documents, photographs and contents of books, and billions of Web pages are useful only if they can be found when needed. Web search engines are the most common way to find such information. They are attracting more than 170 billion queries each month (Bonfils and Yandex, 2013). The field of IR also covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. It is similar to arranging books on a bookshelf according to their topic. Given a set of topics, standing information needs, or other categories (such as suitability of texts for different age groups), classification is the task of deciding which category, if any, each of a set of documents belongs to.

IR systems must have at least three different processes which are, representing the content of documents, representing a user's information need and comparing the two representations (Hiemstra, 2001). IR process begins when a user inputs a query into the retrieval system. Queries are formal statements (in declarative a formal language) of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. An object is an entity that is represented by information in the system database. User queries are matched against the database information.

Full text retrieval systems (FTRS) have become a popular way of providing support for text databases. In a full-text search, a search engine examines all of the words in every stored document as it tries to match search criteria for example (text specified by a user). The main components of a typical search engine according to (Brin and Lawrence, 1998) are: Web Crawler, Indexing and Ranking. Web Crawler according to (Sherman, 2002) are programs which traverse through the Web searching for the relevant information using algorithms that narrow down the search by finding out the most closer and relevant information. Indexing collects, parses, and stores data to facilitate fast and accurate IR. The main purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Ranking is the medium a search engine use to determine which pages are more important than the others, and present them to individual users in order of relevance. The most famous one is the Page Rank Algorithm published by Google founders (Pavalam et al., 2012)

3. METHODOLOGY

There are various search methods to traverse (visit all the nodes) of a graph systematically. A couple of these methods give us some information about graph structure (e.g. connectedness). The key idea behind graph traversal is to mark each vertex when we first visit it and keep track of what we have not yet completely explored. We describe some of the mechanics of these traversal algorithms here. Depth-First Search (DFS) is an algorithm for traversing a finite graph. DFS visits the child nodes before visiting the sibling nodes; that is, it traverses the depth of any particular path before exploring its breadth. A stack is generally used when implementing the algorithm. Breadth-First Search (BFS) uses a queue data structure and it is level by level traversal. Breadth First Search expands nodes in order of their distance from the root. It is a path finding algorithm that is capable of always finding a unique solution, if one exists.

3.1 IR Evaluation Approaches

According to (Agbele, 2014) retrieval effectiveness can be quantitatively measured in a number of ways using a well-known metrics in the IR community to enhance retrieval effectiveness. The most frequently and important basic measures for IR evaluation are precision and recall which are both used in this present study.

3.1.1. Precision

After a search, the user is sometimes able to retrieve relevant information and sometimes able to retrieve irrelevant information. The quality of searching the right information accurately would be the precision value of the search engine.

$$\text{Precision} = \frac{\text{relevant items retrieved}}{\text{total retrieved items}} = \frac{RIR}{TRI}$$

3.1.2. Recall

Recall is the ability of a retrieval system to obtain all or most of the relevant documents in the collection. Also, recall is the fraction of relevant items that are retrieved to relevant items in the database or the probability given that an item is relevant to the retrieved. For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.

$$\text{Recall} = \frac{\text{relevant items retrieved}}{\text{relevant items}} = \frac{RIR}{RI}$$

3.1.3. 11- Point Average Precision

11-point average precision is a measure for representing performance with a single value. In 11-point average precision, we are looking at 11 recall levels (0.0, 0.1, 0.2,... 1.0) and finding the precision at each point. We average these scores across all of the different issued queries from the participants or information needs to validate the retrieval effectiveness of developed system.

3.2. PROPOSED SYSTEM

Figure 1 depicted the architectural design for a Full Text Retrieval System (FTRS) proposed (Aruleba, 2015). This architectural design was implemented using Breadth First Search. The proposed system makes use of a Crawler to gather information from every document on the website and store this information in the index. The index is a structured system of storing the unstructured data returned by the Crawler.

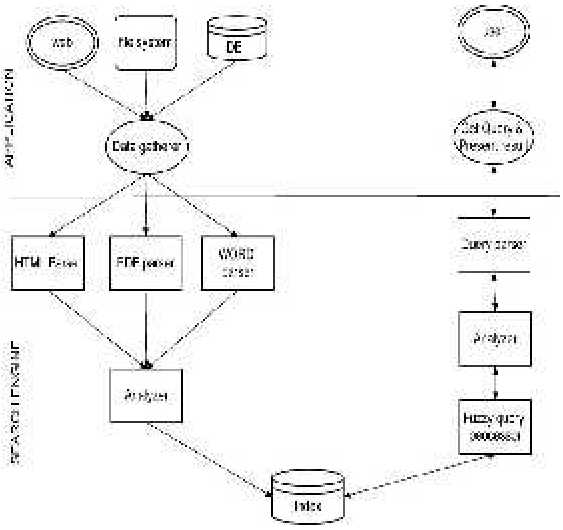

Figure 1: Proposed Architecture for FTRS

4. SYSTEM PERFORMANCE EVALUATION

This section presents the evaluation of the adopted document retrieval search algorithm. The aim of the section is to measure the effectiveness of the retrieval system. In order to test the effectiveness of the full text search system, three departments from Elizade University, that is Departments of Mathematics and Computer Science, Civil Engineering and English were considered. Recall and precision were the two performance parameters used for evaluating the search system. Department of Mathematics & Computer Science, Civil Engineering were chosen to test the computer skills of the users in query formulation and the use of keywords, also English was selected to see how users can construct sentence using keywords and to see how relevant the results of the retrieval system is. The various departments and the total number of participants used in the evaluation is as shown in Table 1.

Table1: Departments and Participants

Department	Participants
Mathematics & Computer	15
Civil Engineering	15
English	15

Total	45

The evaluation had no fixed queries. The Users were asked to perform their daily book searches as usual, based on their daily information needs without any change. The only requirement was that they needed to focus mainly on using search terms related to their departments. The system was developed and implemented with PHP; before the system can be used some requirement (such as software and hardware requirement) must be met. Figure 2 depicts the sample snapshot search system.

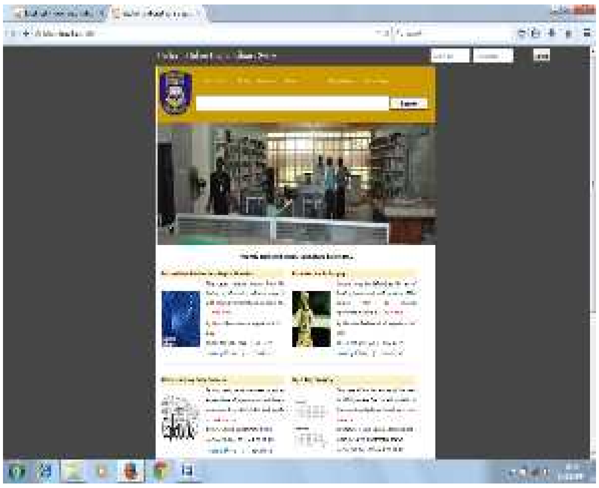

Figure 2: Sample Search Screen

5.1 RESULTS

During the evaluation, users were asked to rate their overall satisfaction with the search engine based on the retrieved results in facilitating their academic work. The results shows that the users in Mathematics & Computer science are more satisfied with the performance of search engines, while the opinions of the users in Civil Engineering and English appeared to be similar to one another as shown in Figure 3.

5.2 DISCUSSION

According to Figures 3, it was observed that there is usually a trade-off between recall and precision i.e. at a high recall value, more documents containing a lot of junks was retrieved by the system and hereby reducing precision while at a high precision value, less but the most relevant documents were retrieved and thereby providing a low recall value. Another observation is that the system seems to perform well on one query than it does to another. This has to do with the query formulation skills of the individual user and how much knowledge a user had about the system content as illustrated in Figure 4.

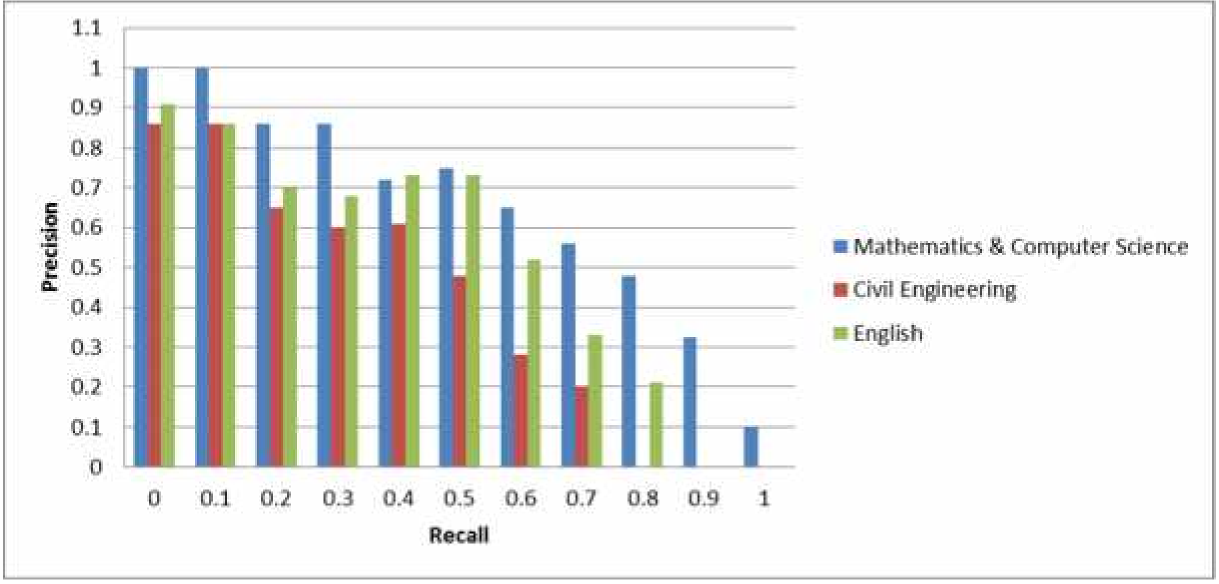

Figure 3: Average 11-point r-p curve across 10 queries using Department of Mathematics and Computer Science, Civil Engineering, English.

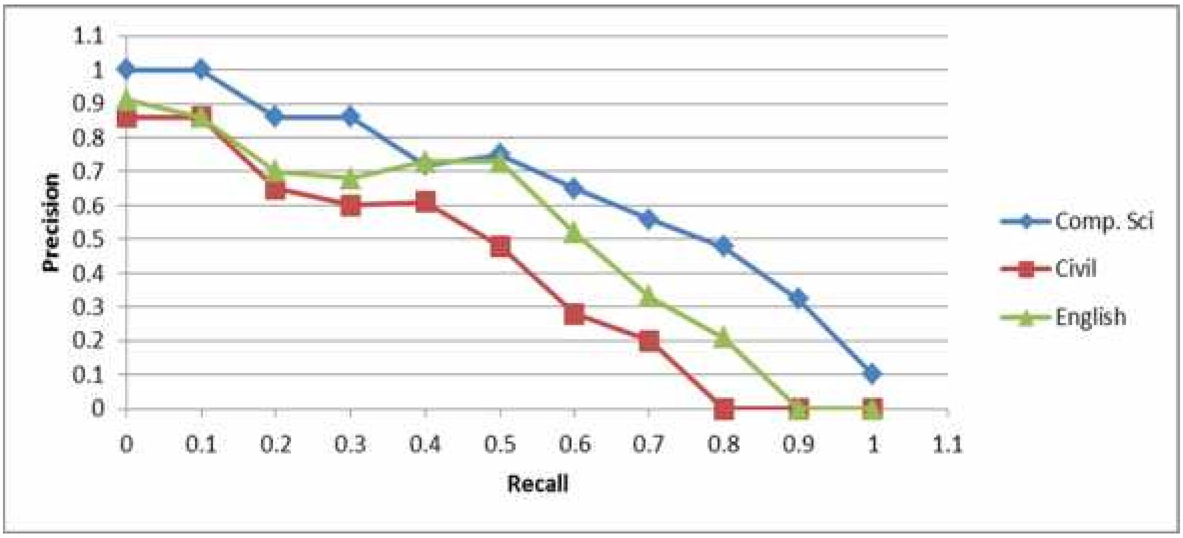

Figure 4: Comparison of 11-point average of mathematics & computer science, civil engineering and English. (NB: The curve closest to the upper right-hand corner of the graph indicates the best performance)

6. CONCLUSION & FUTURE WORK

In conclusion, the analysis of results of the implemented information retrieval system shows that the users of the system find it very effective to use. The implemented information retrieval system enables users to have access to latest learning facilities such as, articles, journals, textbooks, thesis, projects, newspapers, etc. without going through the rigorous steps and routine in the conventional institution libraries. The field of information retrieval is a very interesting research area where improvements can always be made no matter how sophisticated your retrieval application looks. For the future, it remains to be seen whether novel algorithms which may use hybrid techniques and may outperform BFS and DFS individually.

7. REFERENCES

Takkakaw T., (2000) "*Search engines worldwide*" http://www.twics.com/takakuwa/search/search.html (Accessed 15th May 2015)

Dong X., Su L. T. (1997), "*Search engines on the World Wide Web and information retrieval from the Internet: A review and evaluation*." Online and CDROM Review 21: pp. 67-81

Kraits, P.A. (1996) "*Role of the Library in the Scholar's Electronic Research Environment, Digital Libraries and Information Services for the 21st Century*". Seoul, Korea: ICPR.

Kreitz, P.A., and Ogden, A. (1990). "Job *responsibilities and job satisfaction at the University of California Libraries, College & Research Libraries*", 51, 4, 297.

Bonfils M, Yandex (2013), "*Just passed Bing to become 4th largest global search engine*" 2013. URL http://searchenginewatch.com/article/2242374/ (Accessed 20th May 2015)

Hiemstra D., (2001). "*Using language models for information retrieval*." Ph.D. Thesis, Centre for Telematics and Information Technology, 2001.

Page L and Brin S. (1998). "*The Anatomy of a Large-Scale Hypertextual Web Search Engine*", Proceedings of the *Seventh International Conference on World Wide Web 7*, Brisbane, Australia, 30(1-7): 107 – 117, 1998.

Sherman C. (2002). "*Anatomy of a Search Engine: Inside Google*." http://www.searchenginewatch.com/searchday/article.php/2161091. (Accessed 8th July 2015)

Pavalam M., Jawahar M., Felix K., Akorli, S V Kashmir R. (2012). "*Web Crawler in Mobile Systems*" *International Conference on Machine Learning* (ICMLC 2011), Vol. 2, No. 4, pp 531- 534.

Agbele K.K., (2014). "*Context-Awareness for Adaptive Information Retrieval Systems*" Unpublished PhD thesis University of Western Cape, South Africa, 2014.

Aruleba K.D, (2015). "*A full text retrieval system in a digital library environment*" Unpublished MSc dissertation, University of Ilorin, 2015.