Intelligent Information Management, 2016, 8, 1-8
Published Online January 2016 in SciRes. http://www.scirp.org/journal/iim
http://dx.doi.org/10.4236/iim.2016.81001

A Full Text Retrieval System in a Digital Library Environment

Kehinde Daniel Aruleba[1], Dipo Theophilus Akomolafe[2*], Babajide Afeni[3]

[1]Department of Mathematics & Computer Science, Elizade University, Ilara-Mokin, Nigeria
[2]Department of Mathematical Sciences, Ondo State University of Science and Technology, Okitipupa, Nigeria
[3]Department of Computer Science, Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria
Email: arulebakehinde@yahoo.com, *dtakomolafe@yahoo.com, babajideafeni@gmail.com

Received 11 November 2015; accepted 10 January 2016; published 13 January 2016

Abstract

The volume of information being created, generated and stored is huge. Without adequate knowledge of Information Retrieval (IR) methods, the retrieval process for information would be cumbersome and frustrating. Studies have further revealed that IR methods are essential in information centres (for example, Digital Library environment) for storage and retrieval of information. Therefore, with more than one billion people accessing the Internet, and millions of queries being issued on a daily basis, modern Web search engines are facing a problem of daunting scale. The main problem associated with the existing search engines is how to avoid irrelevant information retrieval and to retrieve the relevant ones. In this study, the existing system of library retrieval was studied. Problems associated with them were analyzed in order to address this problem. The concept of existing information retrieval models was studied, and the knowledge gained was used to design a digital library information retrieval system. It was successfully implemented using a real life data. The need for a continuous evaluation of the IR methods for effective and efficient full text retrieval system was recommended.

Keywords

Full Text, Information Retrieval, Library, Digital Library, Queries, Indexing, Catalogue

1. Introduction

For Centuries, libraries have been organizing reading materials on shelves for easy access. However, systematic methods that had been widely adopted for the organization of library materials and their recordings for use by

*Corresponding author.

How to cite this paper: Aruleba, K.D., Akomolafe, D.T. and Afeni, B. (2016) A Full Text Retrieval System in a Digital Library Environment. *Intelligent Information Management*, **8**, 1-8. http://dx.doi.org/10.4236/iim.2016.81001

readers came into being a little more than a century ago [1]. The term digital library is used to refer to a library where some or all of the holdings are available in electronic form, and the services of the library are also made available electronically-frequently over the Internet so that users can access them remotely [2]. The primary purpose of digital libraries is to enable searching of electronic collections distributed across networks, rather than merely creating electronic repositories from digitized physical materials.

An information retrieval (IR) system is designed to retrieve any documents or information required by the user community. It is primarily targeted to make the right information available to the right user at right time. IR is concerned with representing, searching, and manipulating large collections of electronic text data. IR is a discipline that deals with retrieval of unstructured data or partially structured data, especially textual documents, in response to a set of query or topic statement(s), which may itself be unstructured [3]. IR system does not inform *i.e.* change the knowledge of the user on the subject of his enquiry; it merely informs the user of the existence or non-existence and whereabouts of documents relating to the request.

Many problems are associated with the current system of IR and such can be seen from the inability of the system to process request timely and to present inadequate results among others. In view of these inadequacies, it is imperative to develop an IR system that will curtail these inadequacies.

2. Related Work

The importance of 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 easily found when needed.

2.1. Information Retrieval Models

For effectively retrieving relevant documents by IR strategies, the documents are typically transformed into a suitable representation. Each retrieval strategy incorporates a specific model for its document representation purposes. According to [4], the Boolean model is the first model of IR and probably also the most criticized model. Larson [5] shows that much of this criticism seems to be based on lack of knowledge about how to utilise its search possibilities. In this model, we can pose any query which is in the form of a Boolean expression of terms, that is, in which terms are combined with the operators AND, OR, and NOT. The model views each document as just a set of words.

The vector space model (VSM) represents documents and queries as vectors in multidimensional space, whose dimensions are the terms used to build an index to represent the documents. It is used in IR, indexing and relevancy rankings and can be used in evaluation of Web search engines. According to Shang [6], the VSM procedure was divided into three stages. The first stage is the document indexing where content bearing terms are extracted from the document text. The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure.

According to Gonzalez [7], Language models (LM) for information retrieval are retrieval models (taken from the speech recognition field) that do not impose an explicit parametric form for the probability of relevance. Lafferty and Zhai [8] presented a formal connection between probabilistic and language models. The basic idea of the language modelling approach to IR is to assume that a query Q is generated by a probabilistic model of document D. In this context, the generative language models approach estimate £¡ is the probability of the query being generated by a document.

2.2. Query Types

There are many different ways of searching for information. Here we describe the most common ones according to Salerma [9].

A normal query is any query that is not explicitly indicated by the user to be a specialized query. For queries containing only a single term, the desired semantics are clear: match all documents that contain the term. For multi-word queries, however, the desired semantics are not so clear. Some implementations treat it as an implicit Boolean query by inserting hidden AND operators between each search term.

Phrase queries are used to find documents that contain the given words in the given order. Usually phrase search is indicated by surrounding the sentence fragment in quotes in the query string. They are most useful for

finding documents with common words used in a very specific way [9]. For example, if you do not remember the author of a paper, searching for it on the Internet as a phrase query will in all likelihood find it for you.

Boolean queries are queries where the search terms are connected to each other using the various operators available in Boolean logic, most common ones are AND, OR and NOT [10]. Usually parentheses can be used to group search terms. A simple example is software AND database, and a more complex one is software AND database AND data structure.

3. Information Retrieval and Digital Libraries

Libraries have been in existence since the beginning of writing and have served as a repository of the intellectual wealth of society. As such, libraries have always been concerned with storing and retrieving information in the media it is created on. As the quantities of information grew exponentially, libraries were forced to make maximum use of IR methods to facilitate the storage and retrieval process. Some of the IR methods used in digital libraries are described in the following:

Indexing: IR systems need an indexing mechanism for performing efficiently the retrieval process [7]. Indexing is the transformation from the received item to the searchable data structure. Building an index from a document collection involves several steps, from gathering and identifying the actual documents to generating the final data structures [11].

Catalogue records: are short records that provide summary information about a library object. The word catalogue is applied to records that have a consistent structure, organized according to systematic rules. Library catalogues serve many functions, not only information retrieval. Some catalogues provide comprehensive bibliographic information that cannot be derived directly from the objects. This includes information about authors or the provenance of museum artefacts [12]. Descriptive Metadata: Many methods of information discovery do not search the actual objects in the collections, but work from descriptive metadata about the objects [12]. The metadata typically consists of a catalogue or indexing record, or an abstract, one record for each object. Usually it is stored separately from the objects that it describes, but sometimes it is embedded in the objects. Descriptive metadata is usually expressed as text, but can be used to describe information that is in formats other than text, such as images, sound recording, maps, computer programs, and other non-text materials, as well as for textual documents.

Library Digitization

According to Ian and David [13], defined digitization as the process of taking traditional library materials that are in form of books and converting them to the electronic form where they can be stored and manipulated by a computer.

According to Alhaji [14], there are three main reasons for digitization of a library system

1) To make the documents more accessible: This is to serve existing library users better, i.e. to allow users search the full text of documents or to allow users search from remote locations.

2) To preserve the documents: Allow user read older or unique documents without damage to the originals

3) To reuse the documents: Allowing conversion of documents into different formats.

4. Methodology

From the architecture of a FTRS described in Aruleba *et al.* [15], a full text search retrieval system was designed. This section presents the modelling of the system. The modelling is in two parts, which are: Analysis and Design

The analysis of the existing information system in University of Ilorin library was extensively carried out by studying the existing environment. The result of the analysis is presented using the Use-case diagram shown in **Figure 1**.

The result of the existing library information shows that the entire system is made of seven steps. The first step is where the potential library user is registered. The registration allows the user to become a registered and legally authorised user of the library. After the registration, the user is allowed to use the facilities provided by the library. After registration, the registered user is allowed to undertake the remaining steps that is registered user can check available books, read books, check document title, author, publisher, borrow book, return book and pay fine in case of late submission of book(s).

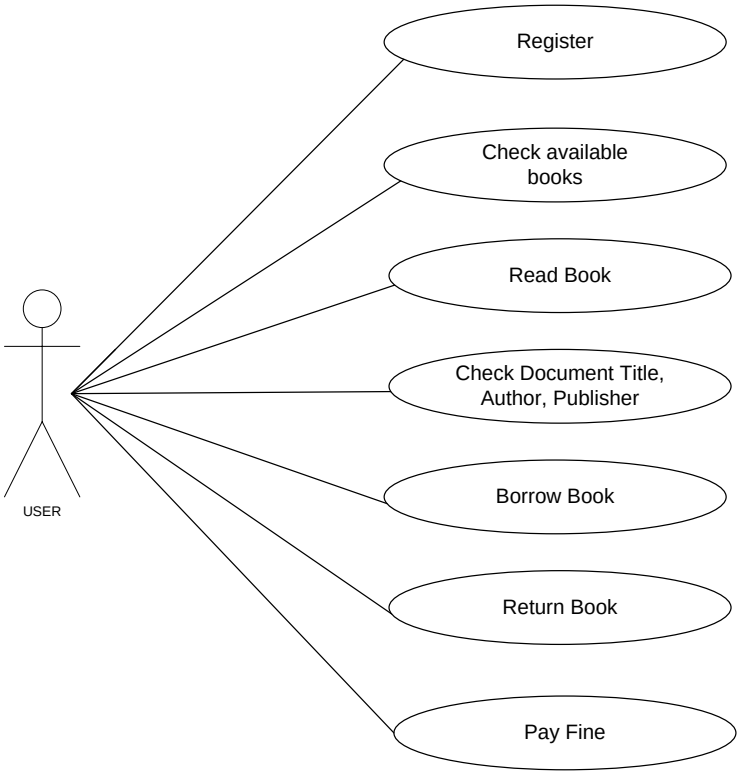

Figure 1. The result of analysis of UNILORIN library system.

The proposed system in addition to the functionality of the existing system allows users to search, modify user details, and upload documents as shown with use-case in **Figure 2**.

From **Figure 2**, the proposed system is made up of eight distinct steps. Though the components are interwoven, each of them performs distinct functions but all work together as a system to process request timely.

4.1. Database Design

Database design mainly includes requirement analysis, concept structure design stage, the logic structure design stage, physical structure design stage, database implementation stage, database operation and maintenance stage, there are six steps altogether.

From the analysis done, **Table 1** was designed for the implementation of the proposed system.

4.2. User Interface Design

There are many factors that must be considered when designing the user interface of a software because the user must be able to interact with the system in a way that the system will understand whatever input given by the user. Therefore, the quality of the interface and software in general must pass the usability testing standard. Some usability factors, such as fit for use, ease of learning, task efficiency, ease to remember, subjective satisfaction and understand ability but all are put into consideration when designing the user interface (**Figure 3**).

The home page screen depicted in **Figure 4**, contains four major modules which are the Search, Registration, Request and Login while the Admin module home page shown in **Figure 5**, contains Sub-module which are view students, view staff, view books, create new book, view book request, create/view facilities, create/view department, logout. Each of them will lead you to its database when clicked and manipulated.

5. System Implementation Phase and Testing

This phase implements what have been discussed in the Section 4. The system was developed and implemented with PHP and MySql Technology.

Figure 2. Use case diagram showing the proposed system.

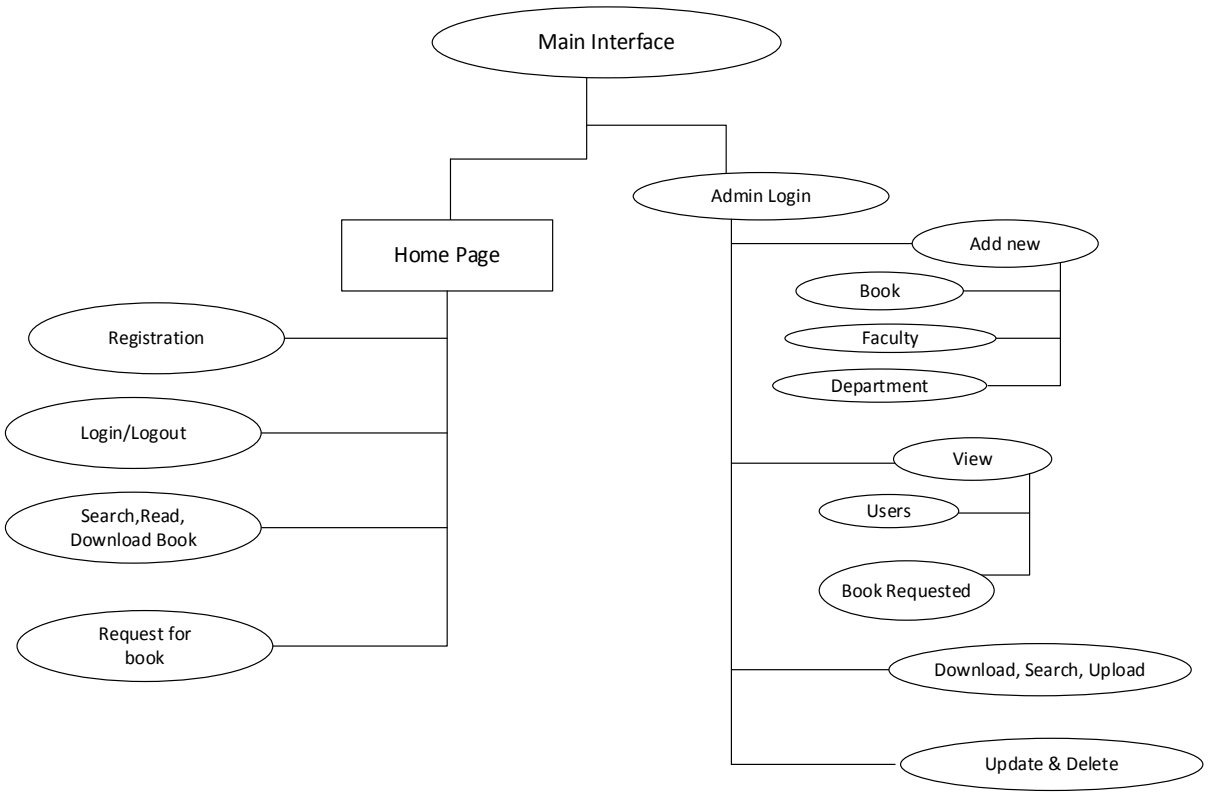

Figure 3. Showing the main interfaces of the system.

Table 1. Generated database.

Table	Action
Admin	To authenticate the system user
Book request	To store information about requested books by the system users
Books	To store book details
Department	To store all the departments available
Faculty	To store all the faculties available
Staff	To store information of staffs using the system
Students	To store information of students using the system.

Figure 4. Home page design interface.

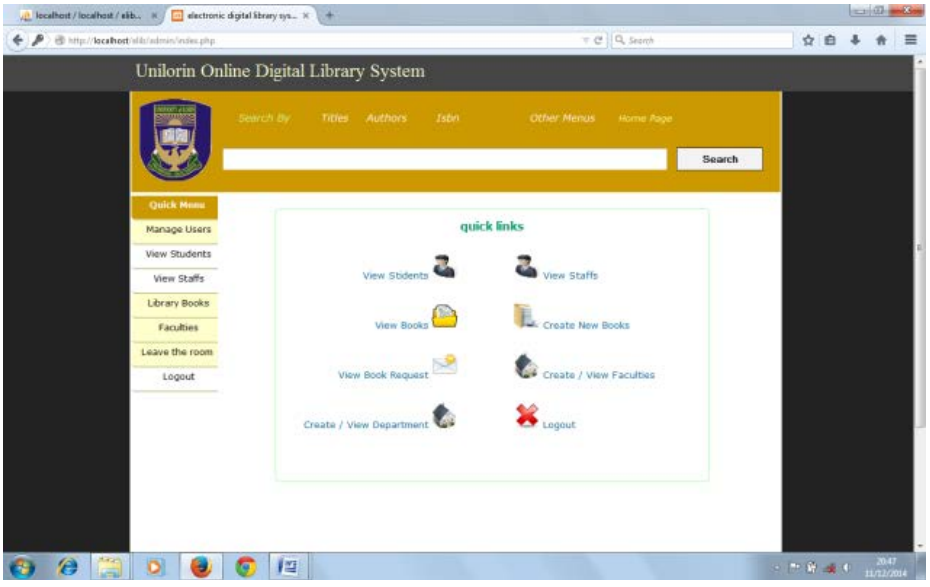

Figure 5. Admin home page design interface.

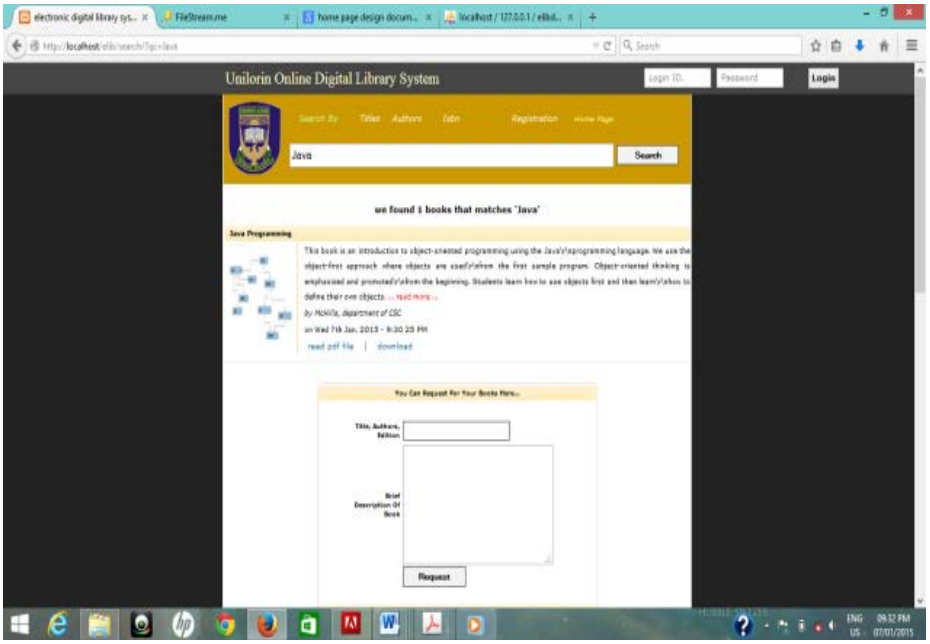

Figure 6. Home page implementation output.

Home Page Interface Implementation

The home page shown in **Figure 6** is the key aspect of the system, because it gives the basic user interface for the full text retrieval digital library. It comprises of: Search, Login, Registration and Request described as follows:

Search: This feature can be used by any user. This module provides a convenient book searching function, the user could search books based on a variety of conditions.

Login: Every user who wants to use the system is authenticated by means of username and password. All entered parameters of the password are matched with information stored in the database, therefore only authenticated users can log on to the program with limited access.

If the login information is wrong, the user will be notified of login failure and would need to try again.

Registration: This involves registering new users. It contains registration form interface with entries like email address, last name, first name, password, password confirmation and sex.

Request: If a user can find the specific book needed, request can be made for such book.

6. Conclusions and Future Work

This study has been able to develop and successfully implement an IR system that reduces the hurdles associated with present system of searching in Libraries. It is shown in the system that users searching full text are more likely to find relevant articles than searching only abstracts. This finding affirms the value of full text collections for text retrieval and provides a platform for aligning searching algorithms that take advantage of rapidly-growing digital archives.

Also, the following areas of the study can be improved upon in future studies to create a more robust IR:

- Increase in the size of the database, this will enable large data storage.
- Integrate advert plans for research materials the institution wants to be selling online.
- Acquire and publish video, audio and heavy graphic research materials.

References

[1] Onwuchekwa, E.O. and Jegede, O.R. (2011) Information Retrieval Methods in Libraries and Information Centers. An *International Multidisciplinary Journal*, **5**, Serial No. 23

[2] Rosenberg, D. (2005) Towards the Digital Library: Findings of an Investigation to Establish the Current Status of University Libraries in Africa. International Network for the Availability of Scientific Publications, Oxford.

[3] Greengrass, E. (2002) Information Retrieval: A Survey by Ed Greengrass. Information Retrieval, 141-163.

[4] Göker, A. and Davies, J. (2009) Information Retrieval: Searching in the 21st Century. Wiley, West Sussex. http://dx.doi.org/10.1002/9780470033647

[5] Larson, R.R. (2010) Information Retrieval Systems. In: Bates, M.J. and Maack, M.N., Eds., *Encyclopedia of Library and Information Sciences*, 3rd Edition, CRC Press, New York, IV, 2553-2563.

[6] Li, L.Z. and Shang, Y. (2000) A New Statistical Method for Performance Evaluation of Search Engines. ICTAI.

[7] González, R.B. (2008) Index Compression for Information Retrieval Systems. Unpublished PhD Thesis, University of A Coruña, A Coruña.

[8] Lafferty, J. and Zhai, C. (2003) Probabilistic Relevance Models Based on Document and Query Generation. In: Croft, W.B. and Lafferty, J., Eds., *Language Modelling for Information Retrieval*, Kluwer, Pittsburgh, 11-56. http://dx.doi.org/10.1007/978-94-017-0171-6_1

[9] Salerma, O. (2006) Design of a Full Text Search Index for a Database Management System. MSc Dissertation, University of Helsinki, Helsinki.

[10] Pazer, J.W. (2013) The Importance of the Boolean Search Query in Social Media Monitoring Tools. Dragon Search.

[11] Manning, C.D., Raghavan, P. and Schutze, H. (2008) Introduction to Information Retrieval. Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9780511809071

[12] Arms, W. (2002) Manuscript of Digital Libraries. MIT Press, Cambridge.

[13] Written, L.H. and Brainbridge, D. (2003) How to Build a Digital Library. Morgan Kaufman Publishers, London.

[14] Alhaji, I. (2005) Digitization of Library Resources and the Formation of Digital Libraries: A Practical Approach. University of Pretoria, Pretoria.

[15] Aruleba, K.D., Aremu, D.R., Oriogun, P.K., Agbele, K.K. and Agho, A.O. (2015) Evaluation of Full Text Search Retrieval System. *Nigeria Computer Society*, **26**, 154-159.