Using Lucene/Solr in E-Learning to Implement Conceptual Extraction and Integration of Multimedia Documents

Titilayo Azeez, Charles Robert

Using Lucene/Solr in E-Learning to Implement Conceptual Extraction and Integration of Multimedia Documents

Titilayo Azeez and Charles Robert
Department of Computer Science
University of Ibadan,
Oyo Road, Ibadan , Nigeria
email: abc.robert@live.com

Abstract
The purpose of this work was to provide relevant integrated multimedia documents to learners. It was situated in an E-Learning environment. Concept extraction methodology based on Lucene and Solr was reviewed and adapted to learning system. Retrieval of relevant materials from a domain was implemented after relevant information was organized and related. Lucene was a key concept that helped us to relate information for providing the relevancy of lessons to the learner. It generated a learner specific e-Learning content by comparing the concepts with vector space model similarity measures. Based on the proposal, an apache projects combining Lucene and Solr were used, and vector space model and Boolean model were applied, for the searching and indexing of learner's query term from multimedia documents. Vector space model was used in information filtering, information retrieval, indexing and relevancy rankings. Vector space model allowed computing a continuous degree of similarity between queries and documents. The work selected query items which were equally weighted compared with Boolean model. The system on lucene/solr allowed multi-media document(s) to be retrieved and provided learner access to multimedia information. The resulting output to queries confirmed that Boolean model has some difficulties that included the fact that (a) it cannot allow result to be effectively ranked (b) it is difficult to rank output, data retrieval rather than information retrieval. Vector space model retrieved fewer documents compared to too many in Boolean model. Retrieving similar documents and matching all the documents so as to produce relevant document is not possible in Boolean model. The results of the proposed e-Learning system under the design of Apache projects similarity measure show a significant increase in performance and accuracy under different conditions. The assessment of the comparative analysis, showed the difference in performance of our proposed method over other methods.

Keywords; E-Learning; Lucene; Solr; vector space model; metadata; index; term; documents;

I. Introduction

E-learning can also be defined as an instruction delivered by any technological mode intended to promote learning (Clark, 2011). E-learning includes numerous types of media that deliver text, audio, images, animation, and streaming video, and includes technology applications and processes such as audio or videotape, satellite TV, CD-ROM, and computer-based learning, as well as local intranet/extranet and web-based learning.With large volume of data being generated every day, resources for E-learning and research is increasing but retrieving relevant resources amidst these large data are getting more complicated. Learners in an E-learning system might spend too long time searching for what they need. They get totally frustrated in their quest for knowledge when E-learning resources (media) retrieval

system does not function as expected. Although search engines provide search tools for relevant text document, other media types are often neglected or not provided unless the media type is specified by the learner.

II. Related work

The work of (Lau et al, 2009) on "*Using concept similarity in cross ontology for adaptive e-Learning systems"* proposed an adaptive e-Learning system, which generates a user specific e-learning content by comparing the concepts with more than one system using similarity measures. A cross ontology measure is defined, which consists of fuzzy domain ontology as the primary ontology and the domain expert's ontology as the secondary ontology, for the comparison process.

The work of (Lau et al, 2009) on "*Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning".* The main contribution of this paper is the illustration of a novel concept map generation mechanism which is underpinned by a fuzzy domain ontology extraction algorithm. The proposed mechanism can automatically construct concept maps based on the messages posted to online discussion forums. By browsing the concept maps, instructors can quickly identify the progress of their students and adjust the pedagogical sequence on the fly.

The report on "*A Query Language for Similarity-Based Retrieval of Multimedia Data"* by (Amato et al, 1996) presented the main features of a Multimedia Query Language tailored for content-based similarity retrieval of multimedia objects. In their opinion, the abstraction mechanisms of a model for multimedia data should be able to represent the features and concepts which are intrinsically present in every multimedia data item. (G. Amato, G. Mainetto, and P. Savino, September 1996).

The research work aims at extracting metadata from multimedia documents using Lucene/Solr which implemented "vector space model" using a similarity measure thereby improving the retrieval of learning contents. The paper is organized as follows: The second section provides models and expectation. The third section describes the methods utilized for the e-Learning system. The fourth section deals with results and performance evaluation of the proposed approach under different criteria. Finally, fifth section is the conclusion.

III. Model and expectations

The approach used was concept extraction (metadata) from multimedia documents to compare the similarities between supplied multimedia documents and learner's query. The method used was Lucene/Solr Apache project; Documents are extracted from the index created by Lucene which the Solr used to perform the searches. A concept map was automated from the index which was manually developed by experts (administrator). An XML file was generated by the index based on a particular concept, which contained the representative and property set of the specific concepts from the documents. The representatives and properties was then processed by vector space model for the generation of the concept similarity measure and retrieval of integrated multimedia documents was supplied matching learners' query.

The concept similarity measure was the main factor which was done by retrieval of multimedia metadata, it defined the most relevant multimedia documents for the learner according to the learners' query.

The main contributions of this paper are as follows;

- This research will contribute to knowledge by improving retrieval system for multimedia in E-learning environment.
- Retrieval of multimedia documents based on concept extraction which helps to build learners' competence through fast access to information.

IV. Methodology, Analysis and Design

The method used for this research was Lucene/Solr which implemented "Vector Space Model" algorithm that compare similarities between learners query and supplied multimedia document.

Documents were viewed as a sequence of terms and each document has its own associated metadata. Metadata is specific form of data describing the document such as its author(s), titles, date of publication. The metadata contain date of creation, format of document, as well as author and possible the title of the document. The retrieval of multimedia document is done by extraction concept from the multimedia metadata.

Extracted multimedia metadata were compared for similarities with the learner's query using Vector Space Model. Closely related media files are selected based on the result of the vector space model and suggested to the users as a resource. Due to the fact that Vector Space Model can only analyze text documents only and multimedia files are in different format, metadata of these media files are extracted so as to be analyzed by the model.

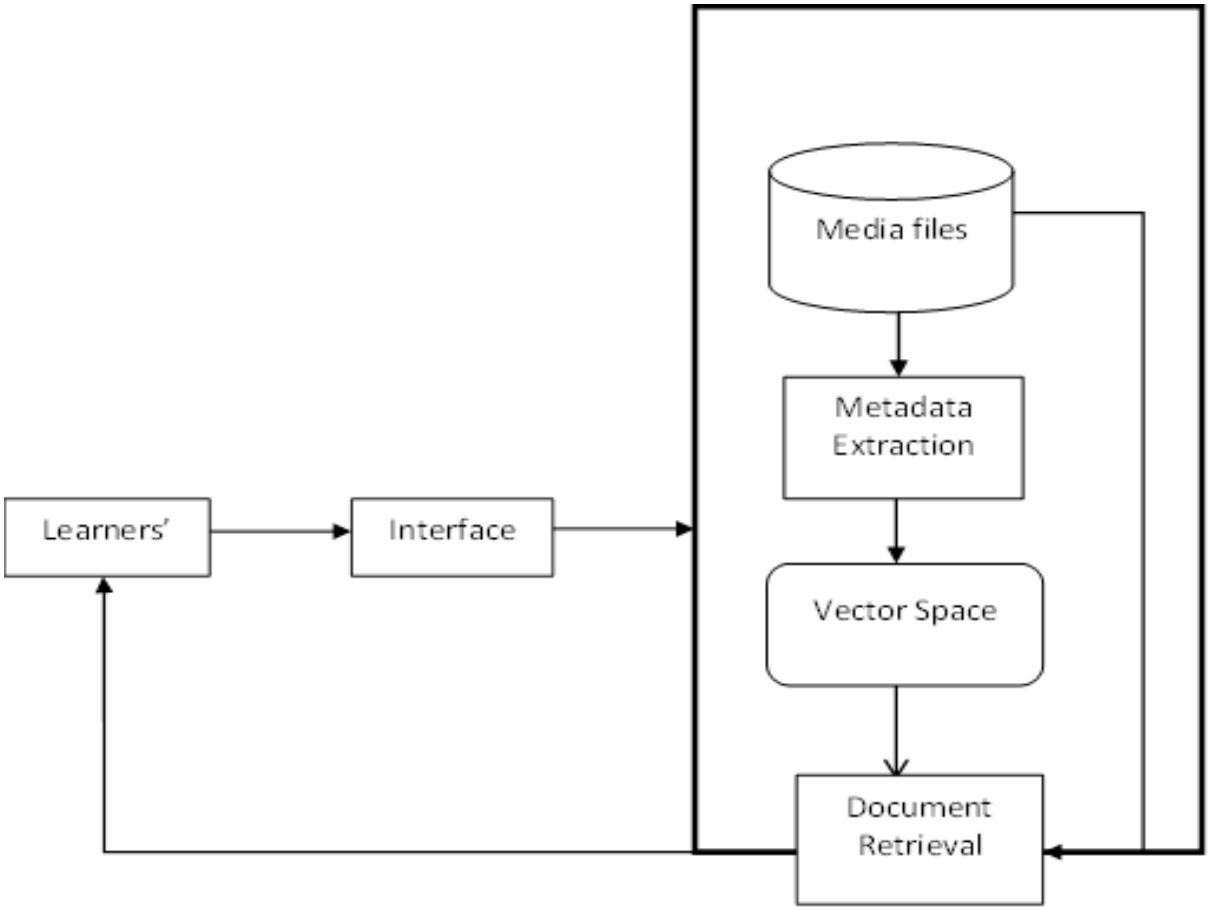

Figure 1*:* Architectural design of the system.

System Analysis entails identifying and defining the problem in hand, carrying out feasibility study of the same and choosing tools to solve the problem. Also, before developing the system a thorough understanding of the framework (Using Lucene/Solr in E-learning system to implement conceptual integration of multimedia documents) was understood.

Most of the software used in development of this system is mainly open source software. It also investigated whether the project being undertaken meets the current technological advances. The project was technically feasible as it used the Lucene and Solr and script programming. The resources required are mainly development tools and a conducive computing and development environment. All the development tools are open source and readily available. Being an academic project, human resource is readily available.

The system design produced the details that state how a system will meet the requirements identified during the system analysis. The database design (Corpus) describes the fields in the documents and then elaborates on the various types of fields in the files. Furthermore, a detailed description of each of the fields follows and the purpose of each field (index). We also identity main fields used in extraction of the documents based on the learner's query identify the document(s).

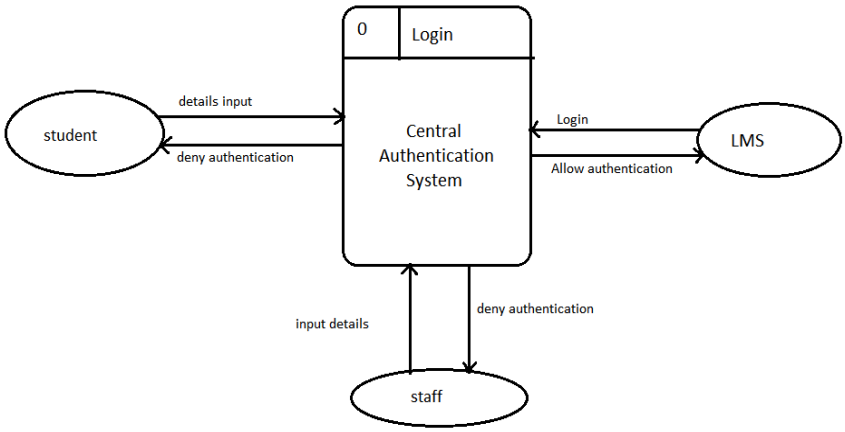

Figure 2: Context diagram for the system.

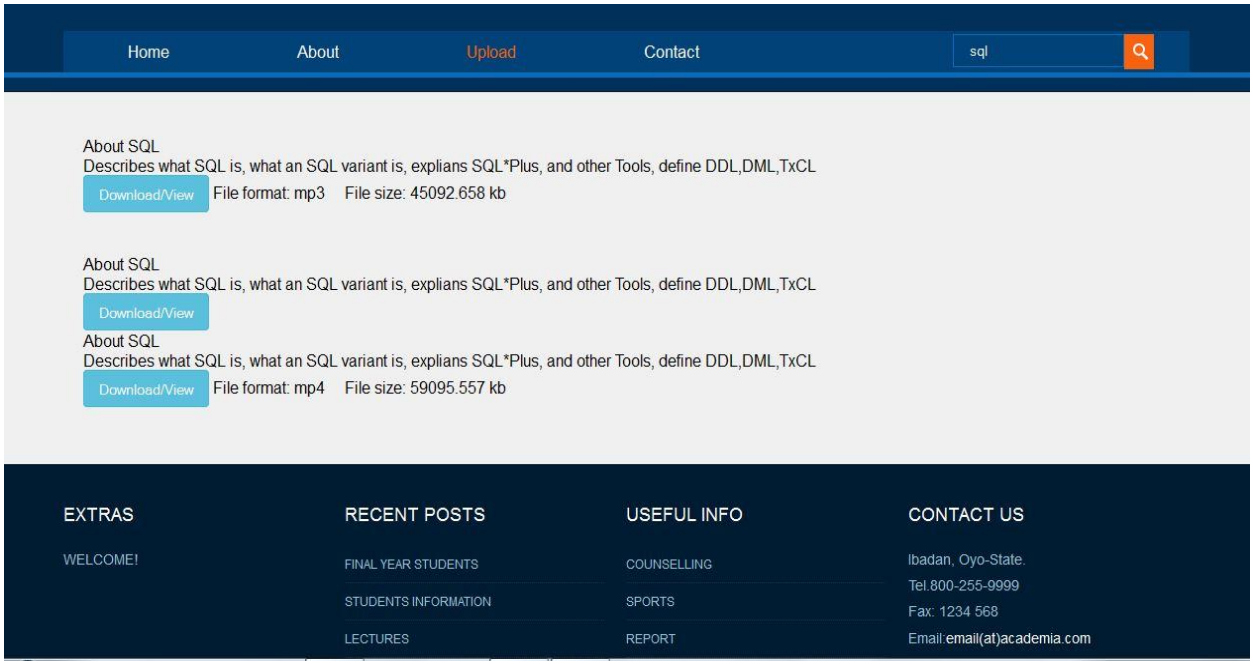

Figure 3: A snapshot of the download page of multimedia documents by learners.

V. Experimental Results and setup

The approach used is subjected to a test with specific documents using Lucene/Solr.Lucene takes all the metadata of documents, splits them into words, and then build an index for each word. Lucene then create a big index which contains word id, the no of documents the word is present, the position of the word in those document. Solr used the index created by Lucene to search.The result is ranked using Lucene / Vector space model algorithms.
This system was implemented using web technologies: PHP, HTML, JAVASCRIPT, LUCENE, and SOLR. Programs were written in PHP, HTML, and JAVASCRIPT for the

execution of LUCENE/SOLR. LUCENE/SOLR performed searches for learners' before multimedia documents were provided. The web technologies were chosen due to their good and advanced capabilities. It was also chosen due to their ease of use and connection to server especially PHP and SOLR.

VI. Conclusion and Future Enhancement

An E-learning system with Apache Lucene and Solr searching and indexing measure has been developed with automatically conceptual integration of multi-media documents as an application to the e-Learning system. The concept was designed with the most innovative and recent techniques. The Apache Lucene and Solr has been incorporated for the method. The major feature in our method was vector space model, which helps in retrieving relevant multi-media documents to learners based on learner's query. Further research can be done on this project so as to make it a world class system i.e. in the areas where it would be useful.

VII. References

1. Bianchi, S., Mastrodonato, C., Vercelli, G., Vivanet, G.. (2009). *Use of ontologies to annotate and retrieve educational contents: the aquaring approach*. International Journal of Computer Science Application Volume 5, No 1, pages 211–220.
2. http://www.cominvent.com/2011/04/solr-architecture-diagram/
3. http://lucene.sourceforge.net/talks/pisa/
4. Lau Raymond, Y.K., Song, Dawei, Li, and Yuefeng, (2009). *Towards a fuzzy domain ontology extraction method for adaptive e-learning*. IEEE Trans. Knowledge Data Engineering
5. (Luhn (1957; 1958)). Computational aspects of vector space scoringearning weights.
6. Salton , B. (1987), Robertson, J. (1976), Croft, H. (1979), Papineni (2001). Extensions and theoretical justifications of idf.
7. (Salton, B. 1988, Singhal et al. 1995; 1996b). The SMART notation for tf-idf term weighting schemes.
8. Sparck Jones (1972). Detailed experiments showing the use of inverse document frequency in term weighting.
9. Zobel, J., Moffat A.,(2006). Inverted files for text search engines. ACM computing Surveys 38(2)