Online Students’ Project Archive with Integration of Document Clustering and Annotation.

Odirichukwu J.C*, Nwokorie E.C. & Asimole I.G.
Department of Computer Science
Federal University of Technology
Owerri, Imo State, Nigeria.

*Chiomajaco6@gmail.com, euphemia.nwokrie@futo.edu.ng, gastonasimole@gmail.com

Daenwe T.L.
Softpointapps
tonye4u@gmail.com

*Correspondence Author.

ABSTRACT

Project work is an indispensible aspect of learning. It impacts knowledge on the student (researcher), supervisor(s), and the entire public which could be used to achieve development goals. Carrying out a project involves an adventure into a task methodologically in order to achieve a mission. Project in most tertiary institutions in Nigeria are not always archived in a proper way. The system does not provide efficient and effective handling of research work by students. Improper documentation of students' research, misplacement of chapters, student's non-implementation of correction made by supervisors, and difficulty in sorting and retrieval of archived projects done by students are some of the problems. In this paper, an online students’ project archiving system with document clustering and annotation is designed and implemented. It provides a web based platform for researchers and their supervisors to interact using internet technologies involving document sharing, short messages service and secured online storage facilities. Object Oriented Analysis and Design Methodology was adopted while internet technology tools like Hypertext mark-up language (HTML), PHP, JavaScript, cascading style sheet,(CSS) and structured query language (SQL) were used. The system clusters each student’s project in the portal for quick and easy access. Implementation was done in www.jacoportal.com.

Keyword: Document Clustering, Annotation, Archive, portal, Internet Technology.

Aims Research Journal Reference Format:

1.0 INTRODUCTION

Projects are undertaken by students as a partial fulfilment of the requirements for the award of degrees (Bachelor, Masters, Doctorate, Post- Doctorate, etc.). Supervisor(s) are usually assigned to each final-year student as early as possible in his or her final year. Students and supervisors communicate through periodic submission of progress reports until the final defence of the project by the student which ends the Project supervision. Copies of students’ projects are kept in the department for future reference and further research. As departments turn out students every year, project heaps increase and hence the need for web based archival system.
The idea here is to produce a web portal to ease the activity and also provide an online archive for past project topics and chapter documentation for future reference. This will enable the students upload their different chapters on the portal and the supervisor download each chapters, reads it annotate it through the portal. After every academic year, one can easily login to the portal and view or research on any topic or past project done by past students. Document clustering algorithm was implemented in the retrieval of chapters. Project chapters are retrieved and clustered within each authors section. The purpose of this paper is to design and implement an online students’ project archive.

This will lead to:
- Effective storage of project topics and chapters.
- Quick referencing and retrieval of project topics.
- Research purpose by either students or staff.

2. REVIEW OF RELATED WORKS

Information retrieval systems use clustering of documents and queries to improve both retrieval efficiency and retrieval effectiveness. Normally, clustering involves grouping together static descriptions of documents by their similarity to each other, though user-based clustering suggest that usage patterns concerning co-relevance can form a basis for clustering. In [1] User-based document clustering was done by re-describing subject description with a genetic algorithm. The article reports that cluster of co-relevant documents obtain increasingly similar description when a genetic algorithm is used to adopt subject description. This makes documents become more effective in matching relevant queries and failing to match non relevant queries. The findings of this work were reached through simulation experiments. A comparative study of generative model-based document clustering was done in [2]. This paper presents a detailed empirical study of 12 generative approaches to text clustering, obtained by applying four types of document-to-cluster assignment strategies (hard, stochastic, soft and deterministic annealing DA based assignments) to each of three base models, namely mixtures of multivariate Bernoulli, multinomial, and von mises-fisher (vMF) distributions.

A large variety of text collections, both with and without feature selection, was used for the study, which yielded several insights including:
- Showing situations wherein the vMF-centric approaches, which are based on directional statistics perform far better than multinomial model-based methods and
- Quantifying the trade-off between increased performance of the soft and DA assignments and their increased computational demands.

Hierarchical Document Clustering is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other cluster, unlike document classification [3]. A survey on optimization approaches to text document clustering was discussed [4]. Text document clustering is one of the fastest growing research areas because of availability of huge amount of information in an electronic form. There are several number of techniques launched for clustering documents in such a way that documents within a cluster have high intra- similarity and low inter-similarity to other clusters. Many document clustering algorithms provide localized search in effectively navigating, summarizing and organizing information. A global optimal solution can be obtained by applying high speed and high quality optimization algorithms. The optimization techniques perform a globalized search in the entire solution space [4]. Word net-based text document clustering can greatly simplify browsing large collections documents by reorganising them into smaller number of manageable clusters as described in [8]. Algorithms to solve this task do exist, however, the algorithm are only as good as the data they work on. Problems include ambiguity and synonym, the former allowing for erroneous grouping and the later causing similarities between documents to go unnoticed.
In this research, naive, syntax-based disambiguation is attempted by assigning each word a part of speech tag and by enriching the bag of words data representation often used for document clustering with synonyms and hypernyms from wordNet [5]. Current clustering research emphasizes the development of more efficient clustering methods without considering the domain knowledge and user’s need. In recent years the semantics of documents have been utilized in document clustering. [6] focused on a clustering model where ontology approach is applied (ontology based model for document clustering). The major challenge is to use the background knowledge in the similarity measure. This article presents an ontology based annotation of documents and clustering system. The semi-automatic document annotation and concept weighting scheme is used to create ontology based knowledge.

The particle swarm optimization (PSO) clustering algorithm can be applied to obtain the clustering solution. The accuracy of clustering has been computed before and after combining ontology with vector space model (vsm). The proposed ontology based framework gives improved performance and better clustering compared to the traditional vector space model [6]. Partitional clustering algorithms have been recognised to be more suitable as opposed to the hierarchical clustering schemes for processing large data sets. A wide variety of distance functions and similarity measures have been used for clustering such as squared Euclidean distance, cosine similarity and relative entropy [7].

An explosive amount of data has been generated in the past two decades and shared by billions of people over the internet, and [8] discusses the history and the current challenges of archiving massive and extremely diverse amounts of user-generated World Wide Web data. This paper also describes the history and the current challenges of archiving massive and extremely diverse amount of user-generated data in an international environment on the World Wide Web and the technologies required for interoperability between service providers and for preserving their content in the future. The field of web archiving exist in a field, fragmented and heterogeneous state. Part of the problem is that this field is relatively new and its literature is scattered across a wide range of journal and conference venues. This makes the state of web archiving as a discipline particularly difficult to ascertain. In [9] an approach to building a collection of web archiving research articles about the subject is presented.

They began with a small data set of articles taken from a web archiving bibliography and then proceeded to expand it by crawling the web and collecting additional documents. The crawled documents are then classified using machine learning classification techniques. They showed that by extracting the documents titles and abstracts and representing them using the bags of words approach, documents can accurately be identified from the web crawler. Several ontology’s have been created in the last years for the semantic annotation of scholarly publications and scientific documents. This rich variety of ontologies makes it difficult for those willing to annotate their documents to know which ones they should select for such activity. A classification and description of these state-of- art ontologies, together with the rationale behind the different approaches with an example of how some of these ontologies can be used for the annotation of a scientific document is shown in [10].

Scholars are increasingly citing electronic web references which are not preserved in libraries or full text archives. There is a new standard for citing web references. This involves archiving the cited web page through www.webcitation.org and citing the website permalink instead of (or in addition to) the unstable live webpage. An archiving system for long-term digital preservation of cited WebPages is described in [11]. The paper discussed the rationale for website, its technology, and how scholars, editors and publishers can benefit from the service. A perspective on archiving the scholarly web on the nature of the archival functions based on a perspective of the developing future scholarly communications infrastructure [12]. Document annotation can be facilitated using content and querying value as a large number of organizations today generate and share textual descriptions to their products, services and actions [15]. Such collections of textual data contain significant amount of structured information, which remains buried in the unstructured text.
While information extraction algorithms facilitate the extraction of structured relations; they are often expensive and inaccurate, especially when operating on top of text that does not contain any instances of the targeted structured information. The approach relies on the idea that humans are more likely to add the necessary metadata during creation time, if prompted by the interfaces; or that it is much easier for human (and algorithms) to identify the metadata when such information actually exists in the document, instead of naively prompting users to fill in forms with information that is not available in the document [15]. Document management system can provide invaluable document searching, versioning, comparison, archiving and collaborations features [16]. For many, logging onto a network or the internet can be like charting the ocean with a row boat. There may be a sea of information at our fingertips, but if we lack the proper vessel to navigate, finding what we need even within our own organisation’s information system can be a significant challenge. [16] describes a better organisation using document management system.

3. METHODOLOGY

The researchers adopted Object Oriented Analysis and Design methodology in carrying out this research. Figure 1 depicts the system architecture for the proposed system.

![Figure 1: Proposed System Architecture](image-url)
3.1 Modeling the functions of the system.

The proposed function was model using use case diagram which is initiated by an actor (administrator) who input the various departments relevant for the usage of the new system and also creates record for new user (lecturers) in other to gain trusted access to the system. Furthermore, authorized users (actors) are given the right to access the module as defined in the security module.

**Figure 2: Use Case Diagram for Proposed System**
3.2 Clustering Techniques
There are two main approaches to document clustering: hierarchical clustering and K-means. The former is thought of as the better approach, but has limitations due to its quadratic time complexity. The latter however, has time complexity which is linear in the number of documents, but produce inferior clusters. Sometimes, agglomerative hierarchical clustering is seen as better than K-means, though slower. However, it has been recently found out that a simple variant of K-means: “bisecting” K-means, could produce better clusters of documents, which are better than those produced by “regular” K-means, and are as good, or better than clusters of documents produced by agglomerative hierarchical clustering.

There are two approaches for creating a hierarchical clustering:
(a.) Agglomerative: - This starts as individual clusters, and joins the most similar pair of clusters at each step.
(b.) Divisive: - This starts with one cluster, and split a cluster at each step, until only clusters of individual points remain.

Unlike the hierarchical clustering techniques, partitional clustering (K means) generates a one-level partitioning of the data points. Supposing that K is the desired number of clusters, then partitional clustering approaches find all K-clusters at once.
In this paper, the researchers adopted K-means clustering algorithm.

4. Analysis of the Proposed System
The proposed system is a portal designed to have a database that will archive project topics and chapters. The researchers adopted document clustering and annotation here such that project chapters submitted by students are clustered according to chapters. Since the process of project supervision involves submission of the same chapters many times until the chapters are declared satisfactory, K-means clustering algorithm was applied such that the same chapters, that is; chapters and its updates are organised in a Last-In-First-Out structure. The annotation by the supervisor (s) would be implemented as an eclipsed of clusters comments with a number indicating the number of annotations made by the supervisor. When pulled down, shows the clustered comments.

When the supervisors assign topics to the students in the portal, each student are required to register and have a password assigned to them. The students and the supervisor need to login it to the portal often, for uploading and downloading of project chapters respectively. The supervisor downloads the documents to a local computer after reading, made corrections or comments and attached as an annotation to the project work. The students are required to always login to the portal to check for their supervisors’ annotations; thereafter makes corrections and uploads it back to the portal for further supervision and approval. Accessible now becomes a reality. This will not eliminate a face to face discussion or interaction between the supervisor and student. This can also create a reference where the student cannot deny that corrections were not made.

4.1 Benefits of the Proposed System
The following are the benefits of the system:
- This new system is designed to reduce the stress which students and supervisors encounter during the course of their project. It will be of good importance to any institution of learning or to any department not only in computer science.
- This design will help new students and their supervisor understand the importance of document clustering whereby as each student uploads their topics to the portal, they are clustered under their own section and not scattered just like in an email box where several mails come in at different times and are not clustered by according to the sender.
This will also help the supervisor have a clearer view of chapters uploaded which clustered under each student name. This design will also provide an avenue for storage and archiving of project topics done by students for easy referencing and research.

4.2 Finding and identifying the objects/class.
The model depicts the proposed system objects/classes within the domain of the system. This shows the different classes that were identified.

![Diagram of object/class model](image-url)

**Figure 3: Finding and identifying the objects/class**
4.3 Program flowchart

Figure 4: Program Flowchart
Table 4.1a Project Chapter table

<table>
<thead>
<tr>
<th>Id</th>
<th>chapters</th>
<th>ptitle</th>
<th>pcomments</th>
<th>suid</th>
<th>Supname</th>
<th>Regno</th>
<th>Student_name</th>
<th>Dc_id</th>
<th>photo</th>
</tr>
</thead>
</table>

This is the database table for storing the chapters and annotation made by supervisors.

Table 4.1b Project Chapter table

<table>
<thead>
<tr>
<th>Id</th>
<th>cid</th>
<th>Ptitle</th>
<th>pcomments</th>
<th>Supid</th>
<th>Regno</th>
</tr>
</thead>
</table>

This is a continuation of the database table which create relationship between students and supervisors.

Table 4.2 Project Student table

<table>
<thead>
<tr>
<th>Id</th>
<th>Regn</th>
<th>student_name</th>
<th>phon</th>
<th>ses</th>
<th>dep</th>
<th>abstrat</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>apr_statu</th>
<th>apr_topic</th>
<th>area</th>
<th>supi</th>
</tr>
</thead>
</table>

This is a database table that stores the students' information.

Table 4.3 Project Chapter log table

<table>
<thead>
<tr>
<th>Id</th>
<th>Chapter</th>
<th>ptitle</th>
<th>pcomments</th>
<th>suid</th>
<th>Supname</th>
<th>Regno</th>
<th>Studentname</th>
</tr>
</thead>
</table>

This is database for the log of table 2.1.1b

5. RESULT DISCUSSION

This section shows the result obtained from the implemented system.

![Figure 5: Successful Registration Output](image)

After the successful registration of any student, a password will be display on the screen.
Figure 6: Student Portal Section
After successful login, the student will have access to the interface in figure 6. For the students to have access to the link for uploading of their chapters, their supervisors must have assigned topic to them.

![Figure 7: The Supervision Section](image)

Figure 7: The Supervision Section
<table>
<thead>
<tr>
<th>Chapter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A computer student must learn to typeset work properly. I believe you are not supervising yourself. You must see me before writing any chapter. See me on Mondays. Once again deadline for submission of chapters is July 31st, 2015.</td>
</tr>
<tr>
<td>Source of data: you must mention all the sources where you got information. This your project is useless if you didn’t conduct any interview with a midwife. I am the truth you present this to external examiner, you’ll score below standard. See me for more correction. Make your format ur work before uploading it to the portal, watch me word videos.</td>
</tr>
<tr>
<td>Methodology: use this methodology: Structured System Analysis and Design Methodology (SSADM) was employed. You can now analyze how you apply the different step in your work.</td>
</tr>
<tr>
<td>Analysis of the Proposed System: This should be elaborated in details. Note that the external examiner focus more on your chapter 3.</td>
</tr>
<tr>
<td>You can add this: Most mothers may not be comfortable asking some questions in the public since antenatal classes are being held as a group class.</td>
</tr>
</tbody>
</table>
6. CONCLUSION

With the issues students and their project supervisors do encounter during their project period, it was very innovative to come out with a new idea and a new system to replace the normal traditional way in which projects are being done in higher institutions. With this design, students’ projects soft copy and the supervision process will be stored permanently in the project portal and this will be a good medium where referencing and research could be made at anytime. In furtherance, this will reduce the issues of late submission of chapters, supervisors will in turn make necessary corrections easily on the work or topics submitted. Also, it will not completely eliminate the normal face to face interaction and explanations between the two parties. This work will help students improve their knowledge of e-learning because, most of the works or chapters will be uploaded on the internet to the students’ project portal, and a link is included in the design where students upload each chapter and abstract. Finally, it will be very much advisable to higher institutions to adopt this new system because not only will it eliminate the old stressful traditional way, but will also be a new development in the university school system.

REFERENCES

Authors’ Biographies

Odirichukwu Jacinta Chioma is on Faculty at the Department of Computer Science, Federal University of Technology, Owerri, Imo State, Nigeria. She obtained her first degree in Computer Science with First Class honours from Madonna University, Okija, Anambra State, and Her Masters degrees in Computer Science from University of Ibadan. Her research interest is Computer Networks, Web development, Embedded System and Robotics. She is a member of Nigeria Computer Society and also a member of Organisation of women in Science for Developing World (OWSDW). She can be contacted by phone on +2348037394691 and info@jacoportal.com.

Dr. (Lady) Euphemia Chioma Nwokorie is a top flight scholar. She graduated with first class honours in computer science from the University of Nigeria Nsukka. At the University of Port Harcourt she had her Masters Degree in Computer Science. She finally got her Doctor of philosophy Degree in Computer Science from the University of Port Harcourt. She is indeed a guru in computer science, widely travelled and has published many articles in local and international journals. She is a member of the Computer Professionals Registration Council of Nigeria (CPN), Nigeria Computer Society (NCS), Institute of Electrical Electronic Engineers (IEEE), Organisation of Women in Science for the Developing World (OWSD) and Nigerian Women in Information Technology (NIWIT). She is presently a Senior Lecturer and Acting Head of Department of Computer Science, Federal University of Technology Owerri. She is happily married and blessed with five lovely children.

Asimole Izunwanne Gaston is a student of the department of Computer science, Federal University of Technology, Owerri, Imo State, Nigeria. He is also a member of National Association of Science Students, FUTO chapter, and also a member of National Association of Computer Science Students FUTO. He can be contact at: 08130426501 and gastonasimole@gmail.com.

Deanwi Tonye Lekara is a computer analyst/programmer. He is the owner of Softpointapps. SOFTPOINT APPS is an information technology company. Focused in providing services in Enterprise Solutions & business specific application development. He can be contact at: 08037892987 and www.softpointapps.com.