Information Retrieval And Indexing Implementation For A Digital Academic Transcript System

Ogwuche Grace
Department of Computer Science, University of Jos, Jos, Nigeria

Oyerinde O.D.
Department of Computer Science, University of Jos, Jos, Nigeria

Researchjournali.com
Exclusive Journal publishing house

ABSTRACT

This article presents a framework for an information retrieval and indexing management system for a digital transcript system which will help overcome the undesirable problem associated with student's grades, slow and strenuous accessibility of student report and records and poor information management within the school. The objective of this research is to propose and implement an algorithm that would optimize data and preserve them in an eco-friendly manner. A review was observed that a strong case has been made for the adoption of indexing in an information retrieval system. Despite the challenges facing users, they have shown a remarkable willingness to adapt to this technology but the time, high cost of infrastructures and complexity involved during the process required for full implementation of indexing has been the major limiting factor. The proposed algorithm developed will ensure easy flow of information and accurate information management for digital transcript systems. The methods of indexing and query formulation have been proposed and implemented. Evaluation also shows that a good percentage have analyzed the system to be efficient, satisfied and fairly easy to learn.

Keywords: Information retrieval system, Indexing, Transcript system, Query formulation

1. INTRODUCTION

Information retrieval (IR) is the field concerned with the organization and retrieval of knowledge-based information focusing mainly on textual information, but with the growing amount of multimedia as well as more complex databases, the nature of IR has changed. An information retrieval system is designed to retrieve the documents or information required by the user community. It should make the right information available to the right user. Thus, an information retrieval system aims at collecting and organizing information in one or more subject areas in order to provide it to the user as soon as it is asked for. (Onwuchekwa & Jegede., 2011)

Increasingly, the activity of retrieving information is mediated by computerized systems to be able to retrieve information desired by the user, representation of the users information must be matched somehow with the representation of the documents contained in the system. The inherent subjective facet inside a user's interest implies that the problem of satisfying a user information need is always going to be open (Birger., 2004).

There are two intellectual processes associated with IR – indexing and retrieval. (Alkafije & Ajam., 2013). Indexing is the process of assigning metadata to content items. Most commonly metadata consists of subjects or terms. These are words or phrases that describe the content. (William & M.D., 2010.)

The transcript system is designed to allow the printing of student transcripts on demand for other educational institutions. It will also allow the tracking of student progress toward fulfilling graduation requirements. The system will store student course information (course final marks) for several years while the student is

attending school. Every year, at the end of a reporting period (semester) when a course is finished, student marks are posted or copied into a transcript data table. This information can then be stored as long as desired. IR problems associated with Digital Transcript systems are:

- Time consumption,
- Lack of proper documentation,
- Poor Storage,
- Intensive cost,
- Inaccurate record keeping and
- Poor information management within the schools.

Today, because of ever growing digital data, it is very important to optimize these data and preserve them in an eco-friendly manner. We present a method to digitize the academic transcript so that the digital data cannot be retrieved by any unauthorized user. In this way, we can save a lot of digital space, which was necessary to save those digital academic records of each student.

2. RESEARCH OBJECTIVE

The objective of this research paper is to propose and implement an algorithm that would optimize data and preserve them in an eco-friendly manner. An information indexing and retrieval system is necessary for academic transcript because it would aid data preservation and optimization making it possible to blend into the ever growing digital data system. This will thus reduce the stress of searching through the whole records just to find a particular record. This system would be aimed at reducing the strenuous accessibility to students report and record. As insights would be gained through more efficient information access (Oyerinde et al., 2013), it would improve the schools transcript management system making it better and user friendly. This research paper is designed to specifically propose and implement an innovative means of indexing and retrieval of archived academic transcripts maintaining integrity and security requirements.

3. LITERATURE REVIEW

Modern document collections often contain groups of overlapped documents. (Andrei et al, 2006) proposed a method made to describe a new document representation model where related documents were organized as a tree, allowing shared content to be indexed just once. The processes index encoding and query evaluation helped support the model by encoding the model in an inverted index and evaluating free text-based queries based on the encoding respectively.

(Birger, 2004) investigated the use of references and citations technique as an integrated part of automatic indexing and retrieval system. He proposed a "boomerang method" which automatically translates the natural

language expression needed into references that are been used as weighted seed documents in a citation search. The method reduces overlap that occurs due to uncertain citations that are been emphasized.

(Roi, 2008) investigated some strategies for index size reduction of IR systems, suggesting several strategies that can efficiently render IR systems by reducing the index size. He addressed two different approaches for index compression:

- *Document reordering:* This enhances the compression of index by reordering the collection of document. It reduces the consumption of resources, in the case of IR, memory and bandwidth transmission compression helps greatly not only to cut off the cost of extra storage, but also to reduce substantially query answering times of an IR system.
- *Static index pruning:* It reduces not only query time, but also disk occupancy, and it is query-independent, hence it can be done off-line without any query information, static pruning can be beneficial for retrieval effectiveness, if handled with care.

(Alharith & George, 2013) proposed a method to analyze documents by using tokenization, preprocessing (converting upper case letters to lower, Unicode conversion, removing diacritics from letters, punctuations, or numbers), stop words removal, and stemming to save indexing time and space. The method proposed was aimed at processing documents and indexing the proposed steps which in turn saves indexing time and space especially for a huge set of data.

(Ian, 2006) developed a newly improved key phrase algorithm called Key Extraction Algorithm KEA++ based on machine learning aimed at achieving the following:

- Eliminating the occurrence of meaningless phrases.
- Yielding a dramatic improvement in performance.
- Lowering the requirements for training data.
- Using a machine technique on terms encoded in a controlled vocabulary.

The algorithm worked in two stages

- Candidate identification which identifies the terms related to the documents content
- Filtering which identifies the most significant terms based on certain features by using learned models.

The study offered support and proposition on the enhancement of automatic key phrase extraction. The research showed that an upgrade can be made that would bring about adaptation between the systems and other structured indexing vocabularies and domains.

(Andrei et al, 2006) introduced and described a Document Representation Model that organized related documents as a tree allowing indexing of shared contents to be done once. The document at a particular node contains shared and private (unique) content. The processes index encoding and query evaluation helped

support the model by encoding the model in an inverted index and evaluating free text-based queries based on the encoding respectively.

The model could be applied to practical IR applications such as web, email. Experimental results showed that the method was able to reduce the size of inverted index and improve query performances due to the synergistic effect of fewer cursor moves on a smaller index.

4. DATA ANALYSIS/ FINDINGS

4.1 METHODOLOGY

The algorithm below gives a summary of the methods of indexing and query formulation used.

1. Begin
2. Login (username and password)
3. If username and password = valid, GOTO 4 else 2
4. Crit = select criteria
5. Switch (Crit)

Case A: select * from table_name where A = "value inputted";
Case B: select * from table_name where B = "value inputted";
Case C: select * from table_name where C = "value inputted";
. . .
. . .
. . .
Case n: select * from table_name where N = "value inputted";

6. Endcase
7. If LOC = N + 1, then set LOC = 0
8. Output result
9. Stop

Note that the criteria Crit to be selected and the various search cases are optional and can be substituted with an individual's choice of criteria that can be used for the search.Also, the table name is not constant as the name can be changed depending on an individual's table name.

Hence, the select statement for the above algorithm can be written using this standard format;

Select * from table_name where Search criteria = value inputted;

The table 4.1 shows the combination of the 4 different fields that can be used for search depending on an individual's choice; it is each represented with alphabets A, B, C, D. It also shows the indexes and queries for the individual combinations gotten and formed during the process of indexing.

4.2 METHODOLOGY IMPLEMENTATION

Table 4.1: Index & Query Table

Indexes	Combinations	Queries
1	A	SELECT * FROM table_name where A=" ";
2	B	SELECT * FROM table_name where B=" ";
3	C	SELECT * FROM table_name where C=" ";
4	D	SELECT * FROM table_name where D=" ";
5	A,B	SELECT * FROM table_name where A=" " AND B=" ";
6	A,C	SELECT * FROM table_name where A=" " AND C=" ";
7	A,D	SELECT * FROM table_name where A=" " AND D=" ";
8	B,C	SELECT * FROM table_name where B=" " AND C=" ";
9	B,D	SELECT * FROM table_name where B=" " AND D=" ";
10	C,D	SELECT * FROM table_name where C=" " AND D=" ";
11	A,B,C	SELECT * FROM table_name where A=" " AND B=" " AND C=" ";
12	A,B,D	SELECT * FROM table_name where A=" " AND B=" " AND D=" ";
13	A,C,D	SELECT * FROM table_name where A=" " AND C=" " AND D=" ";
14	B,C,D	SELECT * FROM table_name where B=" " AND C=" " AND D=" ";
15	A,B,C,D	SELECT * FROM table_name where A=" " AND B=" " AND C=" " AND D=" ";

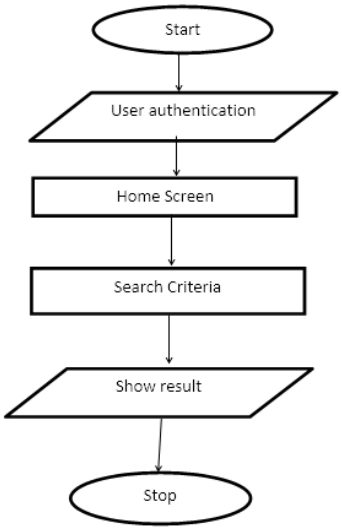

Figure 4.1. Search flowchart

The system being a web based application is meant to be user friendly. When the user runs the application, an interface is displayed. This interface contains a login page where an authorized user (admin) logs in using his password and username uniquely identified to him. The interface contains links corresponding to the various capabilities of the system. The capabilities include searching for a specific record using specified search criteria, viewing all the entries in the database in a tabular form, viewing a student's result and record one after another.

User Authentication

This is the Login part of the system where the user has to input his username and password. If the username and password isn't same as the data inputted in the database, the user is been denied access into the system else he can go proceed and perform the desired task.

Home Screen

This serves as the interface where every action has its individual links. The home page is user friendly; it hence makes the necessary page navigations very much easy.

Search Criteria

This can either be a singular search, that is, if you are searching using a particular detail or it could be a multiple search, if you are searching by combining multiple criteria. This is where table 1 which is the index & query table comes in handy. Each search criteria or combination has its individual query statements performed along. So the user tends to get his result based on the search criteria inputted.

Search Result

This is a display of the result of the inputted search criterion.

From the flowchart in figure 4.1, users search can be easily done and it provides a quick and effective result, thereby improving the indexing and retrieval process.

4.3 CASE ANALYSIS

The algorithm is a linear search algorithm, its complexity is measured by the number of combinations and comparisons required to find ITEM in array, where it is made up of N elements. The linear search algorithm goes through the list from the beginning of the list until reaching the end of list (array).

Data model representing the best case, worst case and average case produces table 4.2 below. For each case, the number of steps is expressed in terms of n, the number of items in the list.

Table 4.2: Model Representing Case Analysis

Model	Number of Comparisons/combinations (for n = 15)	Comparisons as a function of n
Best Case (fewest comparisons/combination)	1 (target is first item)	1
Worst Case (most comparisons/combination)	15 or 16 (target is last item or item not found at all)	n or n+1
Average Case (average number of comparisons/combinations)	8 (target is middle item)	n/2

4.4 RUN-TIME ANALYSIS USING ASYMPTOTIC VALUE

The algorithm which is a linear search can be analyzed and given asymptotic value for the run time using the best, worst and average cases of analysis. The value for each case is shown in table 4.3.

Best case= Θ (1)

For the average, $\sum_{i=1}^{n+1} \Theta(i) / (n+1)$

$= \Theta((n+1)*(n+2)/2) / (n+1)$

$= \Theta(n)$

Worst case= Θ (n)

Table 4.3: Run-Time Analytical Values

Analysis	Asymptotic values
Best case	Θ(1)
Average case	Θ (n)
Worst case	Θ(n)

The rows of Table 4.4 starting from the top, are the array indices, the data and combinations stored at the indexed location, and the search action that would give the fastest result starting from best to worst depending on the item or data to be searched for. Suppose you want to retrieve the result of a certain student and you are provided with a matriculation number, index1 would be the best. To retrieve detail of student name and department, index8 would be best.

Table 4.4: Linear Search Algorithm

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Combinations / Criteria	A	B	C	D	A, B	A, C	A, D	B,C	B, D	C, D	A,B, C	A,B, D	A,C, D	B,C, D	A,B,C, D
Search action	Mat_no							Name,dept							

4.5 USABILITY TEST/ANALYSIS

A Usability test was performed which is a technique used in user centered interaction designed to evaluate a product by testing it on users, this gives direct input on how real users use the system. Usability test focuses on measuring a human-made product's capacity to meet its intended purpose. We created the following tasks for the various users of the system to perform:

- Log in to the system(Admin)
- View all students record(admin)
- Search for a student record(Admin)
- View/Print student transcript(admin)

The following tables highlight the results of the tests performed:

Table 4.5: Task 1

	Tasks	Rating	Admin	User
1.	Log in	Success rate	5/5	None
2.		Task times	0secs-1 minute	None
3.		Error rates		
4.		Problems experienced	Admin inputted wrong username or password	None
5.		Comments/recommendations	Constant use of the system to enhance familiarity	None

Table 4.6: Task 2

	Tasks	Rating	Admin	User
1.	View all students record	Success rate	4/5	None
2.		Task times	0secs-1 minute	None
3.		Error rates		
4.		Problems experienced	Admin could not find the record of any student whose details is not stored in the database	None
5.		Comments/recommendations	Students records and details should be updated and properly stored	None

Table 4.7: Task 3

	Tasks	Rating	Admin	User
1.	Search for a student record	Success rate	4/5	None
2.		Task times	0secs-10 minutes depending on the search criteria provided	0secs-10 minutes depending on the search criteria available
3.		Error rates		
4.		Problems experienced	Admin inputted wrong details due to spelling or typographical error	User provided wrong input details
5.		Comments/recommendations	Admin should be sure and cautious of the details given by the user before search.	User should be sure of the details they provide, if they are not sure of one criterion, another criteria should be tried.

Table 4.8: Task 4

	Tasks	Rating	Admin	User
1.	View and print transcript	Success rate	5/5	2/5
2.		Task times	0secs-15 minutes	0secs-1.5 minutes to confirm the details
3.		Problems experienced	None	User could not see some courses recorded or offered
4.		Comments/recommendations	Constant use of the system to enhance familiarity	User should be sure of the record given.

4.6 FINDINGS

The current system was designed in order to obtain detailed fact about the application area to be redesigned. Investigation also covered looking at the functional requirement of the present system and finding out whether the requirements and objectives of the present system are being achieved.

The proposed system has a single table that contains every possible combination of different search criteria. It could be singly picked or could have a multiple combination which has been individually indexed. Hence, this process of retrieval does not require searching through multiple tables as the existing system but instead, its retrieval will be centered on this new table that is made of records of student all combined and recorded.

Based on the usability test conducted, the following conclusions were drawn:

Learnability: from the analysis above as shown in the various tables 4.5, 4.6, 4.7, 4.8 respectively, the success rate of the individual tasks 1,2,3,4 was observed to be high and provided an optimal value and positive result. As observed, 75% of the process was a success; hence we concluded that the system is fairly easy to learn.

Efficiency: The system provided an optimal effort towards carrying out the function it was designed for. From table 4.7 (task 3), if the user exhibits and observes some form of cautiousness and clarity while on the system, that optimal result would be obtained from such effort, as observed, 70% effort was made.

Satisfaction: As observed from table 4.8 (task 4), 70% 0f the users of the system were satisfied with the way the system worked and are willing to efficiently use the system to enhance familiarity and provide ideas on how to make it better.

5. CONCLUSION

The research work solves the indexing and retrieval problem encountered in a digital transcript system. It seeks to reduce the time consumed in comparing and finding student records saved in different tables.

The system was able to achieve its set objectives which include proposing and implementing an algorithm that would optimize data and preserve them in an eco-friendly manner, improving the indexing and retrieval

process of the system. From the analysis given and evaluated, we can say that the system is fairly easy to learn, provides optimal effort to make it efficient and satisfies the need of the user.

However, as Oyerinde et al., (2013) indicated, implementation is not successful unless the system it produces is accepted and integrated into the work place it was designed for; it is our sincere and earnest desire that this model/framework be used within the Digital Transcript System of the University. This will enable the University benefit from the gains and insights derived from this research.

6. REFERENCES

Alkafije, A & Ajam, G. (2013). Improving Document Processing and Indexing by Preprocessing and Tokenization. *Journal of Babylon University/Pure and Applied Sciences*, 4(21).

Anderson, J. D. (1997). guidelines for indexes and related information retrieval devices.1001-1008.

Bendersky, M., Metzler, D., & Croft, W. B. (2010). Effective Query Formulation with Multiple Information Sources. Department of Computer Science, University of Massachusetts.

Birger, L. (2004). References and citations in automatic indexing and retrieval systems-experiments with the boomerang effect. Department of information studies, royal school of library and information sciences, 297.

Broder, A. Z., Eiron, N., Fontoural, M., Herscovici, M., Lempel, R., McPherson, J., Qil, R. & Shekita, E. (2006). Indexing Shared Content in Information Retrieval Systems.

Callan, J. P & Croft, W. B. (1993). An evaluation of query processing strategies using tipster collection. Department of computer science, university of Massachusetts.93-34.

Gonzalez , B. R. (2008). Indexing compression for information retrieval system. University of a Coruña.

Heide, B., Gerhard, K., & Marc-andre, M. (1999). Document classification methods for organizing explicit knowledge. Oregon Health and Science University.

Jasminka, D., & Bojana, D. B. (2000). Comparison of Information Retrieval Techniques: Latent Semantic Indexing and Concept Indexing.

Leena H. P., & Mohammed, A. (2004). A Semantic approach for effective document clustering using Word Net.

Onwuchekwa, E. O., & Jegede, O. R (2011).Information Retrieval Methods in Libraries and Information Centers. An International *Multidisciplinary Journal, Ethiopia*, 5(6), 108-120.

Oyerinde O.D, Adekunle, A. Y. Ebiesuwa. O. O. (2013). A Superficial Exposé of Data Warehousing: An Intrinsic Component of Modern Day Business Intelligence. *International Journal of Science and Research India Online ISSN: 2319-7064*, *2*(4), 468–473.

Salton, G., Buckley, C. & Allan, J. (1992). Automatic structuring of text files. *Electronic publishing*, 5(1), 1–17.

Saravanan, D. & Somasundaram, V. (2014). Matrix based sequential indexing technique for video data mining. *Journal of Theoretical and Applied Information Technology*, 3(67), 1992-8645.

Soergel, D. (1994). Indexing and retrieval performance; the logical evidence. *Journal of the American Society for Information Science.*

William, H. & M.D. (2010). Information retrieval: access to knowledge-based Resources.

Oregon Health and Science University. Department of Medical Informatics & Clinical Epidemiology, School of Medicine, Oregon Health and Science University, Portland, OR 97239.