A Survey: Framework to Develop Retrieval Algorithms of Indexing Techniques on Learning Material

Zamri Abu Bakar[1], Murizah Kassim[2], Mohamad Norzamani Sahroni[1] and Nurhilyana Anuar[1]
[1]*Center of Foundation Studies, Universiti Teknologi MARA, UiTM Selangor Kampus Dengkil, 43800 Selangor, Malaysia.*
[2]*Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 UiTM Shah Alam, Selangor, Malaysia.*
zamri@salam.uitm.edu.my

***Abstract*—This paper presents a review on indexing techniques to develop retrieval algorithms framework on learning material. Analysis of the framework was drawn from surveys on literature review and experiment on online campus Learning Materials. Data indexing problem of online learning material occurs as online data comprising many types, formats and words of documents on the system become larger daily. Thus, searching capability for relevant information becomes slower. Further, it becomes more difficult to get the correct information as the learning materials consists of multiple forms of documents such as words, images and videos. The objective of this research is to analyze the existing indexing technique in modeling new retrieval indexing algorithms framework mainly for data mining. Four existing indexing techniques for learning material were reviewed. It is identified that the best used technique are Inverted File, Suffix Array, Suffix Tree and Signature File. Based on the four techniques, characterizations and parameters to enhance a new indexing technique (NIT) was identified and five User Acceptance Tests (UAT) were performed. A framework for NIT was designed and experiments are done on a Campus Learning Material. Identified parameters were successfully inserted in the five test experiments. The conceptual framework was continuously applied to develop NIT for retrieval algorithms on learning material. This research is significant for fast accessing on real life campus learning material system that benefits users and fast retrieval of needed information.**

***Index Terms*—Indexing Technique; Data Mining; Retrieval Algorithms; Learning Material; Text; Graphic; Video; Framework.**

I. INTRODUCTION

Document or words indexing is one of an important technique that deals with document search. Today's documents are generated daily at extraordinary rates due to the ease of publishing online. As the volume of documents is increasing exponentially, the ability to search particular information for relevant information is crucial. Indexing is an effective and ubiquitous tool for reducing query execution time. Indexing is data retrieval structure that can reduce the amount of processed data when a query is executed [1]. It is also one of data mining techniques, where feature extraction in data mining reveals the significance of similarity measures based on geometric attributes in detecting the relationships between data [2], [3]. Indexing also helps to identify and point data resource location based on file name or keyword in database records comprising text, video or graphic [4].

Many indexing techniques or algorithms have been invented like inverted files [5], [6] signature files [7], pattern matching [8], suffix arrays [9] and many more. Some of identified important document properties in information retrieval are text or characters [10], link and multimedia [11]. However, the recent documents heterogeneous nature makes it difficult to have uniform indexing solutions across different data sources. Based on its characteristics, information can be designated either structured, semi-structure or unstructured document [12]. Structured implies data identified in an organized structure like relational database table. Unstructured document refers to data that does not have a pre-defined structure like videos, images, and text.

Learning material (LM) is a spectrum of educational resources stored in a server accessed by users to support specific learning objectives [13]. In relation to this, problems may occur when users spend a long time to retrieve the learning materials, such as navigating through a particular folder. User needs to browse different folders to get other relevant learning materials and this seems to consume a lot of time for users. Currently, the learning system lacks of search facilities for users to retrieve learning materials due to many courses, students, lecturer, faculty and program involved in the learning system. Problem also occurs when there is a lot of learning material in the learning database system, such as PowerPoint, PDF files, word documents and others.

This paper presents a review on Indexing Techniques to develop retrieval algorithms framework on Learning Material. Reviews on theories which compared four existing indexing technique for LM by using inverted file, suffix array, suffix tree and signature file has been conducted. Based on the reviews, a framework for indexing algorithms and techniques was designed based on identified characters and parameters found for new retrieving LM algorithms.

II. INDEXING TECHNIQUES ON LEARNING MATERIAL

Current indexing technique for Learning Material (LM) documents, such as Power Point, PDF and Word Processing document has become less effective for retrieval outcome in terms of precision and recall for a LM [14]. LM consists of big data available on the web, and it keeps on growing over the time. The continuous growth of information and the large volume of existing data have caused problems for users to obtain relevant resources effectively [15]. The LM documents which comprises different formats, such as text, multimedia and audio are stored and retrieved educational content [16]. These types of digital resources are also reused

to support learning [17]. The large amount of LM has resulted in the discovery and retrieving of relevant and trusted resources become complex. Using indexing technique from information retrieval can be a solution to the problem of enquiring the resources of interest for learning purposes. Finding and searching information can be more efficient and accurate when the searching is supported by the latest computer technology. In relation to this, researchers have been focusing on ways to improve the measurement of finding relevant information in terms of recall, accuracy and fall-out, where certain tests were conducted [18], [19].

Indexing technique for slide presentation is performed to text only. There is no indexing for text in picture or slide presentation to retrieve the information needed. However, the text in the picture has also become one of the important elements in obtaining the effectiveness of information needs. Due to the abundance of multimodal devices, users need to search information from the learning system which consists of document, images, sounds and video through many possible means [20]. They used the learning system as a platform to find the relevant document to fulfil their interest related to their study. Table 1 shows the review on Indexing Techniques (I-T), Indexing Parameters (I-Par), Algorithms Indexing and Test Collection on type of data collections. Based on Table 1, it is identified that the most used indexing techniques are Inverted File (IF), Suffix Array (SA), Signature File Index (SFI) and Semantic Indexing - Conceptual Indexing-Query Expansion (SCQ) in the reviews.

Table 1
Reviews on Indexing Parameter and Algorithms

I-T	I-Par	Algorithms Indexing	Test Collection
IF	Text	Blocked sort Inversion	Natural Language Words
IF	Text	Naive Bayes Ant Colony	Bank documents
IF	Text	Maximally Stable Extremal Region (MCER) algorithm	Images
IF SA	Document	Single Linkage Clustering Algorithm Estimating DF on TF GREEDY	Web document
IF	Text	Word Count	Web page
IF	Arabic texts	Semantic Index	Arabic Texts
IF	Real-time text	Distributed and Lucene index	Text Document
IF	Bitmap	Adaptive Index BIDS_MapBIDS_Reduce Load Balance Algorithm	Structured data with 2 million tuple
SFI	Text and Image	Codebook Text Matching between image	images
IF	Image	Image indexing	Images
SCQ	Document	SemTree	Web pages, medical report, logs and textual documents
IF	XML Document	Structural index tree Text index tree	Structured data and semi structured data in XML Document
IF	Document	B-tree based provided by MongoDB	SlideShare

III. METHODOLOGY

This section presents five (5) level of process to develop NIT framework. The important process for developing algorithms and framework are: strategize the test plan, test design, test execution and evaluate test result [19]. Thus, before developing the algorithms, there are five levels of process need to be executed. Figure 1 presents the five levels, which are the data collection, Information Extraction from Learning Material, Pre-Processing, Indexing Creation and Retrieval Engine.

Figure 2 presents the data collection process, where a sample of LM data was collected in a day. Then, the data were characterized based on learning material files. With the collected data, document characters and parameters were identified. Later, a query process was carried out to make judgement related to the relevance of each query. This process includes match tracking, feature learning material and files search engine. Lastly, the retrieval results based on similarity ranking of the LM were saved in database.

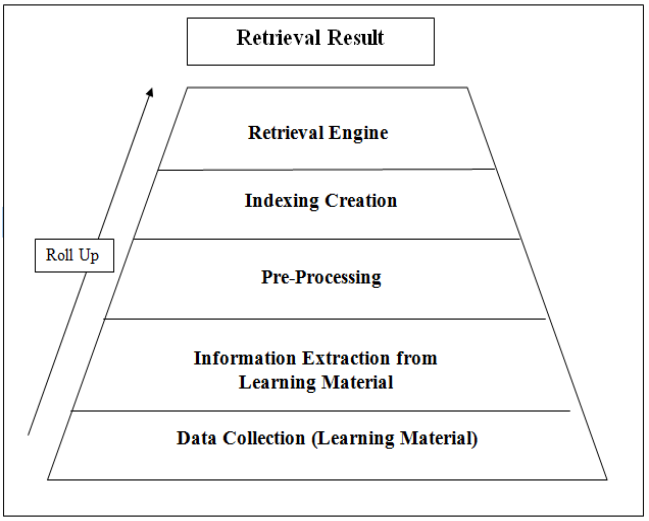

Figure 1: Method Flow in Indexing Algorithm

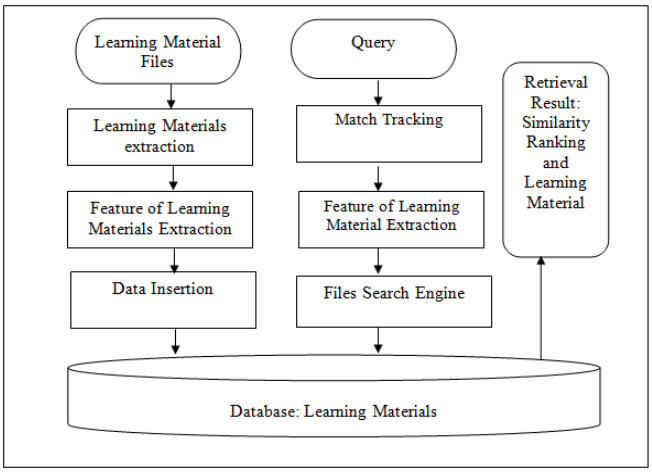

Figure 2: Learning Material Test Collections

Figure 3 presents five phases of the algorithms development for retrieval and indexing. Phase 1 involves the construction of the test collections that is linked to the pro-processing of LM. Phase 2 involves the evaluation of the basic indexing techniques. Phase 3 is the design and construction of the proposed indexing techniques. Phase 4 evaluates the new indexing technique and lastly, phase 5 performs the the User Acceptance Test (UAT). Based on the methodology on the reviews and theories on existing indexing Technique, a framework that Facilitates the

Retrieval Algorithms on indexing is presented, followed by the analyses and results.

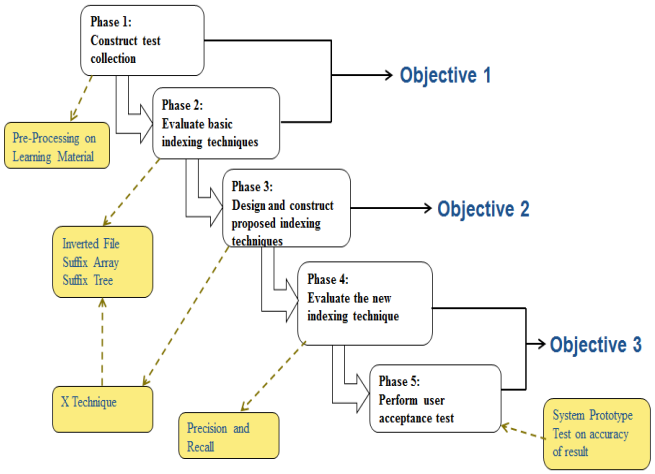

Figure 3: Phases of Algorithms Development

IV. Analysis And Indexing Framework

A framework for new Indexing Technique to Facilitate Retrieval Algorithms is designed and developed. This framework consists of four phases as described below.

A. *Evaluate Basic Indexing Technique*

Figure 4 presents the evaluation of the Basic Index Technique for the proposed online LM. Based on the review techniques theory, the basic indexing technique was identified to test the effectiveness and accuracy of information needs for LM. In the first phase, the chosen Indexing was differentiated according to three types of document properties, which are Text, Links and Multimedia. All of the three types of documents were tested using basic indexing techniques, which are Inverted files (IV), Suffix arrays (SA) and Signature files (SF). All evaluated basic constructed indexed file for each query based on gathered learning material.

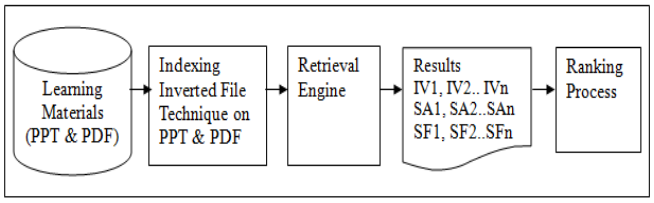

Figure 4: Indexing Technique

B. *Evaluate Modified Indexing Techniques*

Figure 5 shows the second phase of the framework that is the construction of modified indexing techniques on the online LM. This technique evaluated the modified constructed existing indexed files for each query on learning materials.

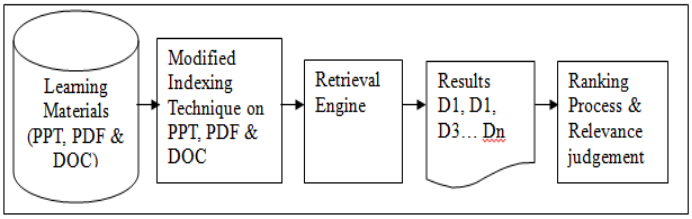

Figure 5: Evaluation of Modified Indexing Technique

C. *Evaluate New Indexing Technique*

The third phase is developing evaluation on the NIT as shown in Figure 6. NIT is enhanced from the existing inverted Index, Suffix Array and Signature File to develop a modified indexing. Info extraction was split into two mediums, which are the text and the Images plus Picture. Then, both were combined to test the Indexing Creation, Index File and Retrieval Engine. Both tests were done on PowerPoint and PDF documents.

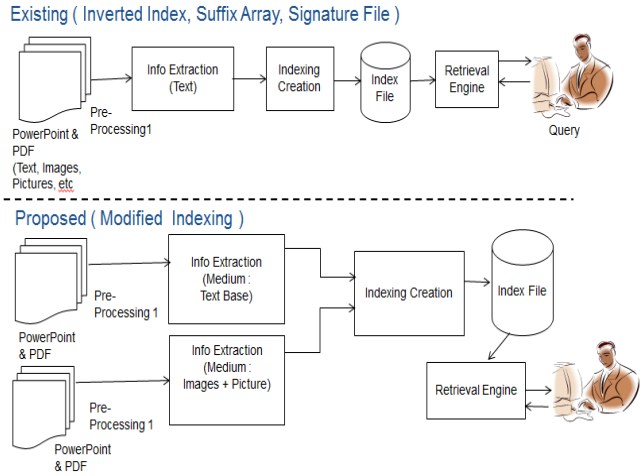

Figure 6: Evaluate New Indexing Technique

The method constructed a new indexing techniques based on basic and new indexing techniques. Then, an evaluation on the new constructed indexed technique for each query of the document was carried out. The proposed modified indexing comprises pre-processing for two different documents, which are the extraction on medium text and image plus pictures. This two indexing were created as parameters for retrieval engine algorithms.

D. *User Acceptance Test*

Finally, User Acceptance Test (UAT) was carried out, where the effectiveness of the new algorithm of indexing technique was measured against the current indexing technique using two (2) processes, which are:

1. Two groups of users (student and lecturer) will conduct the test of searching process in retrieving the information needs.
2. Perform user acceptance test for the new algorithm of indexing technique.
3. Evaluate user searching performance based on recall and precision outcome.

Lastly, both the existing and the new method were compared in experiment or tests. Figure 7 shows experiments 1 to 3. Figure 8 shows experiments 4 and 5, where 400 LM on PPT and PDF were evaluated using the retrieval engine.

Five experiments were conducted and the possible results based on keyword and searching were analyzed.

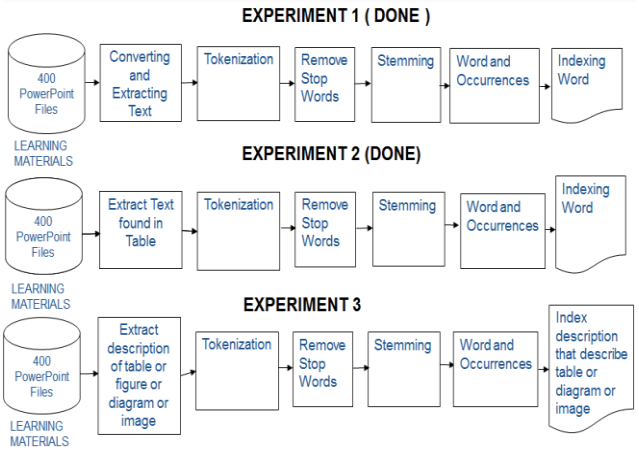

Figure 7: UAT Experiments 1 to 3

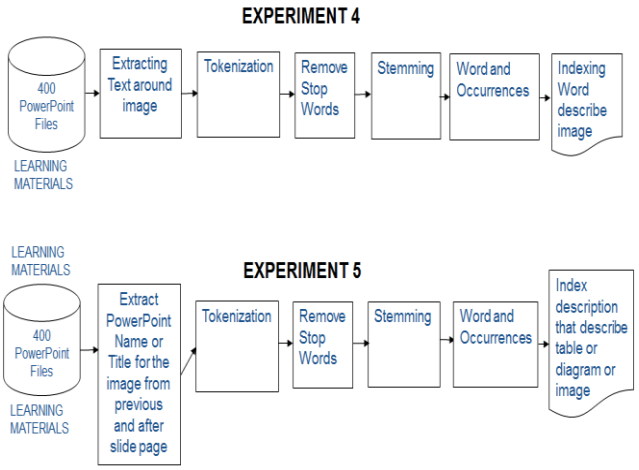

Figure 8: UAT Experiment 4 and 5

V. CONCLUSIONS

This research presented a development of a new indexing technique that can be used for searching technique in multi-discipline learning materials. A Framework was designed based on real learning material and reviews on previously implemented indexing techniques for documents in a system or database. A framework of New Indexing Technique (NIT) has been successful designed and the next phase is to develop an algorithm call Technique Facilitate Retrieval, which focuses on online Learning Material. A new indexing technique will satisfy the users' needs of information by giving relevant information searched by users, followed by important parameters set. By having indexing technique based on Campus life LM, an effective and fast retrieval document system is enhanced, hence it is useful for users such as academic, students or administrators. This helps the process of searching information which benefits users, especially teaching and learning on blended learning environment in a campus university.

ACKNOWLEDGMENTS

Authors would like to thank Universiti Teknologi MARA (UiTM) for the test sampling data on online Learning Material. Authors also would like to thank Center of Foundation Studies, UiTM Selangor Campus for the support grant.

REFERENCES

[1] Day, R.E., An Afterword to Indexing It All: The Subject in the Age of Documentation, Information, and Data. *Bulletin of the Association for Information Science and Technology*, (2016), Vol 42(2): p. 25-28.

[2] Darvishi, A. and H. Hassanpour, A Geometric View of Similarity Measures in Data Mining. *International Journal of Engineering-Transactions C: Aspects*, (2015), Vol 28(12): p. 1728.

[3] Hashemzadeh, E., Hamidi, H., Using a Data Mining Tool and FP-growth Algorithm Application for Extraction of the Rules in Two Different Dataset, *International Journal of Engineering (IJE) Transactions C: Aspects*, (2016) Vol. 29, No. 6.

[4] Golub, K., et al., A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, (2016), Vol 67(1), p. 3-16.

[5] Yadav, A.K., D. Yadav, and D. Rai, Efficient Methods to Generate Inverted Indexes for IR, in *Information Systems Design and Intelligent Applications, Springer*. (2016), p. 431-440.

[6] Bilimoria, D.M., P.A. Patel, and M.S. Rajpoot, Supporting Linked Databases in Keyword Query Searching Using Density Inverted Indexes, *Emerging Research in Computing, Information, Communication and Applications, Springer*. (2016), p. 367-375.

[7] Constantin, C., et al., AS-Index: A Structure For String Search Using n-grams and Algebraic Signatures. *Journal of Computer Science and Technology*, (2016), Vol 31(1): p. 147-166.

[8] Ganguly, A., Shah, R., Thankachan, S.V., Parameterized Pattern Matching--Succinctly. *Cornell University Library, arXiv preprint arXiv*, (2016), 1603.07457.

[9] Ponvert, E., W. Kalter, and J. Szalay, Suffix searching on documents, *Google* Patents. (2016).

[10] Singh, M.P., Dhaka, V., Handwritten character recognition using modified gradient descent technique of neural networks and representation of conjugate descent for training patterns. *Database Journal*, (2008), Vol 5, p. 20.

[11] B'ez, Y.A. and R.C.C. Jiménez. Indexing structured documents with suffix arrays. *2012 12th IEEE International Conference on Computational Science and Its Applications (ICCSA)*, (2012).

[12] Zhu, C., Zhu, C., Li, Q., Kong, L., Wang, X., Hong, X., Associated Index for Big Structured and Unstructured Data, *Springer Web-Age Information Management*, (2015), p. 567-570.

[13] Krašna, M., M. Duh, and T. Bratina. E-learning next step Learning materials for students, *2014 37th IEEE International Convention Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (2014).

[14] Abdullah, M.F. and K. Ahmad. Business intelligence model for unstructured data management. *2015 IEEE International Conference on Electrical Engineering and Informatics (ICEEI)*, (2015).

[15] Ferguson, R. and S.B. Shum, Towards a social learning space for open educational resources. *Collaborative LearningBook, The Open University* (2012), Vol 2: p. 309-327.

[16] Sampson, D.G., Zervas, P., Sotiriou, S., Agogi, E., Sharing of open science education resources and educational practices in Europe. *Open Educational Resources: Innovation, Research and Practice*, (2013), p. 105.

[17] Sicilia, M.A., Garcia, E., On the concepts of usability and reusability of learning objects. *The International Review of Research in Open and Distributed Learning*, (2003), Vol 4(2).

[18] Sharma, M., Patel, R., A Survey on Information Retrieval Models, Techniques And Applications. *International Journal of Emerging Technology and Advanced Engineering*, (2013), p. 2250-2459.

[19] Arbain, A.S., M. Kassim, Saaidin, S., Systematic Test and Evaluation Process (STEP) approach on Shared Banking Services (SBS) System identification. *2010 2nd International Conference on Education Technology and Computer*. (2010).

[20] Smeaton, A.F., Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Information Systems*, (2007), 32(4): p. 545-559.