Vol 8. No. 4 – December, 2015
African Journal of Computing & ICT

www.ajocict.net

A Topic Modelling-Based Framework for Mining Digital Library's Text Documents

***T.A. Olowookere, B.O. Eke. & L.U. Oghenekaro**
Department of Computer Science,
University of Port Harcourt
Choba, Rivers State, Nigeria.
toluwase_olowookere@uniport.edu.ng, eke.bartholomew@uniport.edu.ng, linda.igwebike@uniport.edu.ng

*Correspondence Author

ABSTRACT

The impacts and contributions of scholarly research works in the economic growth and sustainability of any nation cannot be overemphasized. The digital library has emerged as a reliable resource for provisioning researchers with scholarly knowledge (eruditions that result from the research works) which are documented and published in form of journal articles, technical reports or conference proceedings amongst others. However, as academic institutions and publishers around the world are choosing to make their thesis, dissertation and journal articles available in digital form, this electronic repository of knowledge (the digital library), though organized, is flooded with an exploding large collections of documents filled with hidden but useful information in form of the varieties of topics of discourse inherent in them. Thus making it imperative to develop a flexible means to automatically discover the topics that pervade the collections in such digital library. Currently, the application of topic modelling technique holds great promises and has tremendous results in extracting the topical contents of document corporal. In this regard, this paper presents a Topic Modelling-based framework for mining document collections of a digital library for topical structure discovery alongside topic-based similarities search between document collection pairs, by means of integrating the base topic modelling algorithm and inverted Kullback-Leibler divergence mechanism. The framework shows potency in the automatic discovery of topical structures of document collections and it as well describes the capability of finding topic-based similarities between document collection pairs.

Keywords— Digital Library, Document Collection, Text mining, Topic Modeling

African Journal of Computing & ICT Reference Format:
T.A. Olowookere, B.O. Eke. & L.U. Oghenekaro (2015): A Topic Modelling-Based Framework for Mining Digital Library's Text Documents. Afr J. of Comp & ICTs. Vol 8, No. 4. Pp 19-26.

1. INTRODUCTION

Modern day access to document collections of library has almost evolved from the traditional library collections to the digital library collections which tends to supplement the concept of a library as a repository of knowledge. With many academic institutions and journal publishers around the world choosing to make digital versions of their dissertations, theses, articles and other academic documents accessible online, an overwhelming volume of information is becoming available and widely accessible in digital libraries. The digital library as an organized collection of digital documents provides information users with easy access to these collections of documents that can be searched through to retrieve authoritative documents in any specific topic area which is far useful to any researcher working in such a topic area than an unfocused collection like the web [1].

Though the collections of digital libraries are well organized, the fast growing number of documents in the collections has expanded drastically; an increase in the amount of documents that a researcher has to search in other to satisfy his research need(s). Also, owing to the vastness of the documents in a digital library, its collections are consequently filled with vast hidden information in form of the varieties of themes (topical structures) touched in them. One cannot ordinarily and easily identify the topics of discourse within the large collections of a digital library as simple search does not and cannot present us with the knowledge of the topics of interest that pervade these collections. One way for us to be able to have a grasp of the topical information that run through the document collections of a digital library is through the approach of mining the text document collections of such digital library.

Text mining is a growing and exciting research area that tries to solve the information overload problem especially in digital libraries, by using techniques from data mining, statistics, machine learning, information retrieval (IR), natural language processing (NLP), and knowledge management. The process of text mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results [2]. In a manner that is similar to data mining, text mining seeks to extract useful information from sources of data through the identification and exploration of patterns that are interesting. However, in the case of text mining, the data sources are document collections, and interesting patterns are found not among formalized database records but they are found in the unstructured textual data in the documents in these collections. According to [3], the compounded name of text mining suggests that it is either the *discovery* of texts or the *exploration* of texts in search of valuable, yet hidden information.

Techniques of text mining have much to offer digital libraries and their users. In this work, we describe a framework for mining document collections of a digital library for topical structure discovery, such that the topic model based framework allows for the use of loose-coupling integration mechanism for integration with document collections built in a widely used digital library software system such as the Greenstone digital library system [4].

Topic modelling can be described as a form of text mining, a way of identifying topical patterns in a corpus. It is a method for finding and tracing clusters of words (called "topics" in shorthand) in large bodies of texts. A topic is essentially a recurring pattern of co-occurring words, i.e. cluster of words that frequently occur together in documents in statistically meaningful way. Formally, it is a probability distribution over words in a vocabulary. Probabilistic topic modelling is a relatively new approach that is being successfully applied to explore and predict the underlying structure of discrete data, such as text. A topic model, such as the Probabilistic Latent Semantic Indexing (PLSI) proposed by Hofmann [5], is a statistical generative model that relates documents and words through latent variables which represent the topics [6]. By considering a document as a mixture of topics, the model is able to generate the words in a document given a small set of latent variables (or topics). Inverting this process, i.e. fitting the generative model to the observed data (words in documents), corresponds to inferring the latent variables and, hence, learning the distributions of underlying topics.

Topic models (e.g., [5, 6, 7,]) are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model being a *generative model* for documents specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents.

A topic model takes as input a collection of text documents, such as book pages. It outputs a preset number of "topics", which are probability distributions over the words in the collection. Topics are essentially determined by which words occur together across the collection. The most likely words for each topic can then be used to provide human-interpretable keywords for the topic [8]. This work addresses the problem of identifying the specific topics that pervade a digital library's document collections by proposing an enhanced topic modelling-based modular algorithmic framework for useful topical structure discovery in collections and topic-based similarity discovery between pairs of collections in a digital library.

The rest of this paper is structured as follows. Section 2 provides a critical analysis of related work while Section 3 discusses the theoretical framework, providing detailed descriptions of the algorithms that are hybridized in the framework. An activity diagram that depicts the proposed framework is discussed in Section 4, with Section 5 highlighting the contributions to knowledge made by the study. Section 6 concludes the paper by summarizing the study and makes recommendations while Section 7 states the focus of possible future work.

2. RELATED WORKS

Rauber et al. [9] worked on mining digital library collections in the *SOMLIb* digital library system, built on neural networks to provide text mining capabilities. At its foundation they used the *Self-Organizing Map* to provide content-based clustering of documents. The *Self-Organizing Map* is a popular unsupervised neural network model, and a variation of this model, i.e. the *Growing Hierarchical Self-Organizing Map* (*GHSOM*), were used to topically structure a document collection similar to the organization of real-world libraries. By using this extended model, i.e. the *GHSOM*, they further detected subject hierarchies in a document collection, with the neural network adapting its size and structure automatically during its unsupervised learning process to reflect the topical hierarchy (structure).

By mining the weight vector structure of the SOM, using LabelSOM their system was able to select keywords describing the various topical clusters. They demonstrated the capabilities of the *SOMLib* system using collections of articles from various newspapers and magazines in digital library. However their system does not consider any detection of topical similarity among collections of articles but only captures the topical patterns and the keywords describing them, among other issues. Our work proposes a semantic framework to achieve the discovery of topic-based similarity between collection pair through the use of an inverted Kullback-Leibler divergence measure.

Another related work is on mining conference document collections. As various conferences are held every year about different topics and huge volume of scientific literature is collected about conferences in digital libraries, mining conference document collections has become an important problem of attention. Juanzi et al. [10] in their work, noted that previous approaches to conference texts mining which mined conferences by using semantics-based intrinsic structure of the words present between documents (modelling from document level) ignored semantics-based intrinsic structure of the words present between conferences. They therefore proposed a generalized topic modelling approach based on Latent Dirichlet Allocation (LDA) named as Conference Mining (ConMin), by considering semantics-based intrinsic structure of the words present in conferences by modelling from conference level (CL). It provided grouping of conferences in different groups on the basis of latent topics (semantically related probabilistic cluster of words) present between the conferences. They used discovered topics to find associations between conferences (topically related conferences, conferences correlations) and showed temporal topic trends of conferences. Experimental results showed that proposed approach significantly outperformed baseline approach in discovering topically related conferences. This paper however proposes a topic modelling/Kullback-Leibler based framework for topic discovery mechanism which does not only take into consideration, the semantics-based intrinsic structure of words present between document collections but also the document level words structures (topics) with integration to digital library repository.

3. THE THEORETICAL FRAMEWORK

Topic discovery task in documents can be carried out by applying some machine learning and statistical techniques on textual data. The algorithm of our proposed framework integrates topic modelling method and an inverted Kullback-Leibler divergence measure in accomplishing the topical structure discovery task and document collections comparison task. The topic modelling algorithm enhanced is the Latent Dirichlet Allocation (LDA) [7], which the proposed framework employs in the task of discovering corpus-wide and document-level topical structure that is inherent in document collection. The topic modelling technique is an unsupervised machine learning paradigm—no labels are required on documents.

The inverted Kullback-Leibler divergence measure is used to find the semantic similarity between document collections, say documents collection d_i and documents collection d_j, based on their respective topic distributions $p(t|d_i)$ and $p(t|d_j)$. The framework adopts the loose coupling approach for integration with digital library system. The topic modelling algorithm is being enhanced in this work to analyze document collections for the discovery of topics and compare collection pair based on these discovered topics (with the aid of the inverted Kullback-Leibler divergence measure).

3.1 Analysis of the LDA Topic Model Algorithm

The LDA Topic model reflects an intuition that documents contain multiple topics and being a part of the larger family of probabilistic modelling, it assumes that these topics are explicitly specified before any data has been generated (that is, the model assumes the topics are generated first before the documents are). In generative probabilistic modelling, data are treated as arising from a generative process that includes *hidden variables,* thus, the LDA algorithm has a generative process as the algorithm below shows:

The LDA Probabilistic Generative Process

Step 1: For each topic number_1 to topic number_k, [$t_{1:K}$];

a. Draw a distribution over words $p(w|t) \equiv t_k$ (i.e., Per-Topic word distribution).

Step 2: For each document *d* in the collection [1…*D*];

a. Randomly draw a distribution over topics $p(t|d) \equiv \theta_d$ (i.e., Per-Document topic distribution)

b. For each word *w* in the document;

i. Randomly draw a topic from the distribution over topics in Step 2a (i.e., Per-document per-word topic assignment)

ii. Randomly draw a word from the corresponding distribution over the vocabulary (word).

We note that a *Topic*, as used in this study is a distribution over words/vocabulary. The distribution employed in drawing the per-document topic distribution (which is a latent/hidden variable) is called the Dirichlet distribution. Hence, the name Latent Dirichlet Allocation. The generative process of the LDA defines a *joint probability distribution* over both the hidden (latent) random variables and the observed variables. Clearly stating, the observed variables are the *words of the documents* while the hidden variables are the topical structure— *per-topic word distribution* (the *topics*), *per-document topic distribution*, and the *per-document per-word topic assignment*.

Of course the goal of the LDA is not the generation of random documents through these distributions, but rather inferring the distributions from observed document. This inference process uses the observed words of the documents to infer the hidden topical structure, a process which can be seen as reversing the generative process— discovering the hidden structure that likely generated the observed document collection. Therefore, the goal of using topic modelling technique in our proposed framework is to automatically discover the inherent topics from the collection of documents. We emphasize in this paper that the algorithm does not have any information about the themes of the documents and that the documents are not labeled with topics or keywords, hence its unsupervised learning nature.

The inference process uses the joint probability distribution from the generative process to compute the *conditional probability distribution* of the hidden (latent) variables given the observed variables. This conditional distribution is what is referred to as *Posterior Distribution*.

For the generative process, the joint distribution goes thus: ;

$$p(t_{1:K}, \theta_{1:D}, Z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(t_i) \prod_{d=1}^{D} p(\theta_d) \times \left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(w_{d,n}|t_{1:K}, Z_{d,n}) \right) \ldots\ldots\ldots (1)$$

As earlier stated in this analysis, the inference process which is the computational task of inferring the hidden topical structure from the document is concerned with the computation of the posterior distribution (i.e., the conditional distribution of the hidden variables given the observed documents).

For the inference process, therefore, the conditional distribution (Posterior) using the joint distribution is;

$$p(t_{1:K}, \theta_{1:D}, Z_{1:D}|w_{1:D}) = \frac{p(t_{1:K}, \theta_{1:D}, Z_{1:D}, w_{1:D})}{p(w_{1:D})} \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

3.2 Analysis of the Kullback-Leibler (KL) Divergence method

Being a method employed in the measurement of distance between two probability distributions, the Kullback-Leibler (KL) divergence will be used in our proposed algorithm as a means to find the semantic similarity between two document collections *di* and *dj* based on their respective topic distributions. The Kullback-Leibler divergence is actually a distance function rather than similarity function, this is because it usually achieves its minimum when the two probability distributions being compared are maximally similar to each other.

According to Nelken and Shieber [11], Kullback-Leibler divergence is an asymmetric *dissimilarity* measure between two distributions say *x* and *y*, it measures the added number of bits that are needed to encode events that are sampled from *x* using a code based on *y*. We therefore invert the KL divergence in this study, to achieve similarity measure.

For computation of the similarity between two document collections d_i and d_j:

The topic distribution of document collection d_i is $p(t|d_i)$ and,

The topic distribution of document collection d_j is $p(t|d_j)$.

∴ The Kullback-Leibler distance (D_{KL}) between their topic distributions is given as:

$$D_{KL}\left(p(t|d_i)||p(t|d_j)\right) = \sum_{x} p(t|d_i)_x \log_2 \frac{p(t|d_i)_x}{p(t|d_j)_x} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where x here, represents the topics t_k in the individual documents collection.

From the foregoing, the algorithm of the proposed Topic Modelling based documents mining framework is thus made modular and consists of the following steps:

Modular Algorithm of the Topic Modelling based Documents Mining Framework

Module 1

Begin

Step 1: INPUT: Text Document Collection "d_c" from digital library's repository

Step 2: Supply the number of Topics "k" to discover, proportion threshold *h*, number of word to print *r* and number of iterations "*i*"

Step 3: Perform preprocessing on the document collection

a. Tokenize the documents
b. Remove Stop-word using the stop-word_file
c. Perform case folding (optional)

Step 4: Learn the topical structure of the collection from the Preprocessed Vector Space Model

a. Get *i, and k*
b. Compute the joint distribution for the generative process as thus;

$$p(t_{1:K}, \theta_{1:D}, Z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(t_i) \prod_{d=1}^{D} p(\theta_d) \times$$

$$\left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d)\, p(w_{d,n}|t_{1:K}, Z_{d,n})\right) \ldots\ldots\ldots (4)$$

c. Compute the conditional probability for the inference process based on the joint distribution as thus;

$$p(t_{1:K}, \theta_{1:D}, Z_{1:D}|w_{1:D}) = \frac{p(t_{1:K}, \theta_{1:D}, Z_{1:D}, w_{1:D})}{p(w_{1:D})} \ldots\ldots\ldots\ldots\ldots (5)$$

d. Iterate the inference process by *i* and get *h* and *r*

Step 5: Output the topic distributions (in percentage) from the inference process for the text document collection "d_c":

$$p(t|d_c) = p(t_{1:K}, \theta_{1:D}, Z_{1:D}|w_{1:D}) \ldots\ldots\ldots\ldots\ldots (5)$$

Step 6: Repeat step 1 to step 5 for inputted document collections "*C*"

Module 2

Step 7: Choose document collections d_i and d_j to compare

Step 8: Calculate the semantic similarity between the topic distributions of d_i and d_j as obtained in step 5 as thus;

a. Compute Kullback-Leibler distance (D_{KL}) between their topic distributions as:

$$D_{KL}\left(p(t|d_i)||p(t|d_j)\right) = \sum_x p(t|d_i)_x \log_2 \frac{p(t|d_i)_x}{p(t|d_j)_x} \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

b. Compute the inversion of the KL as:

$$S_{KL}\left(p(t|d_i)||p(t|d_j)\right) = e^{D_{KL}\left(p(t|d_i)||p(t|d_j)\right)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Step 9: Output the Similarities from Step 8b (in Percentage)
End.

The descriptions of the mathematical notations and random variables used is as follows:

$t_{1:K}$ represents the topics;
where each t_k is a distribution over the vocabulary/words (i.e., per-topic word distribution).

θ_d represents the topic proportions for the *d*th document (i.e. per-document topic distribution) and $\theta_{1:D}$ for documents 1 to D;
where $\theta_{d,n}$ is the topic proportion for the *k*th topic in document *d*.

Z_d represents the topic assignments (i.e., the per-document per-word topic assignments) for the *d*th document, and $\theta_{1:D}$ is that for documents 1 to D;
where $Z_{d,n}$ is the topic assignment for the *n*th word in the document *d*.

w_n represents the obsereved word for the document *d* and
$w_{1:D}$ is observed words of the documents 1 to D;
where $w_{d,n}$ is the *n*th word in document *d* and is an element from the fixed vocabulary.

Since the resulting value of the KL represents a distance measure, therefore, to achieve the similarity function in this work, such that the function increases as similarity increases (i.e., achieving its maximum when the two distributions being compared are maximally similar), we adopt the inversion mechanism by inverting the Kullback-Leibler distance value as shown in the equation (7), using the exponential method.

4. ACTIVITY DIAGRAM OF THE PROPOSED FRAMEWORK

The activity diagram of the proposed framework is as depicted in figure 1. It shows a model of the important activities within the framework.

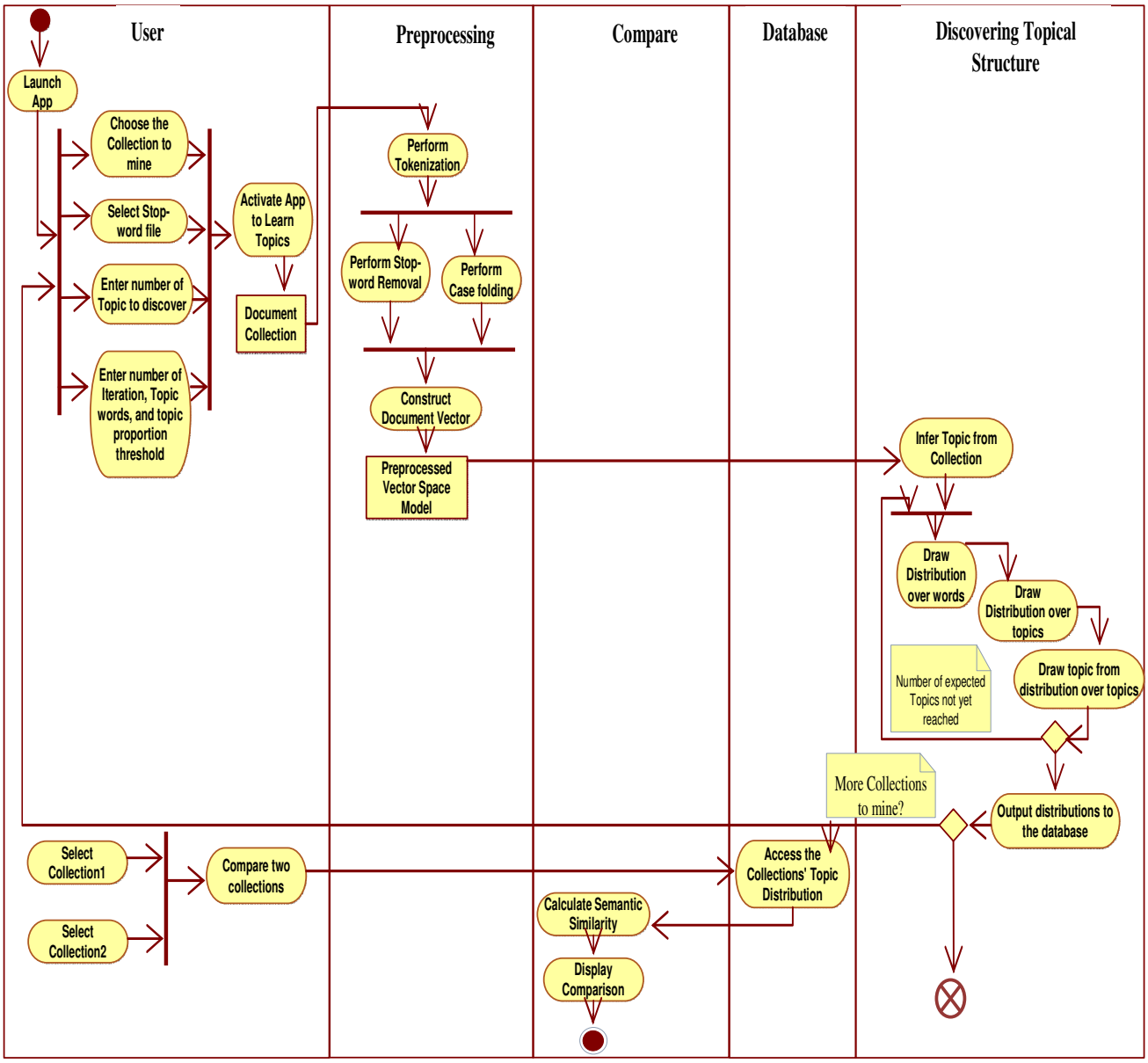

Figure 1: UML Activity diagram of the Proposed Framework

It is essentially a flowchart, showing flow of controls from one activity to another. An activity diagram represents an operation on some classes in a system that results to changes in the state of the system. Five (5) swim-lanes are shown in the activity diagram;

a. User- which depicts those activities that need the direct intervention of the user to trigger some other activities of the system.
b. Preprocessing- which reveals the activities that relate to the preprocessing of document collection.
c. Discovering topical structure- which relates to the activities that occur in the topical discovery process.
d. Database- which shows the activity performed on the database records.
e. Compare- models the operations that are performed by the system in comparing collections pair.

5. CONTRIBUTION TO KNOWLEGDE

The framework is a platform to aid digital and information systems designers and librarians in designing text mining capabilities that focus on topical structures discovery, reducing complexity of available learning materials with respect to their "huge volumes", to avoid information overload on the part of users of such information systems. This would afford librarians the flexibility to create effective application for discovering the topical structures that run through the collections in their digital library archives thereby enhancing content-based organization of the available document collections.

6. CONCLUSION

We have developed a topic modelling-based framework for discovering the topical structure of document collections of a digital library. The framework also describes the capability of finding topic-based similarities between document collection pairs. Adopting a loose-coupling technique, we propose an integration of an application developed based on the proposed framework with a digital library system and such application can then be deployed to mine document collections in the digital library's repository, distinctly showing the topical structure of collections and similarities between collections where such exist. The framework proposed in this work will enhance text mining capabilities in a digital library system, including similarity search for the purpose of easier classification of documents.

The outcome of this study— the proposed framework, holds viable propositions for information system designers and computational researchers alike as a paradigm for developing effective and flexible means to analyze the ever-growing digital library collections based on their inherent topical contents.

It is therefore recommended that:

i. Developers of Information systems should integrate the framework in their design to include text mining capabilities with a focus on topical structures discovery.
ii. Librarians, should ensure deployment of the framework in digital library as an aid to discover the topical structure that run through the collections in their digital archives and thereby achieving content based organization of their collections.
iii. Computational Linguistics researchers should avail themselves with applications developed based on the framework in order to analyze digital library collections based on their topical contents.

7. FUTURE WORK

We intend in future work to investigate the effects of both loose-coupling and tight-coupling integrations of an application developed based on the proposed framework with a digital library environment, preferably an open source digital library system.

REFERENCES

[1] Y. Theng, E. Duncker and N. Mohd, "Design Guidelines for User-Centred Digital Libraries," in *Proc. 3rd Conf. on Research and Advanced Technology for Digital Libraries,* pp. 167-183, 1999.

[2] A. Porter, "Text Mining.," Technology Policy and Assessment Center, Georgia Institute, 2002.

[3] R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2006.

[4] K. Rajasekharan and K. M. Nafala, "Building up a Digital Library with Greenstone, A Self-Instructional Guide for Beginners," Thrissur, India, 2007.

[5] T. Hofmann, "Probabilistic Latent Semantic Indexing," in In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, pp. 50-57, 1999.

[6] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," in In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (ed), Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, 2005.

[7] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[8] D. Mimno and A. McCallum, "Mining a Digital Library for Influential Authors," in In JCDL'07 ACM, Vancouver, British Columbia, Canada, June 18–23, 2007.

[9] Rauber, A., and Merkl, D. (2003). Text Mining in the SOMLib Digital Library System: The Representation of Topics and Genres. *Applied Intelligence, 18*, 271–293.

[10] L. Juanzi, A. Duad, Z. Lizhu and M. Faqir, "Conference Mining via Generalized Topic Modeling," ECML PKDD, pp. 244–259, 2009.

[11] R. Nelken and S. M. Shieber (2006). *Computing The Kullback-Leibler Divergence Between Probabilistic Automata Using Rational Kernels.* Harvard University, Division of Engineering and Applied Sciences, Cambridge.

Authors' Brief

OLOWOOKERE, Toluwase Ayobami is currently a Computer Science PhD student at the Department of Mathematical Sciences, Ekiti State University, Ado- Ekiti in Nigeria. He received a B.Tech. (Hons) Degree in Computer Engineering from Ladoke Akintola University of Technology, Ogbomoso, Nigeria, in 2010. He holds Master of Science (M.Sc.) degree in Computer Science from University of Port Harcourt, Nigeria. His research interest lies within the areas of Text and Data Mining, Process Mining in Business Intelligence and Security, Computer Modelling and Simulation, and Machine Learning. He is a member of Institute of Electrical and Electronics Engineers, IEEE-Computer Society and a graduate member of Nigeria Society of Engineers. He can be reached via *toluwase_olowookere@uniport.edu.ng* or *+2347037986565.*

Dr. EKE, Bartholomew O. is currently the acting Head of Department and a Senior Lecturer at the Department of Computer Science, University of Port Harcourt Nigeria. He is a Practitioner who sees the widening gap between slow classroom and fast paced industrial halls and decided to remain neutral. He obtained his B.Sc. and M.Sc. degrees in Computer Science from the University of Port Harcourt. He received PhD in Computer Science from the same University. His research interest is on "working software", as a signatory to the Agile manifesto. He is a member of the Association for Computing Machinery (ACM). He can be reached via eke.bartholomew@uniport.edu.ng or +234-8037049586.

OGHENEKARO, Linda U. is a Junior Lecturer at the Department of Computer Science, University of Port Harcourt, Nigeria. She obtained her B.Sc. (Hons) degree in Computer Science from the University of Port Harcourt, Nigeria in 2010. She is currently at the concluding stages of her Master of Science (M.Sc.) degree programme in Computer Science at the same university. Her research interests include; Machine Learning, Database Management and Data Security. She is a member of the Nigeria Computer Society. She can be reached via linda.igwebike@uniport.edu.ng or +2348034379834.