Research Journal of Applied Sciences, Engineering and Technology 10(2): 206-213, 2015
DOI:10.19026/rjaset.10.2573
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 29, 2014 Accepted: December 18, 2014 Published: May 20, 2015

Research Article

Generation of Tag Clouds for E-learning Documents-a Value-added Service to Peer Learners

[1]M. Ravichandran and [2]G. Kulanthaivel
[1]Department of Computer Science and Engineering, Sathyabama University,
[2]Department of Electronics Engineering, NITTTR, Taramani, Chennai, Tamilnadu, India

Abstract: In an E-learning environment users (learners/facilitators) have access to a large collection of learning documents stored in various online databases. Retrieving the most relevant learning documents available on database-driven websites is often difficult, as great amounts of textual content is involved. A peer learner may require a general sketch of the digital document content available in the database in order to find out whether the document information is useful for his/her search requirements. In this research study, we present a method for generating tag clouds for E-learning documents as a value added service to peer learners. We propose a system that generates the cluster summary of e-learning document and provides visual representation of these documents stored in a database. Visualization of the content of the database can be helpful during the peer learners search process and to reach their study goals, thus serving as a value-added service. The uniqueness of this method is that it reveals the fundamental structure that provides the text document with certain semantics and is capable of retrieving the most appropriate information. The relationship between the tags is obtained using the Multi Objective Hierarchical Cluster (MOHC) technique. In this study, we tested our proposed method using different datasets and present the tag clouds obtained by the computations for each dataset. The experimental results of tag cloud generation for e-learning documents demonstrate the accuracy and effectiveness of our proposed approach.

Keywords: Document search, E-learning, tag cloud, visualization

INTRODUCTION

In this study, we primarily focus on the problem of information access in e-learning documents. Textual attributes in databases contain useful information that is not structured and hence sometimes the information is not properly processed. This lack of structure presents a disadvantage for users who can only perform syntactic queries in the text as it leads to the retrieval of inexact or wrong information (Campana *et al*., 2009). In the last decade, the amount of digital contents stored in database has increased exponentially. In this scenario of information explosion and several methods have been proposed to present relevant information in a user-friendly manner. For this reason, the concept of relevant information has been studied from the viewpoint of its representation (Harter, 1992) and user perception (Vakkari and Hakala, 2000).

Machine learning algorithms are used to arrange documents into meaningful categories that hold highly related documents, which are distinct from the documents of other groups (Manning *et al*., 2008). The major problem of the clustering is identifying groups of related objects in the data. The similarity between the objects is calculated with the help of suitable similarity function. The process of clustering can be very useful in the text domain as well, where the objects to be clustered are terms, paragraphs, etc. rather than entire documents. Clustering has been particularly useful for arranging documents to improve relevant information retrieval when a user browses a large collection of documents, particularly in the web environment (Anick and Vaithyanathan, 1997; Cutting *et al*., 1993). Hierarchical clustering algorithms have been studied extensively in the literature (Jain and Dubes, 1998; Kaufman and Rousseeuw, 1990) for records of different kinds including multidimensional numerical data, categorical data and text data. In recent years, new techniques have been proposed to aid users in the search, visualization and retrieval of useful information. One of the standard approaches is based on the creation of tag clouds. The tag clouds can be arranged with different visual features, such as a tag for the occurrence of each word, a global tag cloud where the frequencies are organized over all items and users, or a cloud that contains categories, with size of the cloud indicating the number of subcategories. Tag clouds can be used for basic user-centered tasks. The internet holds a lot of information that is

Corresponding Author: M. Ravichandran, Research Scholar, Department of Computer Science and Engineering, Sathyabama University, Chennai, Tamilnadu, India

constantly updated. It is hard for the users to know the recently added information of a particular database, or sometimes even know how to create a suitable query.

A typical scenario is one where a learner who can recognize the query he wants to perform is not proficient to start it by himself. In this situation, it is essential to provide the user with a set of query suggestions that can help in the search process. The tagging method solves this problem by categorizing the document resources using tags structured as a cloud. This tag cloud can be used to recover the categorized information at a later time (Rivadeneira *et al*., 2007). Thus, tag clouds can be serve as appropriate tools for searching the content of a database. Tag clouds fetch relevant information through the visual representation of the most linked tags (Hsieh *et al*., 2009). Tag clouds help users whose search terms are not clearly defined and enable users to recognize the terms from a set of likely queries represented by the tags (Hassan-Montero *et al*., 2010). Tag clouds are useful tools that help users with no previous experience related to document retrieval systems (Leone *et al*., 2011). In (Kuo *et al*., 2007), tag clouds that summarize the information retrieved from a database were presented. This application proposed provides results with tag clouds retrieved from the user queries. Tag clouds help in the query refinement process. In the same year (Watters, 2008), the tool Cloud Mine was presented. These tools perform several analyses of text content and help in global searches. Tag cloud can be seen as a summary visualization tool and we consider a keywords extraction technique to build a summary. The quality of extracting keywords depends on the keyword extraction algorithm used and several methods have been proposed (Kaur and Gupta, 2010). Approaches based on co-occurrence (Matsuo and Ishizuka, 2004) and machine learning algorithm with additional semantic information (Hulth, 2003) has been widely used for extracting keywords from a single document. In (Xu *et al*., 2010) authors suggest using new word features for the extraction of keyword and the generation of headlines, by using Wikipedia for understanding the background of a document.

An iterative approach for document keywords extraction based on the relationship between different subsystems (i.e., the relationship between topic sentence and word) is presented in Wei (2012). In (Kaptein, 2012) the authors explain an application in which tag clouds are used to navigate Twitter search results. This research summarizes tweets sets into word clouds, which can be used to get an overall idea of the contents of the tweets. An approach to summarize selections of text within a document is presented in Bohne *et al*. (2011). The authors proposed a solution based on an extraction algorithm, regardless of language or context. IWISE (Fung *et al*., 2009) is a method depends on an integration of obtainable reducing documents and technology for web summarization. Multilabel document summarization has also been studied recently. Recent research on multilabel summarization adopt various methods to obtain results and innovative techniques to achieve the best results (Tao *et al*., 2008; Wang *et al*., 2008; Wang *et al*., 2009).

Tag clouds have become very popular on the web. They allow the representation of entire websites in a compact way, through a set of tags whose size or color indicates the frequency of use of terms (Viegas and Wattenberg, 2008). Blake Shaw's visualization of the Delicious tags is presented in Shaw (2005). One popular application area for tag clouds is text summarization (Burch *et al*., 2013; Feinberg, 2010; Kuo *et al*., 2007). Here, tag clouds are used to give an intuitive and visually appealing overview of a text by depicting the words that occur more frequently. Such summarization readily provides the user information about the number and kind of topics present in a body of text. Typically, this statistical overview is achieved by positively correlating the font size of the depicted tags with the word frequency. When tag cloud visualization is used this way, the tags are words from a text corpus. The steps for creating a tag cloud include document loading into R, text corpus handling, data preprocessing, metadata management, creation of bag of words, the term-document matrices, frequent term association and document tag cloud visualization of peer learners document search. Our research focuses on the generation of document tag cloud visualization in R using an in-depth description of the modern text mining platform offered by tm package and word cloud (Feinerer, 2008).

METHODOLOGY

In this research, we propose a novel methodology for the generation of tag clouds for e-learning documents. The methodology is composed of the following steps:

- Design of the general architecture of the generation process of the tag for e-learning documents.
- Document preprocessing that includes data cleaning from a syntactic point of view, tokenization, stop word removal, white space removal, stemming and lower case conversion.
- Term-Document matrix construction, term frequency-inverse document frequency (tf-idf)
- Calculation and frequent terms association computation.
- Construction of multiobjective hierarchical cluster structure representation (MOHC) that facilitates grouping of the documents using an unsupervised learning method.
- Generation of tag cloud for documents by taking into account clustered data.

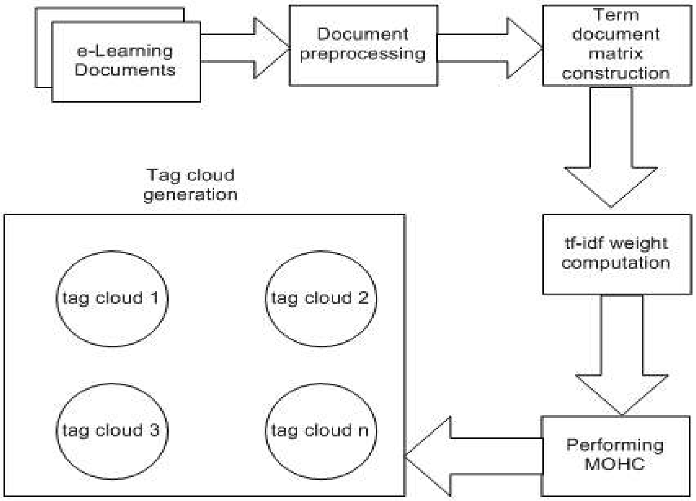

Fig. 1: General architecture of the generation of tag clouds for E-learning documents

General architecture of generation of tag clouds for E-learning documents: The application developed displays a condensed visualization of information represented as tag clouds for e-learning documents using the MOHC structure representation technique. This visualization assists peer learners in retrieving relevant documents. In this study, we present a new method to build a set of tag clouds that combines multiobjective hierarchical clustering task with the tag visualization. The general architecture is presented in Fig. 1. Give a document input dataset, a number of steps are followed to get the visualization of tags. The first step is the loading of the set of e-Learning document for corpus creation. Document preprocessing, metadata management and creation of a bag of words form the second step. At this stage, text cleaning techniques can be successfully applied to the document, such as white space removal, lower case conversion, stop words removal, stemming according to the weight values (tf-idf).The next important step is to construct a multi objective hierarchical cluster structure representation (MOHC) to group the documents. The final step is generating tag clouds to visually present the information obtained from the clustering process.

E-learning: E-Learning is an innovative method that complements the traditional learning system. E-learning enables users to learn the subjects of their choice at their own pace at anytime and anywhere. It can include educational training and the delivery of information and guidance from the facilitator. An information retrieval system is an application that stores and manages information on text documents. The goal of information retrieval (IR) is to provide users with those documents that will satisfy their information need.

Syntactic data preprocessing: The essential first step in the text document mining process is data preparation popularly known as syntactic preprocessing. Preceding the start of text document analysis, this stage typically includes the removal of nonessential characters such as such as white space, conversion of text characters lowercase and punctuation removal. This process is called noisy character data removal.

Term document matrix construction: After syntactic preprocessing, we proceed to construct the term-document matrix associated with text documents. This matrix is a simple vector having frequencies of all terms occurring in the text documents. In a term-document matrix, the rows correspond to the terms in the collection and columns correspond to documents. The term-document matrix contains sparse and non sparse values (zero or non-zero value).

TF-IDF (Term frequency-inverse document frequency) computation: TF-IDF (Term frequency-inverse document frequency) computation: Tf-Idf is a computational statistic that indicates the importance of a word in a document. In Information retrieval, tf-idf is used as a weight calculating factor. The higher frequencies of some words in the corpus indicate that these words are more common than the other words. Tf-idf term weighting methods are used by several search engines to score and rank retrieved documents based on relevance during a search query from the learners. Tf-idf is also used in various areas, including text document classification and summarization.

Let D denote the collection of documents (Corpus); t, d, ti and dj denote the term, document, i-th term and j-th document respectively and ti,dj belongs to D. The total occurrences of ti in dj are referred as the term frequency.

The idf (inverse document frequency) for a term ti is calculated by:

$$\text{idf}_i = \log\left(|D|/|\{d: t_i \in d\}|\,|D|/|\{d: t_i \in d\}|\right)$$

where |D| is the collection of documents in the text corpus. $|\{d\text{: } t_i \in d\}|$ is the number of documents in which the term occurs. The expression to calculate tf-idf is defined as tfij*idf$_i$ (Salton and McGill, 1986). A larger weight term indicates that the term appears less common in the collection of documents.

Multi Objective Hierarchical Cluster (MOHC): Once a dataset preprocessing and tf-idf weight computations are completed, then MOHC methods are applied to cluster the e-learning documents. MOHC used in machine learning algorithm shows the relationship between the terms based on a top-down approach. These clusters of terms are based on the similarity of the documents. The initial step of the hierarchical clustering approach is to create a new pseudo document. Each iterative process step takes next two nearest documents and merges them. These steps repeated until one large cluster having the original documents is formed. MOHC facilitates a tree-based representation to show the relationship of all the original documents visualized by tag cloud:

- Initialize sequence number m = 0 and initial clustering level L (0) = 0.
- Find the smallest dissimilar pair of clusters in the present clustering, say a pair (a), (b), according to:

  dis[(a),(b)] = min d[(i), (j)]

  where, the smallest is over all pairs of clusters in the present clustering.
- Increment the sequence number by one: m = m+1. Merge clusters (a) and (b) into a single cluster to form the next clustering m. The level of this clustering is set by:

  L(m) = d[(a),(b)]

- Change the proximity matrix, Y, by removing the rows and columns corresponding to clusters (a) and (b) and adding a row and a column to the new cluster. The proximity between the new cluster, denoted (a,b) and old cluster (c) is defined as: dis[(c), (a,b)] = min dis[(c),(a)], dis[(c),(b)]
- If all instances are in one cluster, stop the process. Otherwise, go to step 2.

Generation of tag clouds for E-learning documents: The term weight in the MOHC-structure is indicative of the frequency of a term in the document cluster. We generated a tag cloud for this document cluster created using the term frequency. The generation of the tag cloud is performed using an R package. The tag cloud thus generated can be used by peer learners who query the e-learning database.

In a linear normalization method, the weight term$_i$ of a descriptor is mapped to a scale size 1 through *freq*, where term$_{min}$ and term$_{max}$ are specifying the range of presented weights. D$_i$-display fontsize, f$_{max}$-maximum fontsize, term$_i$-count, term$_{min}$-minimum count, term$_{max}$-maximum count.

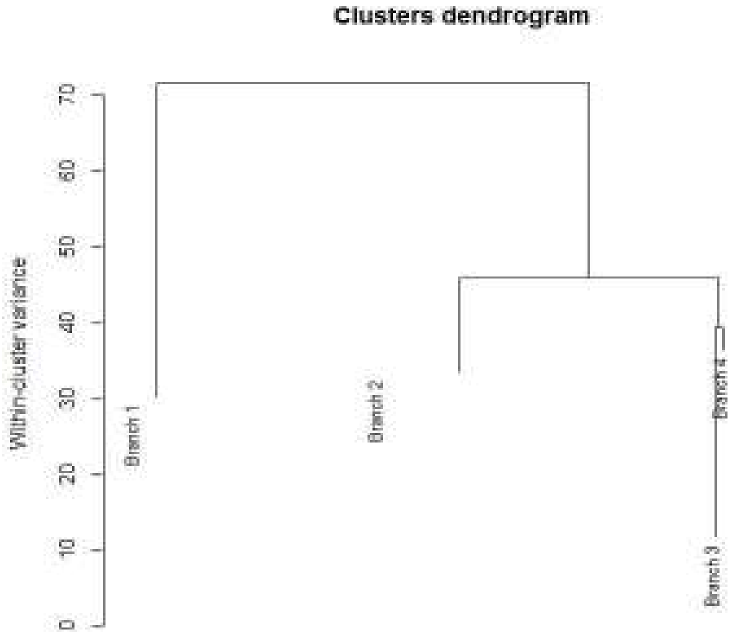

Fig. 2: MOHC cluster dendrogram

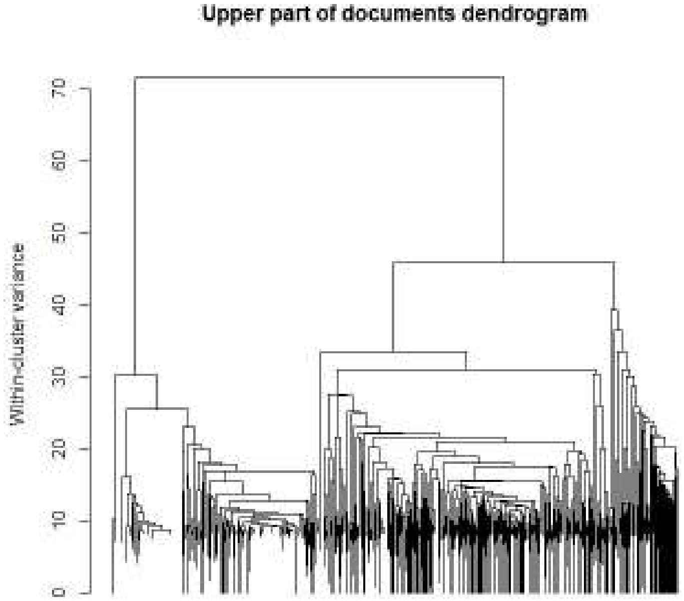

Fig. 3: Upper part of documents dendrogram

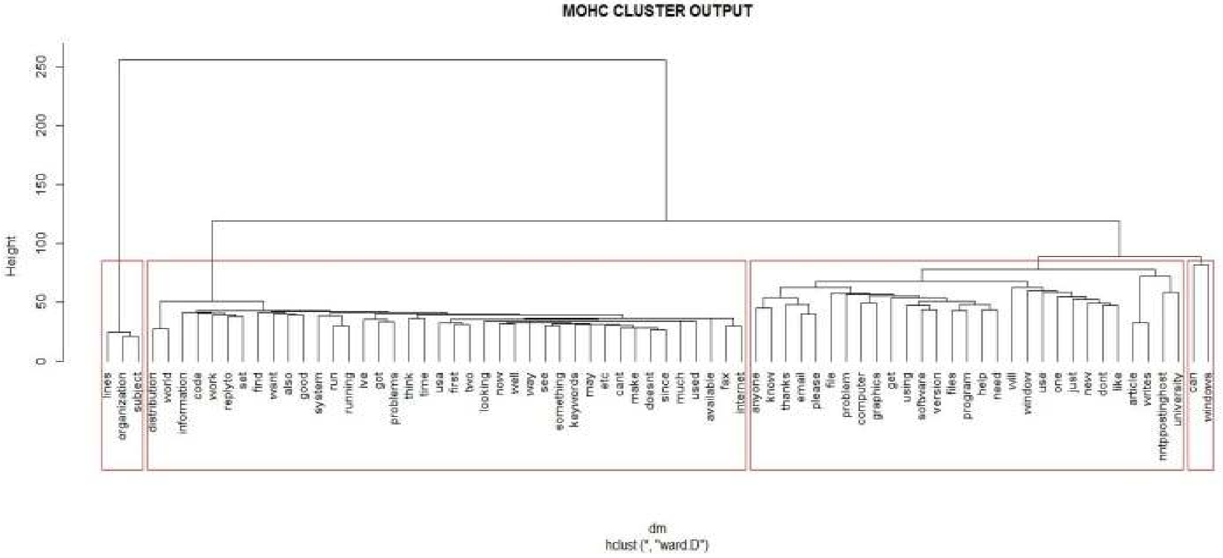

Fig. 4: MOHC cluster output

where,

t_{min}, t_{max} = Specifying the range of presented weights:

D_i = (f_{max}*($term_i$-$term_{min}$)/($term_{max}$-$term_{min}$)) for $term_i$>$term_{min}$; else D_i = 1

EXPERIMENTS AND RESULTS

In this section we explain the method of generation of different tag clouds obtained using our approach. In this study, we show various tag clouds generated using text on various topics. In our experiments, we used 20 news groups as dataset with documents on 4 topics -300 documents each on comp.os.ms.windows, computer. graphics, comp.sys.mac.hardware and comp. windows.x. The term document matrix was constructed. The various preprocessing steps involved to create the term-document matrix included white space removal, stop word removal, stemming, as well as stripping and removing sparse terms. Next, we performed tf-idf weight computation.

This was followed by MOHC representation to cluster the documents. Figure 2 shows the cluster dendrogram and Fig. 3 shows the upper part of the dendrogram. Figure 4 shows the MOHC cluster output. From the computed clustered documents, tag clouds were generated and these are shown in Fig. 5 to 8. The use of MOHC- based tags helps to peer learners better

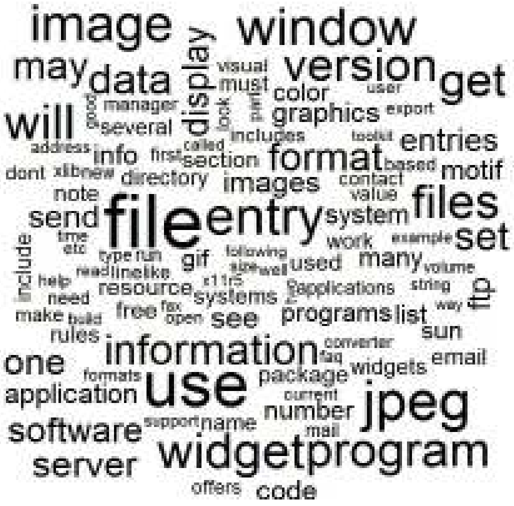

Fig. 5: Tag cloud for cluster 1

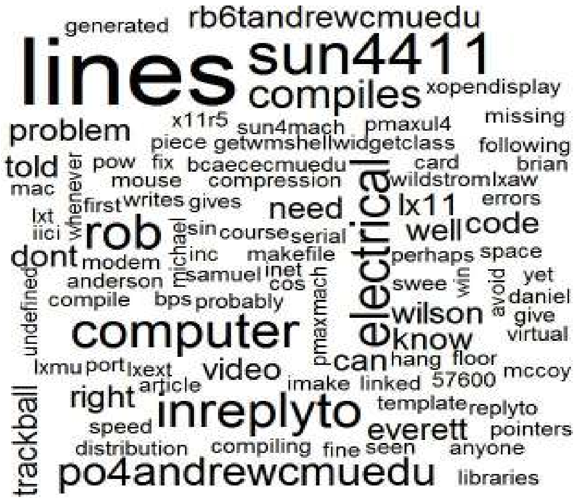

Fig. 7: Tag cloud for cluster 3

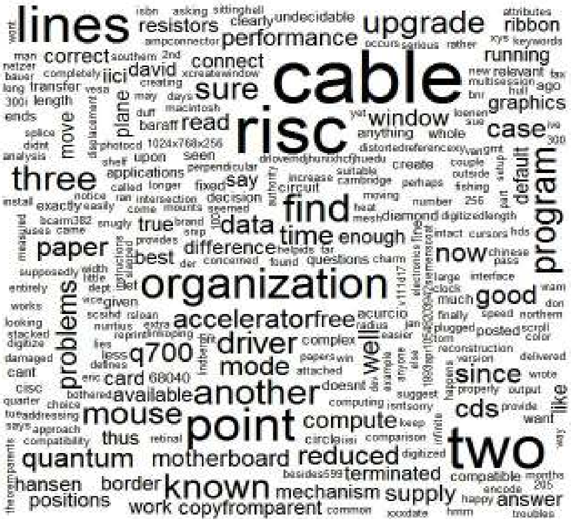

Fig. 6: Tag cloud for cluster 2

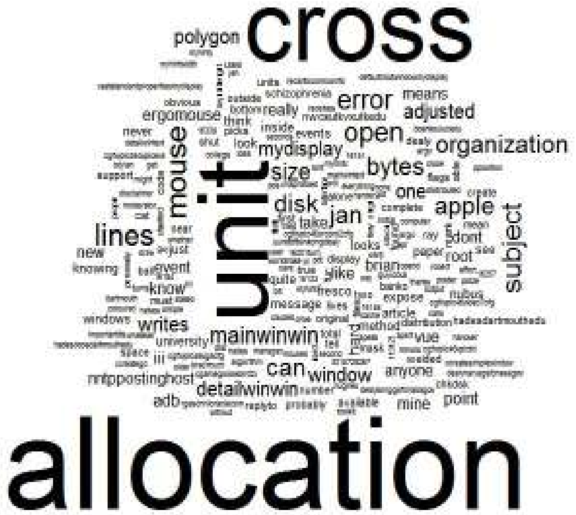

Fig. 8: Tag cloud for cluster 4

Table 1: Cluster summary analysis

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
No. of documents	8.00	261	462	469
% of Documents	0.670	21.8	38.5	39.1
Within cluster variance	16.12	33.5	23.4	27.5

understand the information because of the visualization technique employed. The metric relies completely on the distribution of topics in the text. Some topics may be better suited for visual representation than others because we have not adopted any conditional criteria.

Cluster analysis: The cluster summary is generated based on the number of correctly classified documents and cluster variance. The cluster variance gives the inter gap between the documents in each cluster. If we increase the number of clusters, the variance value is decreased owing to the increase in the number of correctly classified documents. Table 1 lists this summary.

CONCLUSIONS AND RECOMMENDATION

In this study, we have presented a method of e-learning using tag cloud generation. The value-added service system helps peer learners to retrieve the most appropriate learning document by means of visual effects. The new system is built following a modern methodology to extract document tag clouds from unstructured text documents using text mining methods. The methodology involves several steps, including document import, text document corpus handling, data preprocessing and metadata management, creation of bag of words, Term-document matrix construction, tf-idf computation, multi-objective hierarchical cluster structure generation and finally visualization of e-learning document tag cloud generation. This structure is visualized in the form of various tag clouds that represent the information of the key terms in the document and that allows the peer learners querying to successfully retrieve the most relevant e-learning document. To test our research approach, we selected datasets with documents dealing with various topics. For the tag clouds obtained, we calculated several metrics of coverage overlap and balance. As future work, we are planning to modify the sampling model using probabilistic latent feature extraction technique for enabling the dynamic changes in the document corpus and to avoid the sparsity. This work further can be extended to incorporate the Map Reduce model of parallel clustering of documents.

REFERENCES

Anick, P.G. and S. Vaithyanathan, 1997. Exploiting clustering and phrases for context-based information retrieval. Proceeding of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), pp: 314-323.

Bohne, T., S. Rönnau and U.M. Borghoff, 2011. Efficient keyword extraction for meaningful document perception. Proceeding of the 11th ACM Symposium on Document Engineering (Doceng'11). ACM, New York, NY, USA, pp: 185-194.

Burch, M., S. Lohmann, D. Pompe and D. Weiskopf, 2013. Prefix tag clouds. Proceeding of 17th International Conference on Information Visualisation (IV. 2013), pp: 45-50.

Campana, J.R., M.J. Martin-Bautista, J.M. Medina and M.A. Vila, 2009. Semantic enrichment of database textual attributes. In: Andreasen, T. *et al*. (Eds.), FQAS 2009. LNAI 5822, Springer-Verlag, Berlin, Heidelberg, pp: 488-499.

Cutting, D.R., D.R. Karger and J.O. Pederson, 1993. Constant interaction-time scatter/gather browsing of large document collections. Proceeding of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93), pp: 126-134.

Feinberg, J., 2010. Wordle. In: Steele, J. and N. Iliinsky (Eds.), Beautiful Visualization. O'Reilly, Sebastopol, pp: 37-58.

Feinerer, I., 2008. An Introduction to Text Mining in R. R News, 8(2): 19. Retrieved form: http://CRAN.R-project.org/doc/Rnews/. (Accessed on: Oct. 22, 2008)

Fung, C.C., W. Thanadechteemapat and K. Wong, 2009. iWISE, an intelligent web interactive summarization engine. Proceeding of International Conference on Machine Learning and Cybernetics, 6: 3457-3462.

Harter, S.P., 1992. Psychological relevance and information science. JASIST, 43(9): 602-615.

Hassan-Montero, Y., V. Herrero-Solana and V. Guerrero-Bote, 2010. Usabilidad de los tag-clouds: Estudio mediante eye-tracking. Scire: Representación y organización del conocimiento, 16(1): 15-33.

Hsieh, W., J. Stu, Y. Chen and S. Chou, 2009. A collaborative desktop tagging system for group knowledge management based on concept space. Expert Syst. Appl., 36(5): 9513-9523.

Hulth, A., 2003. Improved automatic keyword extraction given more linguistic knowledge. Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03). ACL, Stroudsburg, PA, USA: pp: 216-223.

Jain, A.K. and R.C. Dubes, 1998. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

Kaptein, R., 2012. Using wordclouds to navigate and summarize twitter search. Proceeding of EuroHCIR'12, CEUR, pp: 67-70.

Kaufman, L. and P.J. Rousseeuw, 1990. Finding groups in data: An introduction to cluster analysis. Wiley, New York.

Kaur, J. and V. Gupta, 2010. Effective approaches for extraction of keywords. J. Comput. Sci., 7(6): 144-148.

Kuo, B.Y.L., T. Hentrich, B. Good and M. Wilkinson, 2007. Tag clouds for summarizing web search results. Proceeding of the 16th International Conference on World Wide Web, pp: 1204-1205.

Leone, S., M. Geel, C. Muller and M.C. Norrie, 2011. Exploiting tag clouds for database browsing and querying. Inform. Syst. Evolut., 72: 15-28.

Manning, C.D., P. Raghavan and H. Schutze, 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

Matsuo, Y. and M. Ishizuka, 2004. Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. T., 13(1): 157.

Rivadeneira, A., D. Gruen, M. Muller and D. Millen, 2007. Getting our head in the clouds: Toward evaluation studies of tagclouds. Proceeding of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07), pp: 995-998.

Salton, G. and M.J. McGill, 1986. Introduction to Modern Information Retrieval. McGraw-Hill, ISBN: 0-07-054484-0.

Shaw, B., 2005. Semidefinite Embedding Applied to Visualizing Folksonomies. Manuscript, pp: 9.

Tao, Y., S. Zhou, W. Lam and J. Guan, 2008. Towards more text summarization based on textual association networks. Proceeding of the 2008 4th International Conference on Semantics, Knowledge and Grid. Beijing, China, pp: 235-240.

Vakkari, P. and N. Hakala, 2000. Changes in relevance criteria and problem stages in task performance. JDOC, 56(5): 389-398.

Viegas, F.B. and M. Wattenberg, 2008. Tag Clouds and the Case for Vernacular Visualization. ACM Int., 15(4): 49-52.

Wang, D., T. Li, S. Zhu and C. Ding, 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. Proceeding of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore, pp: 307-314.

Wang, D., T. Li, S. Zhu and Y. Gong, 2009. Multi-document summarization using sentence-based topic models. Proceeding of the ACL-IJCNLP 2009 Conference Short Papers. ACL and AFNLP, Suntec, Singapore, pp: 297-300.

Watters, D., 2008. Meaningful clouds: Towards a novel interface for document visualization. Online Notes, University of Chicago.

Wei, Y., 2012. An iterative approach to keywords extraction. In: Tan, Y., Y. Shi and Z. Ji (Eds.), ICSI, 2012, Part II, LNCS 7332, Springer-Verlag, Berlin, Heidelberg, pp: 93-99.

Xu, S., S. Yang and F.C.M. Lau, 2010. Keyword extraction and headline generation using novel word features. Proceeding of the 24th AAAI Conference on Artificial Intelligence (AAAI-10), pp: 1461-1466.