Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW PAPER
Available Online at www.ijarcs.info

Document representation techniques and their effect on the document Clustering and Classification: A Review

Ksh. Nareshkumar Singh
Dept. Of Computer Science
Manipur University

H. Mamata Devi
Dept. Of Computer Science
Manipur University

Anjana Kakoti Mahanta
Dept. of Computer Science
Gauhati University

***Abstract*:** Text data is the most common form of storing information. When engine search an query, user obtained the large collection of text data. All this retrieve text data are not relevant to the required information. So, it needs to organise the massive amount of text data. Analysing and processing the text data is mainly considered in text mining. Text mining uses the standard data mining methods- classification and clustering. These two methods are used to arrange the documents which are usually represented by hundreds or thousands of texts (words) data. Text data in the document can be represented in various representation methods. In this paper, we have presented a study of various research paper that explore the area of text mining including different document representation methods and their impact on clustering and classification results.

Keywords: Text mining, Document representation, Clustering, Classification.

I. INTRODUCTION

Over the past few years, a great astonishing progress of computer technology has provided more reasonable and powerful computers. The huge amount of data is increasing day by day so we need to maintain and analyse the data for efficient use or processing. Data can be in the form of image, text, spatial form etc. Among this the most natural common data is text data. Text data can be represented in multiple ways - string, words, syntactic structures, entity-relation graphs, predicates, etc. It is used in day today life for example reading the newspaper, postings and messages on social media and even 80% of the company's information is contained in text documents. So, there is great significance in text mining processing and even higher commercial potential than the data mining.

Text mining has been defined as "the discovery by computer of new, previously unknown, information by automatically extracting information from different written resources"[7]. Text Mining also refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents [9]. Now how it differs from the data mining. Text mining is considered as sub-speciality of the broader domain of Knowledge Discovery from Data (KDD)[15]. Basically data mining techniques are used to extract information from structured databases while text mining techniques are used to extract information from unstructured textual data [16]. The major task of text mining includes text classification, text clustering, text categorisation,

sentiment analysis, document summarization, production of granular taxonomies, etc. In this paper we consider only text clustering and text classification because they play important role in organizing the massive amount of document data i.e. text data.

Text document classification is the important task in the text processing field and it is the process of assigning the class labels to the unknown documents based on the model generated in the training phase. Text document clustering is defined as the collection into groups such that documents within the group are similar to each other and dissimilar to those in other groups.

Similarity measures between the objects play a significant role in classification and clustering. Similarity measure is a real valued function that measures the similarity between two objects. Different similarity measures are used to perform text documents clustering and classification. So, choosing good similarity measure is also one of the important steps in the text mining. At the end of the introduction, we can say that text document clustering and classification methods generally depend on the document representation, similarity measure and applying algorithm.

The rest of the paper is organised as follows. Section II describes the document representation methods. In Section III, we present the literature review of various research papers. In Section IV, we give the conclusion.

II. DOCUMENT REPRESENTATION METHODS

Most currently used methods of document representation are Vector Space Model (VSM), Probabilistic Topic Model and Statistical Language Model. In VSM, documents are represented as vectors in an n- dimensional space, where n is the number of unique terms in the vocabulary. The weight value of each term can be computed by different weighted schemes namely Boolean value, Term Frequency (TF), Inverse Document Frequency (IDF), Term Frequency and Inverse Document Frequency (TF*IDF). Most popular and widely use TF*IDF scheme to compute the weight value of each term. The document doc_i can be described as $[W_{i1},W_{i2},.....W_{ij},....W_{in}]$, where W_{ij} is TF*IDF value of j^{th} term in the n-dimensional vector space. This document representation doesn't consider the semantic relations of terms. That is, terms represented in the vector space are

assumed to be mutually independent. Although, it is widely used due to its easy to calculate the similarity between documents, simple and useful for describing document features. In probabilistic topic model, documents are represented as a mixture of topics, where a topic is a probability distribution over words. This model overcomes the problems associated with term as topic. A term can be a word or a phrase. It cannot represent the complicated topics, capture the variations of vocabulary and word sense ambiguity. Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation are the two well known topic modelling methods.

A statistical language model is a probability distribution over sequences of words. A large variety of language technology application needs statistical language model. These include speech recognition, machine translation, document classification and routing, optical character recognition, information retrieval, handwriting recognition, spelling correction, and many more. One of the most successful and popular statistical language model technique is n-gram language models. The n-gram language model assumed that the probability of a word depends on the previous n words. That is, the probability $P(w_{1,}w_{2,}w_{3,}.....w_m)$ of observing the sentence $w_{1,}$ w_2, ...w_m is approximated as

$$P(w_1,w_2, ...w_m)= \prod_{i=1}^{m}(w_i|w_1, \dots \dots \dots \dots \ w_{i-1}).$$

Ah-Hwee Tan[8], presented the text mining framework as consisting of two components namely Text refining and Knowledge distillation. Text refining transforms the unstructured document into an intermediate form and Knowledge distillation deduces pattern or knowledge from the intermediate form. And this paper also illustrated the text mining products and application based on the text refining and knowledge distillation functions.

Tommy W.S. Chow et al.[4], proposed a new document representation with vectorized multiple features including term frequency and term-connection-frequency. Most of techniques are largely based on the term frequency information of "bag of words" model but this model ignores all the semantic or conceptual information. For example two documents containing similar term frequencies may be contextually different when distribution of terms are very different i.e. school, computer and science means very different when they appear in different part of document as compare to they appear together (school of computer science). So "bag of words" model using term frequency information is not the effective way to account the contextual similarity that includes the word inter-connections and spatial distribution of words throughout the document. In this paper introduce graphs for document representation to overcome the word inter-connection problem. The contribution of this paper are:

i) propose a new composite vector for representing a document combined with term frequency and term-connection-frequency extracted from graph.

ii) extract vectorized graph connectionist which facilitate the matching of complex graph especially when consider the large datasets.

The experimental results show that vectorized multiple features which are extracted from different graphs of document representation are able to improve the retrieval accuracy.

Zakaria Elberrichi and Karima Abidi [5], studied on the categorization of the Arabic texts which are represented in three different mode of representations. They are

i) bag of words representation : to transform the texts into vectors of words

ii) N-grams representation : an N-gram is a sequence of N characters just like moving a window of N boxes on the text body.

iii) Concepts representation : vector representation contains only the concepts associated with the words of the text represented, or it contains both concepts and the words.

From the experimental result, it is clear that representation based on concept is the better way of representing the Arabic texts than the other two representation.

Anisha Mariam Thomas and Resmipriya M G[1], proposed a text classification method which is performed by using semi-supervised clustering. Semi-supervised clustering means both labelled and unlabelled data are using for doing clustering. Clustering makes use the labelled text to create the silhouettes (outlines) of text cluster and the unlabelled texts are used to capture the centroids of the text components. Each text cluster category is labeled by the label of text in it and the category of new incoming unlabelled text, measure its similarity with the centroids of the text clusters and give its label with that of the nearest text cluster. In this paper the similarity measures they use are Euclidean distance, Cosine similarity measure, SMPT(Similarity Measure for Text Processing) and Dice coefficient. And they calculate the accuracy value of correctly classified documents.

$$\text{Classification Accuracy} = \frac{\text{No.of correctly classified documents}}{\text{Total no. of documents considered}}$$

When Euclidean measure is used as the similarity measure for classification of given 100 documents as input, 88 documents are correctly classified so its accuracy is 0.88. Similarly Cosine similarity, Dice coefficient and SMPT gives an accuracy of 0.89, 0.81, 0.93 respectively. From the experimental result it is clear that SMPT gives better performance.

Milios et al.[2], studied three well known dimension reduction techniques on three different document representation methods. The three dimensionality reduction methods are

- Independent Component Analysis(ICA)
- Latent Semantic Indexing(LSI)
- Document Frequency(DF)

These three methods are applied on the three document representation methods based on the idea of Vector space model namely word, term and N-Gram representation. Traditional representation of documents is word representation and it is also called as "bag-of-words" or "set of words" . In this representation, every document is considered as a vector where its components are represented by the term in the document collection. Next document representation is term representation method. In this representation, the multi-word terms or sometimes called phrases can be used as features in document vectors. N-Gram representation is the language independent text representation technique. Each document is represented as

feature vector where each feature corresponds to a contiguous substring. Above mentioned dimension reduction techniques are applied on the five different dataset(namely Classics, NG, RD256, RD512 and URCS) represented in above three text representation techniques in order to compare their performance. From experimental results, ICA demonstrates good performance and more stability than LSI in most of the dataset. Both ICA and LSI can effectively reduce the dimensionality from thousands to the range of 10-100. DF can get close to best performance of two other methods at very high dimension but at lower dimension its performance is very lower than two others. In general they made the rank of these three dimension reduction techniques in the order of ICA > LSI > DF. And experiments also show that the word representation method gives better clustering results as compared to two other methods namely term and N-Gram representation.

S M Ruger and S E Gauch [3], discussed feature reduction techniques and their application to the document clustering and classification and also showed that feature reduction improves the efficiency as well as accuracy. They examined the effect of the amount of training data on the classification accuracy. The experimental result show that classification accuracy using vector space approach are more than the Naive –Bayes approach. They also investigated the effect of using only subset of words rather than using the complete test documents as input to the vector space classifier. The words are weighted by using tf.idf scheme and selected the top weighted words of the document which can convey the enough information for classification. They concluded the power of dimensionality reduction in two ways:

i) Best result was found with only six training pages, adding more pages did not improve the classification
ii) Representing pages to be classified by more than the 40 most important words (features) did not improve the classification

They suggested the ranking of the important words in the document. Suppose D be the set of documents and H be the subset that is matched by a particular query. The weight of word j is

$$w_j=(h_j/d_j).h_j \log(|H|/h_j)$$

where h_j is the number of documents in H containing the word j, and d_j is the number of documents in the whole document collection D containing j. Each matched document i is represented as a k-dimensional vector v_i, where the j-th component v_{ij} is a function of the number of occurrences t_{ij} of the j-th ranked related word in the document i:

$$v_{ij}= \log(1+t_{ij}).\log(|D|/d_j)$$

This is the variation of tf-idf weight that stresses the term frequency less.

The reduced feature representations reduce the computational time and also show a better discriminatory behaviour.

Shutian Ma el at.[6], focused on document representation methods and their effects on the quality of clustering results. They compared four different types of document representation:

i) Vector Space Model (VSM) : in this model, documents are represented as vectors. Terms represented in the vector space are assumed to be mutually independent and the weight of each term is computed by using tf*idf (term frequency and inverse document frequency) scheme.

ii) Latent Semantic Indexing (LSI) : In the VSM, document representation method doesn't give the semantic relations of term. This LSI method overcomes the limitation of VSM. LSI is the approach that use the particular mathematical technique called singular value decomposition (SVD) with a raw term by document matrix to get reduced document matrix under certain single value.

iii) Latent Dirichlet Allocation (LDA) : LDA is a generative probabilistic model. It represents the documents as a random mixtures of topics over the latent topic space, where each topic is characterized by a distribution over words.

iv) Doc2vec (Para2Vec): it represents the large blocks of texts such as sentences, paragraphs even the entire documents. It consists of three main stage. First stage is unsupervised training to get word vectors that is same with Word2Vec. Second stage is to get paragraph vector. Third stage is paragraph vector to make a prediction about some particular labels.

From the experimental results they found that document representation method should be chosen according to the corpora characteristics and size to get the better clustering results.

Sunita Sankar el at.[10], presented the comparative analysis of three different clustering algorithm namely K-means, Particle Swarm Optimization (PSO), and hybrid PSO+K-means for clustering the text documents (in Nepali language) which are represented in terms of synsets corresponding to a word. Basically the quality of clustering mainly depends on two factors:

--text document representation
--select of suitable clustering algorithm

They made the effort to improve the quality of clustering by representing the text using the semantics relation. In this study, they measured the quality of clustering according to the two criteria:- to maximize the intra-cluster similarity and to minimize the similarity between the clusters(Inter-cluster similarity). From the experimental results, hybrid PSO+K-means gave the best quality of clustering according to the above two criteria than the separate individual algorithms.

Wen Zhang et al.[11], studied the comparision of TF*IDF, LSI and multi-word methods for text representation and examined their performance of information retrieval and text categorization on the Chinese and English document collection. This paper discussed indexing and term weighting as two components of text representation scheme but here, not considered their effectiveness individually. They mentioned that the basic criterion of text representation is the semantic and statistical qualities which are the two basic properties of the indexing term.

The experimental results show that in text categorization, LSI method gave the best performance than other two methods. LSI method also produced best performance in English information retrieval, but in case of Chinese information retrieval, TF*IDF gave the best performance than two other methods. As an overall outcome they conclude that LSI method was favourable for both semantic and statistical quality.

F. Amin et al. [12], presented three different type of constraints namely Instance level constraint, Corpus level constraint and Cluster level constraint. They used the

instance level constraint as the prior knowledge to propose a new constraint HAC (Hierarchical Agglomerative Clustering) algorithm. This instance level constraint consists of two kinds of constraints – Must-link(ML) and Cannot-link(CL). They used the approach of Wang et al.[17] that provided the graph based document representation in which documents were represented as a dependency graph. Then, they extended this graph based document representation with constraints to improve the quality of document clustering.

The experimental results show that the quality of document clustering using graph based document representation with constraints gave better results than graph based document representation without constraints when the results were measured by the standard cluster quality measure like F-Measure, Purity and Entropy.

Fengxi Song et al.[13], considered mainly the text representation factors namely "stop words removal", "word stemming", "indexing", "weighting" and "normalization" and the effectiveness of these factors to text classifier. Here they used linear SVM (Support Vector Machines) as a text classifier in all the experiment. At the end, they concluded following points from the experimental results:

i) there are strong interactions among the above mentioned text representation factors and the best text representation schemes are dependent on the characteristic of corpus.

ii) Among the above five factors, normalization is the most important factor that may affect the text representation greatly.

iii) Removing the stop words from the vocabulary are not harmful if it is not helpful, no matter in the view of classification efficiency or effectiveness

iv) No definite conclusion can be drawn about word stemming as it is helpful in some dataset while in some other dataset it is harmful. But it can reduce the dimensionality of text feature space greatly, so it can be used in the efficiency of text classification system.

Wen Zhang et al. [14], concerned about the effectiveness of using multi-words for text representation on the performance of text classification. They developed two strategies based on the different semantic level of the multi-words for multi –words representation. First is the decomposition strategy using general concepts for representation and second is the combination strategy using subtopics of general concepts for representation. They carried out text classification experiment on the Reuters-21578 documents using the representations with multi-words. And they also analysed the effect of the linear kernel and non-linear polynomial kernel in support vector machine (SVM) on classification performance. Finally, they also compared the classification performance on the effect of applying different representation strategies and applying the different SVM kernels.

The experimental results show that in multi-word representation, subtopic of general concepts representation perform better than the general concept representation and linear kernel outperforms the non linear kernel of SVM in classifying the Reuters data. The effect of applying different representation strategies is greater than the effect of applying different SVM kernels on the classification performance. In conclusion, they mentioned about the benefit of using multi-words representation:

--lower dimensionality than individual words

--it includes more semantics and large meaningful units than individual word

--multi-word is easy to acquire from the documents.

A.K. Abdulsahib et al.[17], proposed a graph based text representation method, namely dependency graph, in order to reduce sparsity and semantic problem in the textual document. The dependency graph representation scheme is created through the accumulation of syntactic and semantic analysis with the aim of improving the accuracy of document clustering. The basic idea is to convert documents to its syntactic structure. At the semantic level, sentence structure is represented by dependency graphs. Each document is represented as a graph with words as nodes and relationship as edges. They made the comparison of the clustering result obtained by the dependency graph representation method with other popular text representation methods, namely TF-IDF and Ontology based text representation.

After documents represented as dependency graphs, text clustering is performed with the K-means algorithm and with cosine similarity as a similarity measure. They evaluated the performance of different text representation methods based on four measures, namely, precision, recall, F measure, and accuracy. The results shown that proposed text representation method leads to more accurate document clustering results than TF-IDF and Ontology based text representation.

IV. CONCLUSION

In this paper, we have observed the effect on document classification and clustering results due to the different document representation methods. We could mention two other factors that also effect on the result of document clustering and classification are i) similarity measure, for measuring the pair-wise of similarity of document and ii) their applying clustering/classification algorithms. Document representation gives more impact on the classification and clustering results because it captures the semantics of document and also contribute to reduce the problem of high dimensionality. No single document representation method can be recommended as a general method for any application. Different representation methods perform differently depending on collection of data. At the end, what we could conclude is that document representation method should be chosen according to the characteristics of corpora.

V. REFERENCES

[1] Anisha M Thomas and Resmipriya, "An efficient text classification scheme using clustering," Procedia Technology 24, pp.1220-1225, October 2016 [International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST-2015)].

[2] Evangelos E. Milios, M. Mahdi Shafiei, Singer Wang, Roger Zhang, Bin Tang and Jane, " A systematic study on document representation and dimensionality reduction for text clustering," Technical Report CS-2006-05, Faculty of Computer Science, Dalhousie University, July 11,2006.

[3] S M Ruger and S E Gauch, " Feature reduction for document clustering and classification," [kmi.open.ac.uk] / [www.doc.ic.ac.uk], November 2000.

[4] Tommy W.S. Chow, Haijun Zhang and M.K.M. Rahman, "A new document representation using term frequency and vectorized graph connectionists with application to document retrieval," Expert Systems with Applications, Vol. 36, Issue 10, December 2009, Pages 12023-12035.

[5] Zakaria Elberrichi and Karima Abidi, "Arabic text categorization: a comparative study of different representation modes," The International Arab Journal of Information Technology, Vol.9, No. 5, September 2012.

[6] Shutian Ma, Chengzhi Zhang and Daqing He, "Document representation methods for clustering bilingual documents," Proceedings of the 79th ASIST&T Annual Meeting:Creating Knowledge, Enhancing Lives through Information & Technology, Article No. 65, Copenhagen, Denmark-October 14-18, 2016.

[7] Hearst, M. What is Text Mining?; people.ischool.berkeley.edu/~hearst/text-mining.html, October 17, 2003.

[8] Ah-Hwee Tan, "Text Mining: The state of the art and the challenges," www.ntu.edu.sg, In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases.

[9] Feldman, R. & Dagan, I. "Knowledge discovery in textual databases (KDT)," In proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21,1995, AAAI Press, p112-117.

[10] Sunita Sarkar, Arindam Roy and B.S. Purkayastha, "A comparative analysis of Particle Swarm Optimization and K-means algorithm for text clustering using Nepali Wordnet," International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014.

[11] Wen Zhang, Taketoshi Yoshida and Xijin Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification", Expert Systems with Application, Vol. 38, Issue 3, March 2011, Pages 2758-2765.

[12] F. Amin, M. Rafi and M. Shahid, "Document clustering using graph based document representation with constraints," Pak. J. Engg. & Appl. Sci. Vol. 18, January 2016 , p. 56-68.

[13] Fengxi Song, Shuhai Liu and Jingyu Yang, "A comparative study on text representation schemes in text categorization," Pattern Analysis & Applications, Vol. 8, Issue 1, September 2005, Springer-Verlag London, UK, Pages 199-209.

[14] Wen Zhang, Taketoshi Yoshida and Xijin Tang, "Text classification based on multi-word with support vector machine",Knowledge-Based Systems , Vol. 21, Issue 8, December 2008, Pages 879-886.

[15] Elizabeth D. Liddy, "Text Mining," http://surface.syr.edu/cnlp/9, Center for Natural Language Processing, Paper 9, Oct/Nov 2000.

[16] M. Rajman and R. Besancon, "Text Mining : Natural Language Techniques and Text Mining applications " S.Spaccapietra & F.Maryanski(Eds.) , IFIP -The International Federation for Information Processing 1998, Published by Chapman and Hall, Data Mining and Reverse Engineering pp 50-64 .

[17] A.K. Abdulsahib and S.S. Kamaruddin, "Graph based text representation for document clustering," Journal of Theoretical and Applied Information Technology, Vol.76. No.1, 10th June, 2015.

