ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Relevant Feature Discovery and Document Clustering Using Text Mining-A Survey

S.Nithya[1], N.Kamal Raj[2]

M.Phil Scholar, Dept. of Computer Science, Dr .SNS Rajalakshmi College of Arts & Science, Coimbatore, India [1].

Head of the Department, Dept. of Information Technology, Dr .SNS Rajalakshmi College of Arts & Science, Coimbatore, India [2]

ABSTRACT:Clustering is the unsupervised learning process, which is used for data analysis and document classification. And the document clustering is the method of grouping all relevant text documents based on its similarity level. As like clustering, classification is the supervised learning process. The classification involves in data classification based on complete training samples. The main use these two techniques in text document are to enhance and empower the document management and document searching process. It also plays an important role in information retrieval and extracting useful information from huge amount of documents. There are n numbers of algorithms and techniques are available for document clustering and classification. This survey explores the popular document clustering as well as the classification techniques along with its pros and cons.

KEYWORDS: Text Mining, Clustering, Pattern discovery, Text classification.

I. INTRODUCTION

In the recent scenario all domains and departments are rehabilitated into digital. Maximum all documents are converted into electronic format. For example, the students from an institution can refer their academic books online, several institutions providing digital libraries with more and more features. Due to this huge digitalization, documents counts are huge and became unmanageable. To manage such documents, data mining algorithms are used. The most popular techniques are clustering and classification. These techniques are used to summarize, search and manage text documents effectively [1].

Text documents are semi structured in nature, i.e., it is neither completely un-structured nor completely structured. A document may include a few structured topics, such as title, abstract and index terms domain etc., and it also contains some huge unstructured textual topics, such as introduction and contents. In recent research on text document management, several studies have been done to build and develop semi-structured data. To handle the un-structured documents, text indexing and IR (Information Retrieval) techniques have been developed. But those IR and other traditional techniques are not sufficient to handle vast amount of data [2]. This paper discusses the various techniques and tools have been used for document management in the recent years and finally provide an outline about this proposed work. This survey also handles the important process of document clustering, which is known as feature discovery process. This relevance feature discovery process helps to identify the useful features available in the text documents at the time of training [3].

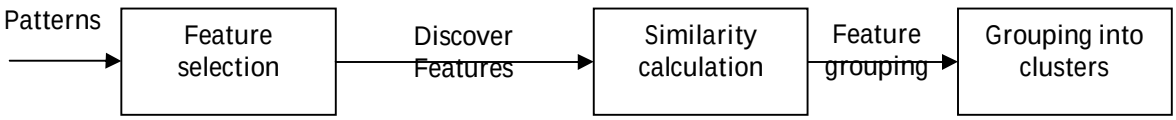

Fig 1.0 Feature discovery and document grouping process

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

The above fig 1.0 represents the basic process involved in document clustering. In order to group the document based on the features, this is important to find the relevant and appropriate features. The above process shows the feature selection and feature based document similarity calculation and based on the similarity score, the documents are grouped together.

II.DOCUMENT CLUSTERING AND CLASSIFICATION TECHNIQUES

In this chapter, we provide a detailed study about the traditional document clustering, classification and feature discovery techniques.

A. Document clustering algorithms:

Document clustering is the process of segmenting text documents into different groups based on its similarity level. The clustering is the unsupervised learning process, where it won't need any training samples for the grouping process [4]. The followings are the popular clustering algorithms are used for document management.

i. Hierarchical clustering Algorithms:

The hierarchical clustering algorithm is a group of data objects forming a tree shaped structure. It can be generally classified into two categories which are as follows [5].

- Agglomerative clustering and
- Divisive clustering.

The agglomerative approach is also called as the bottom up approach. In this approach each data point is considered to be a separate group, and after single iteration the groups are clustered again. In document clustering agglomerative hierarchical clustering plays a vital role. Many authors [6][7]used agglomerative hierarchical clustering for document clustering process. In the divisive clustering all data points are considered as a single group, and they are separate into a number of clusters. This separation is based on certain criteria. This divisive clustering approach is also called as the top down approach.

ii. K-Means Clustering:

Another popular text clustering algorithm is k-means clustering. This algorithm is known to be efficient in clustering large data sets. It is one of the simplest and the best known unsupervised learning algorithms that solve the well known clustering problem. But several researches [8] [9] proves the hierarchical clustering is better than K-means clustering.

iii. **Expectation Maximization** (EM):

The EM algorithm is another popular approach for text document clustering. An EM algorithm is used to find maximum likelihood estimates (MLE) of parameters in probabilistic models, where the model depends on unobserved hidden variables [10]. Clustering text documents requires different features and representations of documents. These features are analyzed and filtered before applying to the clustering process. The Set of features can set of words, word similarity, number of sentences, etc

B. Document classification techniques:

The document classification is a supervised learning process, which needs a complete and semi training process.

i. SVM (Support Vector Machine)

SVM is commonly used to segment a individual data input from a set of documents into two distinct groups. And SVM exploits a supervised learning method, which means it learns to classify unseen data based on a set of labeled

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

training data. The initial set of training data is typically identified by domain experts and is used to build a model that can be applied to any other data outside the training set. The effort required to construct a high quality training set is quite modest, particularly while measure up to the volume of data that may be eventually classified next to the training set. It defines that the learning algorithms such as SVM offer an exceptionally cost effective method of text classification for the massive volumes of documents produced by modern organizations [11].

ii. Vector Space Document (VSD) model:

The VSD represents the text documents as vectors of identifiers. It is widely used in information retrieval and document similarity matching. Using VSD, the sentences are extracted and finally converted into the summary of the document. This is a widely used popular similarity calculation model in document clustering method. Finally this performs the ranking process to filter the undesired documents and features [12].

iii. KNN (K-Nearest Neighbor):

KNN is another popular algorithm for document classification. It is a non parametric based method, which has k-closest neighbors. Some authors in the literature have proposed the combination of KNN algorithm with TF-IDF method and framed a new text classification approach. This approach increases the speed and quality of document classification [13]

III.LITERATURE REVIEW

Chim, Hung, and Xiaotie Deng [14] proposed phrase-based document similarity, which helps to compute the pair wise relationship between multiple text documents. The work by the authors is based on the STD (Suffix Tree Document model). In general, phrase in a documents have been considered as a more informative feature. And it improves the effectiveness of text document clustering. The STD model is a phrase based approach, which inherits the tf-idf weighting scheme. The authors have applied group average Hierarchical agglomerative clustering algorithm to develop a new clustering algorithm. The experiments are conducted in RCV1 and corpora datasets. The authors proved that the phrase based document similarity works better than the single word tf-idf measure. The new phrase based document similarity successfully connects the two document models and inherits their advantages. The concept of STD is simple and effective, but the implementation is become very difficult and tough. The STD structure is used n-gram method to identify and extract phrases from the documents.

Lan, Man, et al[15] proposed supervised term weighting method, i.e., tf:rf, to improve the terms' discriminating power for text categorization task. This paper utilizes vector space model for text representation, this transforms the content of a text document into a vector. In this study, the authors investigated numerous unsupervised and supervised term weighting methods with SVM and kNN algorithms. Finally the paper shows the supervised term weighting methods are good in performance. The implementation of the tf:rf in text summarization and IR are left for future work. And the proposal has been experimented in a sample synthetic dataset and the proposal is not efficient for huge dataset like RCV1 benchmark corpus. This doesn't have the ability to handle huge text collections.

Shehata, Shady, FakhriKarray, and Mohamed Kamel[16] proposed a new concept-based mining model, which analyzes terms on the sentences. This work bridges the gap between NLP and text mining regulations. A new concept based mining model composed of four components. The authors proposed a method to enhance the text-clustering eminence by exploiting the semantic structure of the sentences in documents. A better text clustering result is achieved by using the above concept based mining model. The first one is sentence based concept analysis, document based concept analysis, corpus level concept analysis and finally concept based similarity detection. The authors used tf method for all the above analysis. This has the ability to calculate pairwise document similarity accurately. It is very robust and accurate. However, the work affected by several issues, such as this work only handles homogeneous text documents and tough for real time implementation.

Jiang, Jung-Yi, Ren-JiaLiou, and Shie-Jue Lee [17] proposed a fuzzy similarity-based self-constructing algorithm for feature clustering, because the feature clustering is the most powerful tool for text classification. Based on the words, the documents are grouped in the same cluster. The proposed FFC (Fuzzy Feature Clustering) is an incremental clustering approach to reduce the dimensionality of the features in text classification. Here statistical mean deviation has been used. While grouping the texts based on the words, outliers are considered as new clusters. The similarity

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

measure is calculated using the variance. The authors show the FFC is a feature reduction technique, which facilitates the fast clustering process. The FFC method is only good for text categorization problems due to the suitability of the distributional word clustering concept.

Zhang, Taiping, et al[18] proposed a new spectral clustering method called CPI, which is abbreviated as correlation preserving indexing. This CPI method is performed in the correlation similarity measure space. The authors are much concentrated on the intrinsic geometrical structure of the document. The work is not sufficient to apply on the huge datasets.

Later **Zhuang, Fuzhen, et al[19]** developed a two phase cross domain method for text classification. Particularly, a CD-PLSA (Collaborative Dual Probabilistic Latent Semantic Analysis) model is first learned to effectively capture the distinction and commonality from corner to corner numerous domains in a collaborative nature. In that case, the authors further mined the intrinsic composition of desired domains by refining the outputs from CD-PLSA, which also called RCD-PLSA. The work by [19] is based on EM (expectation Maximization) algorithm, and provided various training samples for almost all domains. But the paper suffers from accuracy issues.

Skabar, Andrew, and Khaled Abdalgader [20] presented a novel fuzzy clustering algorithm that operates on relational input data. The relational input data are such as in the form of a square matrix of pair-wise similarities between data documents. The proposed algorithms utilize the graph representation and operate in the EM algorithms as like the previous paper. The likelihood function is created using the graph centrality. In the paper, the concepts present in natural language documents (NLD) usually display some type of hierarchical structures, while the algorithm presented in this paper identifies only flat clusters rather than the hierarchical. So it doesn't supports hierarchical structure.

Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee [21] proposed a new similarity measure algorithm for text classification and clustering. This takes many cases for similarity calculation, which are features from both documents, features from a single document and features are not in the given documents. The authors generated the awareness of detecting presence and absence of features, features have non-zero values. This has been applied in hierarchical clustering and KNN clustering algorithms. But the proposed work has been investigated only few clustering algorithm and doesn't provide accuracy in similarity finding.

Li, Zechao, et al[22] developed a novel unsupervised feature selection algorithm, named as clustering guided sparse structural learning (CGSSL). This integrates the cluster analysis and sparse structural analysis as a joint framework. The authors used the nonnegative spectral clustering for accurate cluster label detection. The cluster labels are predicted using non-negative analysis.

VI. CONCLUSION

With the use of data mining approach, data management becomes easier and convenient. Due to digital document process, the size of documents is very huge and very tough to manage. In such environment, the effective and fast grouping is more important because the data should retrieve quickly and effectively. In this survey, the different document clustering techniques and algorithms are discussed. This survey gives the overall summary of the review by different metrics and parameters. In this survey, various techniques in text mining for document management is discussed, this paper shows the pros and cons of several traditional feature discovery algorithms based on different techniques.

However, the techniques almost concentrated on general text mining process, where the document clustering needs additional concentration and work to improve the following problem. The first problem is discovering appropriate features with less effort and validation on discovered features are not yet studied. And there is a need for a new system to handle the above problem in document management and information retrieval.

REFERENCES

[1] Murua, A., et al. "Model based document classification and clustering."*Manuscript in preparation* (2001).
[2] B. Croft, D. Metzler, and T. Strohman, Search Engines: Information Retrieval in Practice. Reading, MA, USA: Addison-Wesley, 2009.
[3] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.
[4] Neto, Joel Larocca, et al. "Document clustering and text summarization." (2000).
[5] Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." *KDD workshop on text mining*. Vol. 400. No. 1. 2000.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

[6] Voorhees, Ellen M. "Implementing agglomerative hierarchic clustering algorithms for use in document retrieval." *Information Processing & Management* 22.6 (1986): 465-476.
[7] Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003.
[8] Richard C. Dubes and Anil K. Jain, *Algorithms for Clustering Data*, Prentice Hall, 1988.
[9] Paul Bradley and Usama Fayyad, *Refining Initial Points for K-Means Clustering*, Proceedings of the Fifteenth International Conference on Machine Learning ICML98, Pages 91-99. Morgan Kaufmann, San Francisco, 1998.
[10] Jajoo, Pankaj. *Document clustering*. Diss. Indian Institute of Technology Kharagpur, 2008.
[11] Yu, Chun-Nam John, and Thorsten Joachims. "Learning structural svms with latent variables." *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009.
[12] Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." *Communications of the ACM* 18.11 (1975): 613-620.
[13]Trstenjak, Bruno, SasaMikac, and DzenanaDonko. "KNN with TF-IDF based Framework for Text Categorization." *Procedia Engineering* 69 (2014): 1356-1364.
[14] Chim, Hung, and Xiaotie Deng. "Efficient phrase-based document similarity for clustering." *IEEE Transactions on Knowledge and Data Engineering* 20.9 (2008): 1217-1229.
[15] Lan, Man, et al. "Supervised and traditional term weighting methods for automatic text categorization." *IEEE transactions on pattern analysis and machine intelligence* 31.4 (2009): 721-735.
[16] Shehata, Shady, FakhriKarray, and Mohamed Kamel. "An efficient concept-based mining model for enhancing text clustering." *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010): 1360-1371.
[17] Jiang, Jung-Yi, Ren-JiaLiou, and Shie-Jue Lee. "A fuzzy self-constructing feature clustering algorithm for text classification." *IEEE transactions on knowledge and data engineering* 23.3 (2011): 335-349.
[18] Zhang, Taiping, et al. "Document clustering in correlation similarity measure space." *IEEE Transactions on Knowledge and Data Engineering* 24.6 (2012): 1002-1013.
[19] Zhuang, Fuzhen, et al. "Mining distinction and commonality across multiple domains using generative model for text classification." *IEEE Transactions on Knowledge and Data Engineering* 24.11 (2012): 2025-2039.
[20] Skabar, Andrew, and Khaled Abdalgader. "Clustering sentence-level text using a novel fuzzy relational clustering algorithm." *IEEE transactions on knowledge and data engineering* 25.1 (2013): 62-75.
[21] Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee. "A similarity measure for text classification and clustering." *IEEE transactions on knowledge and data engineering* 26.7 (2014): 1575-1590.
[22] Li, Zechao, et al. "Clustering-guided sparse structural learning for unsupervised feature selection." *IEEE Transactions on Knowledge and Data Engineering* 26.9 (2014): 2138-2150.

.