IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue, 5, March 2019 E-ISSN: 2347-2693

Survey on Feature Selection Techniques towards Text Mining in Cloud

Balavinothini[1*] , Gnanambigai [2]

[1]Research Scholar, Bharathiar University, Coimbatore, India
[2]Department of Computer Science, Indira Gandhi Arts and Science College, Puducherry, India

Corresponding author: balavinothini@yahoo.co.in

DOI: https://doi.org/10.26438/ijcse/v7si5.101105 | Available online at: www.ijcseonline.org

Abstract**—** Cloud computing is a technology that provides efficient services to the users over internet. Users stores volumes of data in cloud which is rendered as data as a service (DaaS) on demand and charged as per usage. Text mining is a technology that is used to retrieve data from a massive set of database. Cloud uses Text mining to retrieve data efficiently from various cloud data centres. Text classification is a technique used for discovering classes of indefinite data. Prior to applying any mining technique, trivial features should be filtered. Feature selection is capable of improving learning process, lesser computational complexity, organizes better general models, and decreasing required storage. We analyses towards effectiveness of the clustering based feature selection method. This paper is to analysis on different techniques used for feature selection. Further survey on Feature selection and Feature extraction technique has been extract the features from the documents, which results in single and multi-label document classification. Based on the extracted features the survey is done on multiple-feature based projective nonnegative matrix factorization technique to cluster the documents.

Keywords**—** Data mining, Feature selection, Text mining, Filtering, factorization

I. INTRODUCTION

Cloud computing has emerged as one of the real world technology that are hugely in use in recent decades. It provides resources and applications on-demand as a pay-per-use service over internet. The advent of the big data storage possibility over the cloud has created a large scope of cloud data storage and cloud data management features. Huge size data stored in the cloud can be categorized into structured, semi-structured, unstructured and heterogeneous, which is the combination of any of the above types of data. Documents are the text data, which is a subcategory of structured data, which can be are stored in cloud database. Users can also access this cloud database, which is referred as a Database-as-a service. One of the major concern for the users who access the cloud database is the relevancy problem and high document retrieval time.

In order to solve this issue, the documents are clustered before uploading into the cloud database based on their respective categories [3]. Though document clustering is possible, still there is no possible Standard Document clustering strategy for cloud storage and hence it has been emerging as a one of the interesting research areas in current trends. Document clustering can be described as a method of organizing huge volume of text documents into sub categories based on the topic about which it speaks which paves way to content retrieval and summarization of data. This document clustering technique can lead to further deeper studies like topic detection and tracking problems. Feature selection improve the processing learning in a lesser computational complexity and organizes a better general models, and decreasing required storage [8]. Here the survey analyse the effectiveness of the clustering method and different techniques used on Feature selection and Feature extraction technique. this has been extract the features from the documents, which results in single and multi-label document classification on multiple-feature based projective nonnegative matrix factorization technique [5]. Section I contains the introduction of the cloud to create a large scope of cloud data storage and cloud data management features, Section II contain the related work of Feature Selection Techniques towards Text Mining, Section III contain the some measures of Different Clustering Techniques, Section IV contain the architecture and essential steps of Document Clustering Framework structure technique section V explain the concludes research work with future directions. Section VI describes results and discussion.

II. RELATED WORK

Document clustering is possible, still there is no possible Standard Document clustering strategy for cloud storage and hence it has been emerging as a one of the interesting research areas in current trends. Based on the extracted features the survey is done on multiple-feature based projective nonnegative matrix factorization technique to cluster the documents. Document clustering can be described as a method of organizing huge volume of text documents into sub categories based on the topic about which it speaks which paves way to content retrieval and summarization of data.

Table 1 Analysis of Different Clustering Techniques

S.N	Algorithm / Methodology	Advantages	Limitation /Proposed Idea
1	Domain-specific ontology , Combination with the vector space model (VSM) with singular value decomposition (SVD) and Fuzzy equivalence relation.[1]	1. Reduces the dimensionality of the original data and considers the correlation between the terms. 2. Encoding the ontologies in the aggregation process provides better clustering results. 3. This work is applied to food safety supervision which is beneficial for government and society	1. One limitation of our fuzzy clustering method is that its performance depends on the comprehensiveness of the ontology used, for obtaining better results in other specific domains, more extensive ontologies should be incorporated. 2. Natural language processing methods can be improved for feature identification. 3. Addition of visualization techniques alongside our fuzzy clustering method can further aid user navigation of document collections
2	CDIM, an algorithmic framework for partitional clustering of documents that maximizes the sum of the discrimination information provided by documents.[2]	1. Produces high-quality clusters and an advantage of using a measure of discrimination information is that it quantifies the degree of relatedness of a term to its context in the collection. 2. It produces clusters that are readily interpretable by their highly discriminating terms. 3. The superior understanding provided by CDIM's output is also demonstrated, enabling documents in the clusters to be identifiable as belonging to specific contexts or topics	1. Use of external sources like WordNet and Wikipedia with CDIM, CDIM with intelligent seeding, CDIM using item sets and sequences are also some promising future extensions. 2. In addition to implementing soft-CDIM, the soft clustering version of CDIM, future work can be investigated with other measures of discrimination/relatedness information, extend and evaluate CDIM for constraint based clustering, and co-clustering.
3	Hard Diagonal Double K-means (DDKM) and Fuzzy Diagonal Double Kmeans (F-DDKM).[3]	1. More effective for document term partitioning (both on binary data and with TF-IDF transformation). 2. DDKM requires less time to converge; up to 20 times less time than DKM and 40 times less time than SpCo commonly used in the domain of document clustering	1. For further research, it will be worthwhile to investigate an efficient theoretical way to choose the values of the parameters α and β and to study in more detail their impact on the clustering Performance. 2. The knowledge of the number of co-clusters is mostly required. Another initiative will be to investigate an efficient way to assess this parameter
4	A Frobenius based inner product that allows defining kernel functions for spectral clustering.[4]	1. The results obtained from this method exceeds the results of the tfIdf method and the classic BoW representation in the information retreival task.	1. Variation of kernel functions can be integrate and used with the proposed technique. 2. Different clustering algorithms along with the different parameters can be tested since this work is done with the simple k-means algorithm for better results.
5	Semi-Supervised Concept Factorization (SSCF) algorithm.[5]	1. Effective for complicated document clustering 2. It inherits the strength of CF and avoids its weakness of CF. 3. It can incorporate various forms of background knowledge into clustering to improve the clustering performance	1. This SSCF algorithm is generic and can be applicable to many other research fields such as image data representation, DNA gene expressions and multimedia processing in future.
6	Integrating WordNet with lexical chains to address clustering issues.[6]	1. The combination of explicit and implicit semantic relationships in Word Net pays a positive role to the assessment of word sense similarity. 2. Estimation of the true number of clusters is possible by observing the obtained results, which is valuable for deciding the value of k in K-means clustering algorithms. 3. Lexical chain features (the core semantics) improves the quality significantly with a reduced number of features in the document clustering process.	1. Some important words which are not included in WordNet lexicon will not be considered as concepts for similarity evaluation. 2. The proposed method can obtain better clustering results only if the explicit and implicit relationships between words are thoroughly represented in WordNet. 3. In future ,this method can be performed on a larger knowledge base, such as Wikipedia
7	CROCS framework for unsupervised CCR(Co-reference Resolution).[7]	1. Computational cost reduction. 2. Avoids model-selection parameters	1. Feature generation from multiple Knowledge Base's and catering to streaming scenarios (e.g., news feeds or social media) are directions of future work.
8	Big text document clustering algorithm using terms of class label and semantic feature based on Hadoop framework.[8]	1. Cost effective and It clusters the big data size of document using the distributed parallel processing based on Hadoop. 2. Normalized mutual information value is higher when implementing this algorithm and compared with the other document clustering algorithms	1. The semantic feature extractions are limited when the word semantics are not available in the dictionary. 2. Big data clustering still have some issue in scheduling of the parallel processing techniques

`

Lin Yue et al [1] had developed new method Do-main-specific ontology with the combination of vector space model (VSM) and singular value decomposition (SVD). This method reduces the dimensionality of the orig-inal data and considers the correlation between the terms. Encoding the ontologies in the aggregation process provides better clustering results. But the performance of this method depends on the comprehensiveness of the ontology used, for obtaining better results in other specific domains, more extensive ontologies should be incorporated and can further support user navigation of document collections

In CDIM method [2] Malik Tahir et al introduce a high-quality cluster using a measure of discrimination in-formation is that it quantifies the degree of relatedness of a term to its context in the collection. But the use of external sources like WordNet and Wikipedia with CDIM, CDIM with intelligent seeding, CDIM using item sets and se-quences are need to extended.

Charlotte Laclau and Mohamed Nadif [3] developed DDKM and F-DDKM methods for more effective for document term partitioning (both on binary data and with TF-IDF transformation). It is not worthwhile to investigate an efficient theoretical way to choose the values of the parameters α and β and so it is needed to study in more detail their impact on the clustering Performance and the knowledge of the number of co-clusters is mostly required. Another initiative will be to investigate an efficient way to assess this parameter

Vıctor Mijangos et al [4] explain about frobenius based inner product that allows defining kernel functions for spectral clustering and the results obtained from this method exceed the results of the TFIDF method and the classic BoW representation in the information retrieval task. Different clustering algorithms along with the different parameters need to be tested since this work is done with the simple k-means algorithm for better results

Mei Lua et al [5] had introduced an effective for complicated document clustering Semi-Supervised Concept Factorization (SSCF) algorithm which inherits the strength of CF and avoids its weakness of CF. This SSCF algorithm is generic, which can be applicable to many other research fields such as image data representation, DNA gene expres-sions and multimedia processing in future.

Tingting Wei et al [6] explain about Integrating WordNet with lexical chains to address clustering issues. The combination of explicit and implicit semantic relation-ships in WordNet pays a positive role to the assessment of word sense similarity. Estimation of the true number of clusters is possible by observing the obtained results, which is valuable for deciding the value of k in K-means cluster-ing algorithms. Lexical chain features (the core semantics) improves the quality significantly with a reduced number of features in the document clustering process. Some im-portant words which are not included in WordNet lexicon will not be considered as concepts for similarity evaluation. This method can obtain better clustering results only if the explicit and implicit relationships between words are thor-oughly represented in WordNet.

Sourav Dutta and Gerhard Weikum [7] had intro-duced CROCS framework for unsupervised CCR(Co-reference Resolution) which reduce computational cost and also avoids model-selection parameters. Feature generation from multiple Knowledge Base’s and catering to streaming scenarios (e.g., news feeds or social media) are need to analysis.

Yoo-Kang et al [8] explains about Big text docu-ment clustering algorithm using terms of class label and semantic feature based on Hadoop framework which are cost effective and it is clusters the big data size of document using the distributed parallel processing based on Hadoop. The semantic feature extractions are limited when the word semantics are not available in the dictionary. Big data clustering still have some issue in scheduling of the parallel processing techniques.

III. ANALYSIS OF DIFFERENT CLUSTERING TECHNIQUES

The traditional document clustering method used Bag of words model, which extracted the features, for term frequency and weight computation, which lacked in semantic relationship computation. Distance based measures are computed to cluster the documents, but this technique also had issues since, distance, as a single parameter cannot determine the exact relationship between the documents. To overcome the issue in traditional model, Internal and external knowledge based document clustering is introduced. Internal knowledge based clustering uses factorization techniques but they lack in construction of semantic features, on the other hand, external knowledge based clustering constructs term ontologies but it also ends up in some amount of information loss due to the difficulty in locating the comprehensive ontology that covers all the concepts given in the documents. It also takes high cost of construction of ontologies.

Domain-specific ontology method

Domain-specific ontology method is a combination with the vector space model (VSM) with singular value decomposition (SVD) and Fuzzy equivalence relation. [1] This method reduces the dimensionality of the original data and considers the correlation between the terms. Encoding the ontologies in the aggregation process provides better clustering results. This work is applied to food safety supervision which is beneficial for government and society. One limitation of fuzzy clustering method is that its performance depends on the comprehensiveness of the ontology used, for obtaining better results in other specific domains, more extensive ontologies should be incorporated.[1] Natural language processing methods can be improved for feature identification and in addition of visualization techniques alongside the fuzzy clustering method can further aid user navigation of document collections.

Clustering by Discrimination Information Maximization method

CDIM is an algorithmic framework for partitioned clustering of documents that maximizes the sum of the discrimination information provided by documents. This method produces high-quality clusters. The advantage of using a measure of discrimination information is that it quantifies the degree of relatedness of a term to its context in the collection. It produces clusters that are readily interpretable by their highly discriminating terms. The superior understanding provided by CDIM's output is also demonstrated, enabling documents in the clusters to be identifiable as belonging to specific contexts or topics. Use of external sources like WordNet and Wikipedia with CDIM, CDIM with intelligent seeding, CDIM using item sets and sequences are also some promising future extensions. In addition to implementing soft-CDIM, the soft clustering version of CDIM, future work can be investigated with other measures of discrimination/relatedness information, extend and evaluate CDIM for constraint based clustering, and co-clustering.[2]

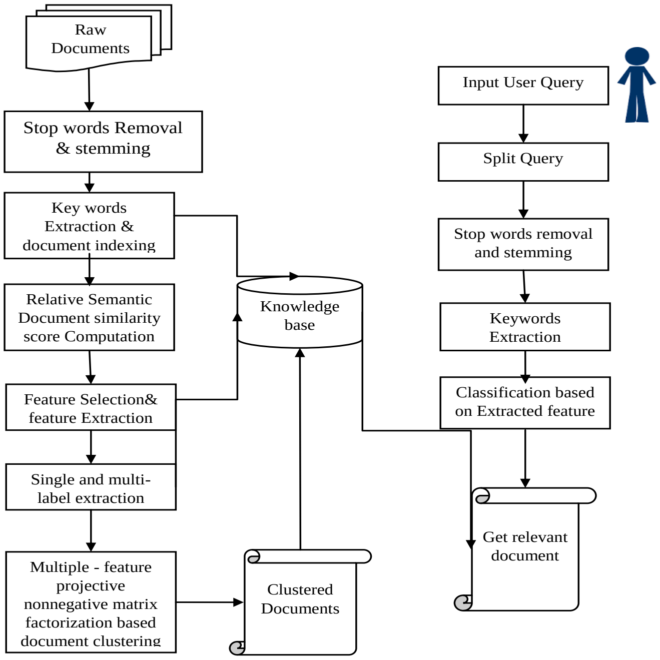
Raw Documents
Stop words Removal & stemming
Key words Extraction & document indexing
Relative Semantic Document similarity score Computation
Feature Selection& feature Extraction
Single and multi-label extraction
Multiple - feature projective nonnegative matrix factorization based document clustering
Knowledge base
Clustered Documents
Input User Query
Split Query
Stop words removal and stemming
Keywords Extraction
Classification based on Extracted feature
Get relevant document

Figure 1. Document Clustering Framework structur

IV. Document Clustering Framework structure

To overcome the above mentioned issue, a Novel Document Clustering Framework has been proposed that will outperform comparatively from the existing techniques. Initially the documents to clustered and uploaded to the cloud storage are preprocessed. Some of the preliminary preprocessing techniques are stop words removal, stemming, keywords extraction, indexing and etc. The preprocessed documents are processed to compute the similarity. A novel relative semantic similarity score computation methods are survived towards to compute a similarity between the documents based on the distance additionally considering the Natural language processing techniques which computes the relative semantic similarity score.

Further the survey is based on Feature selection and Feature extraction technique to extract the features from the documents, which results in single and multi-label document classification. Based on the extracted features a novel multiple-feature based projective nonnegative matrix factorization technique to cluster the documents. On the request of user query, the cluster documents can be uploaded to the cloud storage by the user- friendly relevant document retrieval model.

V. Conclusion and Future Scope

In this paper, the survey provides efficient services to the users over internet for cloud users to stores huge volumes of data in cloud. Many text mining, Text classification and Feature selection techniques are used to retrieve data efficiently from various cloud data centres. Feature selection improves learning process in lesser computational complexity and decreases required storage in cloud. This survey analyses the effectiveness of the clustering based feature selection method. Further survey on Feature selection and Feature extraction technique has been extract the features from the documents, which results in single and multi-label document classification on multiple-feature based projective nonnegative matrix factorization technique to cluster the documents from cloud environment.

References

[1] Lin Yue, Wanli Zuo , TaoPeng , YingWang, Xuming Han A fuzzy document clustering approach based on domain-specified ontology, " Data & Knowledge Engineering", 100 (2015) 148-166.

[2] Malik Tahir Hassana, Asim Karim, Jeong-Bae Kim, Moongu Jeon CDIM: Document Clustering by Discrimination Information Maximization, "Information Sciences", 316 (2015) 87–106.

[3] Charlotte Laclau, Mohamed Nadif "Hard and fuzzy diagonal co-clustering for document-term partitioning "Neurocomputing", 193 (2016) 133–147

[4] Vıctor Mijangos, Gerardo Sierra, Azucena Montes Sentence level matrix representation for document spectral clustering "Pattern Recognition Letters, Elsevier", 20 November 2016.

[5] Mei Lua, Xiang-Jun Zhao, Li Zhang, Fan-Zhang Li, Semi-supervised concept factorization for document clustering, “Information Sciences”, 331 (2016) 86–98.

[6] Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, Xianyu Bao, A semantic approach for text clustering using WordNet and lexical chains, “Expert Systems with Applications”, 42 (2015) 2264–2275.

[7] Sourav Dutta, Gerhard Weikum, Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment, “Transactions of the Association for Computational Linguistics”, 3(2015)15–28.

[8] Yong-Il Kim, Yoo-Kang Jiand Sun Park, Big Text Data Clustering using Class Labels and Semantic Feature Based on Hadoop of Cloud Computing, “International Journal of Software Engineering and Its Applications”, 8(2014),1-10.

Authors Profile

Mrs. B.Balavinothini pursed Bachelor of Science from BharathidasanUniversity, Trichy in 2004 and Master of Information Technology in Bharathidasan University, Trichy in 2006. She is currently pursuing Ph.D.in Bharathiar University in Coimbatore since 2012. She has 8 years of teaching experience and 6 years of Research Experience.

Dr. N. Gnanambigai received Ph.D in Computer Science, M.Phil in Computer Science and Master of Computer Science. She is having 17 Years of Experience in Teaching and working as Assistant Pro-fessor in Computer Science, Indira Gandhi college of Arts and Science College, Puducherry, India. Her research interests include Artificial Neural Networks, Distributed Systems, networking and Cloud Computing.