ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 9, September 2013

INCREMENTAL TEXT MINING USING MODIFIED FUZZY BASED NAVIE BAYESIAN CONCEPT FOR SEMANTIC REPRESENTATION

[1]Dr. M. Thangamani, [2]A. Kalayanasaravanan, [3]Dr. E.T. Venkatesh & [4]Dr. P. Thangaraj
Assistant professor, Department of Computer Science and Engineering,
Kongu Engineering College, perundurai, Erode-638 052, Tamilnadu, India,

Abstract

Document clustering techniques are used to group up the documents with reference to the similarity. It is widely used in web mining and digital library environment. Documents are represented in vector space model. Each document is a vector in the word space and each element of the vector indicates the frequency of the corresponding word in the document. Documents are presented as high dimensional data elements. It is very complex task to cluster documents using K-means clustering algorithm. The sub space clustering schemes can be adopted to cluster documents. The document clustering uses the term weights from the similarity measure. The sub space model uses the relevant attributes for the similarity estimation. The fuzzy logic is used to cluster the documents. Semantic analysis is carried out with the support of the ontology. The ontology is used to maintain term relationships. Term relationships are represented using the synonym, meronym and hypernym factors. Ontology is manually collected by the users. Domain based ontology is used for the document clustering process. The system uses the data mining domain based ontology for the semantic analysis. Semantic weights are used in the similarity measure. Fuzzy based text document clustering scheme uses the stop word filters and stemming process under the document preprocess. Term clustering and semantic clustering operations are performed in the system. The system is designed as graphical user interface based stand alone application. The memory requirement for the term clusters and semantic clusters are analyzed. The process time indicates the time taken to finish the clustering process.. These three factors are used to decide the efficiency of the clustering schemes. The document repository for the system is constructed with 1000 documents.. The java language and oracle back end are used for the system development.

Index Terms— Text Clustering, Co-occurrence Probability, Naive Bayesian Concept, Cluster Conditional Independence, Fuzzy Clustering

1. Introduction

Document clustering has been studied intensively because of its wide applicability in areas such as web mining and information retrieval. In document clustering, unlabeled documents are typically represented in vector space model (VSM), where each document is a vector in the word space and each element of the vector indicates the frequency of the corresponding word in the document. Generally, the data are of very high dimensional and sparse, which poses a big challenge to conventional clustering algorithms such as *k*-means.

In high dimensional data, clusters often exist in subspaces rather than in the entire space. For example, in document clustering, clusters of documents of different topics are categorized by different subsets of keywords. Moreover, the keywords for one cluster may not occur in the documents of other clusters. One solution to this problem is text subspace clustering, which aims to discovering the document clusters in different subspaces of the original word space. In the past few years, soft subspace clustering algorithms have been developed and successfully applied to clustering large document collections. Examples include LAC, FWKM, and EWKM etc. In these algorithms, each term is assigned with a desired set of weighting values to distinguish its different contributions to document categories. Since the weighting values are ranged between 0 and 1, the subspaces discovered by these algorithms are of soft. With *k*-means type methods the algorithms iteratively group the documents into *hard* partitions.

In many applications, a document may include multiple topics and thus may relate to multiple categories at the same time, resulting in the requirement of *fuzzy* document clustering. On the other hand, due to its effectiveness in discovering clusters with overlapping boundaries, fuzzy clustering algorithms are able to reveal more accurate cluster structures within the document collections. In a feature weighting algorithm combined with the fuzzy *k* prototypes algorithm was presented. The steps of feature weighting and data partitioning are separated in this algorithm. Recently, an algorithm named fuzzy W-*k*-means was proposed. In this algorithm however, the dimensions are assigned with

a uniform value for different clusters. Additionally, the fuzzy W-*k*-means introduces two user defined parameters α and β, which are difficult to estimate in practice.

2. Related works

In many applications, a document may include multiple topics and thus may relate to multiple categories at thesame time, resulting in the requirement of fuzzy document clustering. On the other hand, due to its effectivenessin discovering clusters with overlapping boundaries, fuzzy clustering algorithms are able to reveal more accurate cluster structures within the document collections (Q.Wang, Y.Ye, and J.Z.Huang. 2006). In (J.Li, X.Gao, andL.Jiao. 2005), a feature-weighting algorithm combined with the fuzzy k prototypes algorithm was presented. Thesteps of feature weighting and data partitioning are separated in this algorithm. Recently, an algorithm namedfuzzy W-k-means (Q.Wang, Y.Ye, and J.Z.Huang. 2006) was proposed. In this algorithm however, thedimensions are assigned with a uniform value for different clusters. Additionally, the fuzzy W-k-means (Q.Wang,Y.Ye, and J.Z.Huang. 2006) introduces two user defined parameters α and β, which are difficult to estimate inpractice. In order to perform fuzzy clustering on high dimensional data, a new algorithm named FPC (FuzzyProjected Clustering) was studied in our previous work (L.Chen, Q.Jiang, and S.Wang. 2008). This parameter free algorithm can generate "soft" partitions of the high dimensional data. (M. Thangamani and P. Thangaraj, 2010, 2012 & 2013) proposed d document clustering for individual and distributed environment to achieve semantic clustering.

Statement of the Problem:

The Existing document clustering techniques uses the term weights. Stop word elimination is applied to reduce the vector size. The term frequency is used for the term weight estimation process. But, vector space model is used for the data clustering process. Term relationship is not used in the similarity measurement. Vector size is high. Process time is high and Cluster accuracy is low.

3. Proposed System

The proposed system is designed to perform the document clustering using the semantic analysis mechanism. The ontology is used for semantic analysis. The fuzzy logic technique is used for the clustering process. The fitness analysis is performed to verify cluster accuracy. The sub space clustering scheme is used in the system. The document attributes are collected and grouped with relevancy. The similarity measurement is estimated on the sub space model. The sub space similarity model reduces the computation complexity and increases the accuracy. The sub space model also reduces the process time.

4. System Methodology

Currently we are using TF-IDF matrix. We present the algorithm to build the co-occurrence matrix from the TF-IDF matrix in [2]. In this algorithm, we study the co-occurrences between terms. When two terms co-occur in a document, we take the minimum of the number of their occurrences as a co-occurrence measure. Terms which are linked semantically will be grouped under one cluster. We assume that terms which have a high degree of co-occurrence are likely to be linked semantically. In our work, we assume that one term may belong to only cluster. We uniquely assign a term to a single cluster. This is done by the application of conditional probability and the naïve Bayesian concept. We calculate the conditional probabilities of a term belonging to each of the possible clusters and assign it to the cluster with the highest probability. From the co-occurrence matrix obtained, we come to know which terms co-occur. Initially, each term is treated as a cluster centre and all terms cooccurring with this term are put into the cluster corresponding to this term. Terms which do not co-occur with any other term are the singular terms in their respective clusters.

The system is divided into four major modules.

Document Preprocess

The documents are maintained in text file format. The contents of the documents are parsed and converted into the vector space model. The stop word elimination and stemming process are used to reduce the vector size. The system maintains a stop word repository. The stop words in the documents are removed using the repository. The stemming process analyzes the suffix value for the terms. The base term is extracted using the stemming process. The porter-stemming algorithm is used in the system. The document details are updated into the database. The system also updates the term list into the database.

Term Cluster

The system performs two types of clustering operations. They are term clustering and the semantic clustering. The term clustering task is

performed using the term weights. The term frequency is estimated and updated into the database. The term frequency and inverse document frequency are calculated for each term. The term weights are used for the similarity measurement process. The fuzzy clustering scheme is applied on the sub space of the term collection. The term weights are used for the comparison process. The term cluster requires high vector size for the clustering process.

Semantic Cluster

The semantic clustering is performed with the term relationship based comparison. The term cluster does not consider the term relationship. The semantic cluster uses the term relationship for the clustering process. The ontology is used to maintain the relationship for the term collection in a domain. The terms are maintained with synonym, meronym andhypernym relationships. The term category is used for the weight estimation process.

5. Result

The below figure show Term Cluster vs. Semantic Cluster.Semantic Cluster performance the better result than term cluster. The Performance analysis gives the result of the Fuzzy Term Cube with the Fuzzy Semantic Cube. For every increase in transaction data, the fitness of the Fuzzy Semantic Cube is higher than the Fuzzy Term Cube.

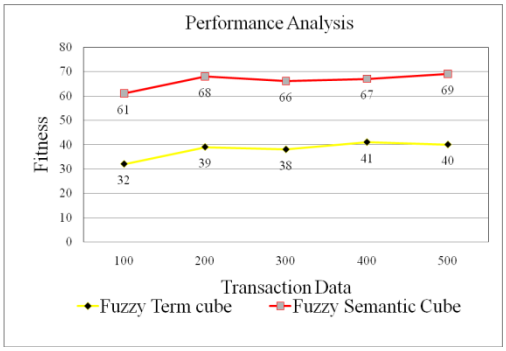

Figure 5.1 Term cube Vs. Semantic Cube

6. Conclusion and future Enhancement

The data clustering techniques are applied on the structured databases. The documents are unstructured databases. The document clustering is a complex task. The document clustering requires a preprocessing task to convert the unstructured data values into a structured one. The documents are large dimensional data elements. The dimension is reduced using the stop word elimination and stemming process. The clustering process is applied on the preprocessed document collection. Term weights are estimated using the term frequency values. The ontology is used for semantic weight estimation process. Concept relationship is considered in the semantic analysis. The fuzzy logic technique is used to optimize the weight levels between 0 to 1 boundaries. It is applied to improve the clustering accuracy. The fuzzy document clustering uses the sub space-clustering model. The relevant attributes are used for the comparison process. The semantic analysis is used to reduce the vector size. The system is tested with different cluster counts and document count. The cube size and vector size are analyzed for different document count. The memory, process time and fitness levels are compared in the performance analysis. The accuracy level is analyzed with fitness point values. The fuzzy based semantic clustering scheme improves the clustering accuracy. The system presents feasible solution for document clustering requirement.

The fuzzy document-clustering scheme is enhanced with semantic analysis mechanism to improve cluster accuracy. The system reduces the process time in a considerable manner. The benchmark datasets are used in the system. Data mining domain-based ontology is used in the system. Different parameters are used in the performance analysis. Process time and memory parameters areused to measure the scalability effects. The fitness point is used to measure the accuracy values. The system can be enhanced with the following features.

- ✓ The fuzzy based document clustering system can be enhanced to perform text categorization and text summarization tasks.
- ✓ The single domain ontology model can be enhanced with multi domain ontology to cluster documents in all disciplines.
- ✓ The system can be enhanced to perform document clustering under distributed environment.
- ✓ The clustering scheme can upgrade to cluster multilingual documents.
- ✓ The system uses the text documents as the input source. In the future the system can be improved to cluster other document forms such as web documents, rich text format (RTF) and portable document format (PDF).
- ✓ The clustering scheme can be integrated with rule mining techniques to mine association between the documents.
- ✓ The text document-clustering scheme can be improved to cluster text and image contents.

References

C.Domeniconi, D.Gunopulos, and S.Ma. (2006). "Locally adaptive metrics for clustering high dimensional data,"
Technical Report ISE-TR-06-04, George Mason University, 2006.

H.Sun, S.Wang, and Q.Jiang.(2004). "Fcm-based model selection algorithms for determining the number ofclusters," Pattern Recognition, vol. 37(10), pp. 2027–2037, 2004.

J.Li, X.Gao, and L.Jiao. (2005). "A novel feature weighted fuzzy clustering algorithm," LNAI, vol. 3641, pp.412–420, 2005.

J.Z.Huang, M.K.Ng, H.Rong, and Z.Li. (2005). "Automated variable weighting in k-means type clustering,"
IEEE Transactions on Knowledge and Data Engineering, vol. 27(5), pp. 657–668, 2005.

L.Chen, Q.Jiang, and S.Wang. (2008). "A probability model for projective clustering on high dimensional data,"
Proceeding of the IEEE ICDM, pp. 755–760, 2008.

L.Chen, Y.Ye, and Q.Jiang. (2008). "A new centroid-based algorithm for text categorization," Proceeding of theAINAW, pp. 1217– 1222, 2008.

L.Jing, M.K.Ng, and J.Z.Huang.(2007) "An entropy weighting kmeans algorithm for subspace clustering ofhigh-dimensinoalsparese data," IEEE Transactions on Knowledge and Data Engineering, vol. 19(8), pp. 1–16.

L.Jing, M.K.Ng, J.Xu, and J.Z.Huang. (2005). "On the performance of feature weighting k-means for textsubspace clustering," Proceeding of the WAIM, pp. 502–512, 2005.

Lifei Chen, Shengrui Wang and Qingshan Jiang.(2009). "A Robust Algorithm for Fuzzy Document Clustering",2009.

Q.Wang, Y.Ye, and J.Z.Huang. (2008). "Fuzzy k-means with variable weighting in high dimensional dataanalysis," Proceeding of the WAIM, pp. 365–372, 2008.

S.B.Kotsiantis and P.E.Pintelas. (2004) "Recent advances in clustering: A brief survey," WSEAS Transactions onInformation Science and Applications, vol. 11(1), pp. 73–81.

Thangamani .M and Thangaraj (2010) P,"Ontology Based Fuzzy Document Clustering Scheme", Modern Applied Science,vol.4(7), pp.148-153.

Thangamani .M and Thangaraj (2010) .P,"Survey on Text Document Clustering",International Journal of Computer Science and Information Security,vol.8(4).

Thangamani, M. and Thangaraj, P, (2010), "Integrated Clustering and Feature Selection Scheme for Text Documents", International Journal of Computer Science, Vol.6, Issue 5, pp.536-541.

Thangamani.M and Thangaraj.P, (2012) "Effective fuzzy semantic clustering scheme for decentralized network through multidomain ontology model", International Journal of Metadata, Semantics and Ontologies, Interscience Vol.7, Issue 2, pp.131-139, December 2012 Interscience publication

Thangamani.M and Thangarj.P. (2013) "Fuzzy ontology for document clustering based on genetic Algorithm", International Journal of Applied mathematics and information science, Vol.4, Issue 7, pp.1563-1574.