Journal of Computer Science 6 (5): 536-541, 2010
ISSN 1549-3636

Integrated Clustering and Feature Selection Scheme for Text Documents

[1]M. Thangamani and [2]P. Thangaraj
[1]Department of Computer Technology,
[2]School of Computer Technology and Applications,
Kongu Engineering College, Perundurai-638052, Erode-(DT), Tamilnadu, India

Abstract: Problem statement: Text documents are the unstructured databases that contain raw data collection. The clustering techniques are used group up the text documents with reference to its similarity. **Approach:** The feature selection techniques were used to improve the efficiency and accuracy of clustering process. The feature selection was done by eliminate the redundant and irrelevant items from the text document contents. Statistical methods were used in the text clustering and feature selection algorithm. The cube size is very high and accuracy is low in the term based text clustering and feature selection method. The semantic clustering and feature selection method was proposed to improve the clustering and feature selection mechanism with semantic relations of the text documents. The proposed system was designed to identify the semantic relations using the ontology. The ontology was used to represent the term and concept relationship. **Results:** The synonym, meronym and hypernym relationships were represented in the ontology. The concept weights were estimated with reference to the ontology. The concept weight was used for the clustering process. The system was implemented in two methods. They were term clustering with feature selection and semantic clustering with feature selection. **Conclusion:** The performance analysis was carried out with the term clustering and semantic clustering methods. The accuracy and efficiency factors were analyzed in the performance analysis.

Key words: Clustering, text mining, ontology, feature selection, document clustering.

INTRODUCTION

The term Data Mining generally refers to a process by which accurate and previously unknown information can be extracted from large volumes of data in a form that can be understood, acted upon and used for improving decision processes. Data Mining is most often associated with the broader process of Knowledge Discovery in Databases (KDD), "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data". By analogy, this system defines Textual Data Mining as the process of acquiring valid, potentially useful and ultimately understandable knowledge from large text collections.

How to explore and utilize the huge amount of text documents is a major question in the areas of information retrieval and text mining. Document clustering is one of the most important text mining methods that are developed to help users effectively navigate, summarize and organize text documents (Al-Mubaid and Syed Umair, 2006). By organizing a large amount of documents into a number of meaningful clusters, document clustering can be used to browse a collection of documents or to organize the results returned by a search engine in response to a user's query. It can significantly improve the precision and recall in information retrieval systems and it is an efficient way to find the nearest neighbors of a document.

Statement of problem: The problem of document clustering is generally defined as follows: Given a set of documents, would like to partition them into a predetermined or an automatically derived number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. In other words, the documents in one cluster share the same topic and the documents in different clusters represent different topics.

In most existing document clustering algorithms, documents are represented using the vector space model which treats a document as a bag of words. A major characteristic of this representation is the high

Corresponding Author: M. Thangamani, Department of Computer Technology, Kongu Engineering College, Perundurai-638052, Erode-(DT), Tamilnadu, India

dimensionality of the feature space, which imposes a big challenge to the performance of clustering algorithms. They could not study efficiently in high-dimensional feature spaces due to the inherent sparseness of the data. Another problem is that not all features are important for document clustering. Some of the features may be redundant or irrelevant. Some may even misguide the clustering result, especially when there are more irrelevant features than relevant ones. In such case, selecting a subset of original features often leads to better clustering performance. Feature selection not only reduces the high dimensionality of the feature space, but also provides better data understanding, which improves the clustering result (Sebastiani, 2002). The selected feature set should contain sufficient or more reliable information about the original data set. For document clustering, this will be formulated into the problem of identifying the most informative words within a set of documents for clustering.

MATERIALS AND METHODS

Feature selection has been widely used in supervised learning, such as text classification. It is reported that feature selection can improve the efficiency and accuracy of text classification algorithms by removing redundant and irrelevant terms from the corpus. Traditional feature selection methods for classification are either supervised or unsupervised, depending on whether the class label information is required for each document. Those unsupervised feature selection methods, such as the ones using document frequency and Term Strength (TS), can be easily applied to clustering. However, supervised feature selection methods using the information gain and the X^2 statistic can improve the clustering performance better than unsupervised methods when the class labels of documents are available for the feature selection. However, supervised feature selection methods cannot be directly applied to document clustering because, usually, the required class label information is not available. The Iterative Feature selection (IF) method is proposed, which utilizes the supervised feature selection to iteratively select features and perform text clustering. In many previous text mining and information retrieval research, the X^2 term-category independence test has been widely used for the feature selection in a separate preprocessing step before text categorization. By ranking their X^2 statistic values, features that have strong dependency on the categories can be selected and this method is denoted as CHI. Two variants of the X^2 statistic have been proposed recently (Liu, 2003). Correlation coefficient is proposed, which could be viewed as a "one-sided" X^2 statistic. Galavotti *et al*. (2000) went further in this direction and proposed a simplified variant of the X^2 statistic, which was called GSS coefficient. Feature selection methods based on these two variants of the X^2 statistic were tested on improving the performance of text categorization.

RESULTS AND DISCUSSION

TCFS algorithm: In most existing text clustering algorithms, text documents are represented by using the vector space model. In this model, each document is considered as a vector in the term-space and is represented by the following Term Frequency (TF) vector:

$$dtf = [t_{f1}, t_{f2}, \ldots, t_{fh}] \quad (1)$$

Where:
t_{fi} = The frequency of the i^{th} term in the document
h = The dimension of the text database

which is the total number of unique terms. Normally, there are several preprocessing steps, including the stop words removal and the stemming, on the documents. A widely used refinement to this model is to weight each term based on its Inverse Document Frequency (IDF) in the corpus. To account for the documents of different lengths, the length of each document vector is normalized to a unit length. Normalized vector space model is weighted by TF-IDF is used to represent documents during the clustering.

For the problem of clustering text documents, there are different criterion functions available (Ienco and Meo, 2002). The most commonly used is the cosine function. The cosine function measures the similarity between two documents as the correlation between the document vectors representing them. For two documents di and dj, the similarity between them can be calculated as:

$$Cosine\ (di, dj) = di*dj/||di|||dj|| \quad (2)$$

Where:
X = The vector dot product
|di| = The length of vector di

The cosine value is 1 when two documents are identical and 0 if there is nothing in common between them. The larger cosine value indicates that these two documents share more terms and are more similar.

The K-means algorithm is very popular for solving the problem of clustering a data set into k clusters. If the data set contains n documents, d_1; d_2, d_n, then the clustering is the optimization process of grouping them into k clusters so that the global criterion function is:

$$\sum_{j=1}^{k}\sum_{i=1}^{n} f(d2, cenj) \quad (3)$$

either minimized or maximized. Centroid represents the centroid of a cluster cj, for j = 1, ..., k and f (di, cenj) is the clustering criterion function for a document di and a centroidcenj. When the cosine function is used, each document is assigned to the cluster with the most similar centroid and the global criterion function is maximized as a result. This optimization process is known as an NP complete problem and the K-means algorithm was proposed to provide an approximate solution. The steps of K-means are given as follows:

1. Select k initial cluster centroids
2. For each document of the whole data set, compute the clustering criterion function with each cluster centroid Assign each document to its best choice
3. Recalculate k centroids based on the documents assigned to them
4. Repeat step 2 and 3 until convergence

Ontology based term analysis: There are several approaches for machine-processing information such as the statistical approaches and the knowledge representation approaches. The major difference between the two approaches concerns their handling of incoming or already existing data. The former approach is free of any structural information (Wenliang *et al*., 2004). Systems that are developed according to this approach, such as those that involve machine-learning methods, analyze a given piece of data and try to learn rules out of the regularities in it. As a next step the learned rules are applied to the new incoming data for processing. The latter on the other hand does make use of structural representation. The systems that base on this approach have an internal representational structure of how the data to be processed ought to look like. The incoming data will be then processed according to this representational structure. Therefore, the second approach implies knowing beforehand, whereas the first one implies learning.

An ontology is a "a specification of a conceptualization" whereby a conceptualization is a collection of objects, concepts and other entities that are presumed to exist in some domain and that are tied together with some relationships. A conceptualization is a simplified view of the world, a way of thinking about some domain.

Ontologies belong to the knowledge representation approaches that have been discussed above and they aim to provide a shared understanding of a domain both for the computers and for the humans. Thereby, ontology describes a domain of interest in such a formal way that it can be processed by computers. The outcome is that the computer system knows about this domain (Forman, 2005). Ontology is a formal classification schema, which has a hierarchical order and which is related to some domain. Ontology comprises the logical component of a "Knowledge Base". Typically, a knowledge base consists of ontology, some data and also an inference mechanism. Ontology, comprising the logical component of the knowledge base, defines rules that formally describe how the field of interest looks like (Lindeberg, 2004). The data can be any data related to this field of interest that is extracted from various resources such as databases, document collections and the Web. The inference mechanism would deploy rules in form of axioms, restrictions, logical consequences and other various methods based on the formal definition in the ontology over the actual data to produce more information out of the existing one.

Semantic based TCFS: The web documents are composed using HTML files with textual contents and tag elements. The data mining applications are applied to extract knowledge from the web contents. The web contents are passed into the data cleaning operation before the mining process. The textual web content refers the cleaned web documents. The web documents with out HTML tag elements are denoted as textual web contents. The textual web contents are maintained in text document formats. This system is designed to perform clustering process and feature selection on text documents.

The semantic clustering and feature selection on textual web contents system is designed with three major phases. They are document preprocessing, document clustering and clustering with feature selection process. The document preprocessing is the initial phase for the system. Stop words elimination and stemming process are carried out in the document preprocessing phase (Liu and Yu, 2005). Term weight and semantic weight estimation are also done at the document preprocessing phase. The document clustering process is applied to group the documents with relationship to the document contents. The semantic relations are identified in the rule mining process on

clustered documents. The feature selection process is applied on each text documents under each cluster.

The semantic clustering and feature selection system is designed as a graphical user interface tool. The system is divided into four major modules. They are Text documents, document analysis and document clustering and feature selection. The text documents module is designed to maintain the text documents and ontology. The document analysis module is designed to perform document preprocessing and weight estimation process. The term cluster and semantic cluster are constructed in the document-clustering module. The term features and semantic features are extracted from the clustered documents under the feature selection module.

The system is developed using the Java language and oracle database. The user interface forms are designed using swing package. The modules and sub modules in the system are connected with the graphical menu items. The modules are designed in dialog boxes. The graphical menu options are designed using the Ulead Photoimpact graphical designing software. All the sub menus and sub modules are connected with the main menu form. The text documents are maintained in the documents folder under the application environment.

Text documents: The text documents module is the initial module in the system. The text documents module is designed to maintain the text documents and ontology. The text documents are collected from the IEEE web site. Data mining domain related journal collection is downloaded from the web. The journal abstract page is designed using HTML. The HTML pages are downloaded and converted into text documents. The text document conversion is done by remove the HTML tag elements from the web documents. The text contents are maintained in separate text files.

The text documents module is divided into three sub modules. They are text document list, text document view and ontology view. The text documents list shows the list of text documents that are collected from the web. The document name and size details are displayed in the text document list. The text documents are maintained in the documents folder. The content view sub module is designed to display the contents of the selected text document (Liu *et al*., 2003). The ontology maintains the concepts, synonyms, meronyms and hypernyms collection. The ontology is constructed for different domains. This system uses the ontology for data mining domain. The ontology view shows the ontology contents.

Document analysis: The document analysis module is designed to perform document preprocessing and weigh estimation process. In the document preprocess stop word elimination and stemming process is carried out on the text documents. The stop word elimination is done with a stop word collection. The stemming process is applied using the porter stemming algorithm. The documents contents are reduced in a considerable way. The weight estimation process is done in two methods. They are term weight and semantic weight (Li and Luo, 2008). The term weight is estimated using the Term Frequency (TF) and Inverse Document Frequency (IDF) values. The term and its count are used in the term weight estimation process. The semantic weight estimation is done with the support of the ontology. The concept identification and weight estimation are done in the semantic weight estimation process. The clustering process is done using the term weight and semantic weight. The semantic weight values are used in the rule mining process.

The document analysis module is divided into two sub modules. They are term analysis and semantic analysis. The term analysis is designed to estimate term frequency and integrated document frequency values. The semantic analysis is designed to calculate semantic weights. The term cube is constructed in the term analysis. The semantic cube is constructed in the semantic analysis. The term cube and semantic cube are used in the clustering process. The term analysis and semantic analysis are performed after the document preprocessing.

The term analysis module is divided into three sub modules. They are term extraction process, term weight estimation and term cube construction. The term extraction process is applied to produce the list of term and their count in a document (Martin Law and Mario Figueiredo, 2004). The term weight estimation is done using the term and its count. The terms and its TF/IDF values are produced in this module. The term cube is constructed for entire document collection. All terms and document details are updated into the term cube.

The semantic analysis module is divided into three sub modules. They are concept extraction, concept weight estimation and semantic cube. The concept extraction module is designed to identify concept in each document. This process is done with the help of the ontology collection. The terms are matched with concepts, synonyms, meronyms and hypernyms in the ontology. The concept weight is estimated with the concept and its element count. The concept weight is also called as semantic weight. The semantic cube is constructed with concepts and semantic weight.

Clustering process: The clustering process module is designed to cluster the documents with reference to its relationship. The clustering process groups the documents. The clustering process is divided into two major modules. They are term cluster and semantic cluster (Dash *et al*., 2002). The term cluster module is designed to cluster the document with the term weights. The semantic cluster groups the document with semantic weights. The document cluster details are updated into the database. The cluster process filters irrelevant terms and concepts.

The term cluster module is divided into two sub modules. They are term cluster view and term cluster details. The term cluster view module is designed to initiate the clustering process. The user can select the cluster count and initiate the cluster process. The cluster name and its document count are displayed in the form. The cluster details form shows the list of documents that are arranged in each cluster. The semantic cluster module is divided into two sub modules. They are semantic cluster view and semantic cluster details. The semantic cluster view initiates the clustering process. The cluster details are displayed with the cluster name and its relevant documents. The system also filters the irrelevant concept from the list.

Feature extraction: The feature extraction module is designed to extract frequent and highlighted terms from the text documents. The feature selection process is done for all the documents (Chen *et al*., 2005). The clustering process and feature selection process are done in parallel order. The term feature selection is carried out during the term clustering process. The semantic feature selection is performed during the semantic clustering process. The feature selection results are displayed in separate forms. The term feature selection results are displayed with feature summary and feature list. The feature summary the cluster name and feature count for each cluster is displayed (Das, 2001). The feature list shows the list of features for the selected cluster. In the same way the semantic feature is also produced. The feature analysis module is designed to perform the comparison for the feature selection techniques.

The performance analysis is carried out graphical analysis. The result graph shows that the semantic clusters require minimum cube size than the term clusters. So the memory requirement and process time are reduced in the semantic based analysis.

The text documents are denoted as unstructured databases. It is very complex to group the text documents. The feature selection is the process of extracting the frequent and popular contents of the text document collection. Both the document grouping and feature selection tasks require the content relationship factors. The semantic analysis is the technique that uses the term and its relationship with a collection of terms. The relationships are represented as synonym, meronym and hypernym. The system is implemented to perform text document grouping and feature selection with the support of semantic analysis.

The system is tested with 1000 text documents collected from the IEEE web site. Initially the documents are updated to the database with preprocessed information. The stop word elimination and stemming operations are performed in the preprocess. All the document analysis operations are carried out on the database information. The weight estimation process is done with term analysis and semantic analysis tasks. The system also tested with term clusters and semantic clustering operations. The feature selection is done for both term clustering and semantic clustering operations. The feature analysis is also performed to evaluate the performance of the feature selection operations.

The performance analysis is carried out for different document collections. The cube analysis is conducted to compare the cube size requirement for the term clustering and semantic clustering process. The cube size analysis results are represented in Fig. 1 and Table 1. The results show that the semantic clustering process reduces the cube size in a large way. 90% of term cluster cube size is reduced in the semantic cluster. The feature analysis results are shown in Fig. 2 and Table 2. The results show that the semantic clustering with feature selection reduces the irrelevant feature in a considerable manner.

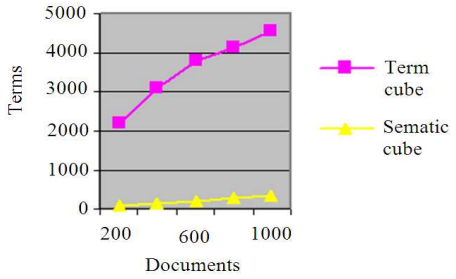

Fig. 1: Analysis on term cube Vs semantic cube

Table 1: Analysis on term cube Vs semantic cube

Documents	Term cube	Semantic cube
200	2176	98
400	3098	137
600	3807	192
800	4152	268
1000	4548	310

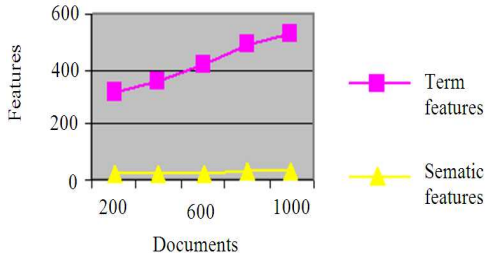

Fig. 2: Analysis on term cube Vs semantic cube

Table 2: Analysis on term cube Vs semantic cube

Documents	Term cube	Semantic cube
200	317	17
400	359	21
600	418	23
800	491	27
1000	524	30

CONCLUSION

The text grouping and feature selection system is developed to carry out the text document grouping and feature selection operations. The text document contents are optimized with semantic analysis. The system is analyzed with 1000 text documents. The data mining domain based ontology is used for the system. The clustering and feature selection is performed for different cluster count. The results show that the proposed scheme reduces the cube size. The feature selection process also produces more accurate results. The text document clustering and feature selection system can be upgraded to process all type of documents such as Rich Text Format (RTF) documents and Portable Document Formats (PDF). The system uses single domain based ontology for data mining domain only. The system can be enhanced with multi domain ontology to analyze documents with any domain.

REFERENCES

Al-Mubaid, H. and A. Syed Umair, 2006. A new text categorization technique using distributional clustering and learning logic. IEEE Trans. Knowl. Data Eng., 14: 1156-1165.

Chen, J., D. Ji and C.L. Tan, 2005. Unsupervised feature selection for relation extraction. Neural Computat., 13: 2573-2593.

Das, S., 2001. Filters, wrappers and a boosting-based hybrid for feature selection. Proceeding of the 18th International Conference on Machine Learning, June 28-July 1, Morgan Kaufmann Publishers Inc., San Francisco, CA., USA., pp: 74-81.

Dash, M., K. Choi, P. Scheuermann and H. Liu, 2002. Feature selection for clustering a filter solution. Proceeding of the International Conference on Data Mining, IEEE Xplore Press, USA., pp: 115-122.

Forman, G., 2005. Feature selection: We've barely scratched the surface. IEEE Syst., 17: 4.

Galavotti, L., F. Sebastiani and M. Simi, 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. Proceedings of 4th European Conference on Research and Advanced Technology for Digital Libraries, Lisbon, Portugal, pp: 415-424. DOI: 10.1007/3-540-45268-0

Ienco, D. and R. Meo, 2002. Exploration and reduction of the feature space by hierarchical clustering. Department of Informatics, University of Torino, Sala Conferenze, Italy.

Li, Y. and C. Luo, 2008. Text clustering with feature selection by using statistical data. J. Cognit. Neurosci., 20: 2125-2136. DOI: 10.1109/TKDE.2007.190740

Lindeberg, T., 2004. Feature detection with automatic scale selection. Int. J. Comput. Vis., 30: 77-116. DOI: 10.1023/A: 1008045108935

Liu, H. and L. Yu, 2005. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng., 17: 491-502.

Liu, H., 2003. Evolving feature selection. IEEE Intell. Syst., 20: 64-76.

Liu, T., S. Liu, Z. Chen and W.Y. Ma, 2003. An evaluation on feature selection for text clustering. Proceeding of the 20th International Conference on Machine Learning, Washington, DC., pp: 415-424. http://www.hpl.hp.com/conferences/icml2003/papers/15.pdf

Martin Law, H.C. and A.T. Mario Figueiredo, 2004. Simultaneous feature selection and clustering using mixture models. IEEE Trans. Patt. Anal. Mach. Intell., 26: 1154-1166.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Computer. Surveys, 34: 1-47.

Wenliang, C., C. Xingzhi and W. Huizhen, 2004. Automatic word clustering for text categorization using global information. Lecturer Notes Comput. Sci., 3411: 1-11.