A New Hybrid Feature Selection using Natural Language Processing for Text Clustering

Mrs. Rashmi G. Dukhi
Mtech(CSE),GHRIETW,Nagpur

Ms. Pratibha Mishra
Assistant Professor,GHRIETW,Nagpur

ABSTRACT

Text clustering is unsupervised machine learning method.It needs representation of objects and similarity measure. which compares distribution of features between objects. For the high dimensionality of feature space performance of clustering algorithms decreases.Two techniques are used to deal with this problem: feature extraction and feature selection.In this paper, we describe the hybrid method used for text clustering which is the combination of active feature selection,genetic algorithm and bisecting K-means.Internal quality measures computes the effectiveness of clustering.Our method is compared with K-means.

Keywords -summarization; unsupervised; similarity measures; classifier.

I. INTRODUCTION

Internet contains vast amount of unstructured text. The unstructured texts contains massive amount of information which cannot be used for further processing by computers. To extract useful patterns from documents,specific processing methods and algorithms need to be used.Huge information lies in collections of documents in the form of digital libraries and repositories, and digitized personal information such as blog articles and emails. Clustering of text documents plays a vital role in efficient Document Organization, Summarization, Topic Extraction and Information Retrieval. Text document clustering groups similar documents that to form a coherent cluster, while documents that are different have separated apart into different clusters. Clustering is used in information retrieval and information extraction, by grouping similar types of information sources together.

NLP tools extracts novel knowledge out of very large unstructured collections of text documents (text data mining). NLP organizes the documents into meaningful groups according to their content and to visualize the collection, providing an overview of the range of documents and of their relationships, so that they can be browsed more easily.Natural language processing approaches can be applied both to feature extraction and feature reduction phases of the text classification process. Linguistic features can be extracted from texts and used as part of their feature vectors. Natural language processing can be used in ways that encompass both feature extraction and reduction, tools can be used to identify keywords from a text document or even create a semistructured summary of the text. Feature extraction from such condensed forms of the original documents reduces the dimensionality of the input vector without reducing the classification performance.

Feature selection is a process that chooses a subset from the
original feature set according to some criterions. The selected
feature retains original physical meaning and provides a better
understanding for the data and learning process.The process of text classification is shown in figure1.

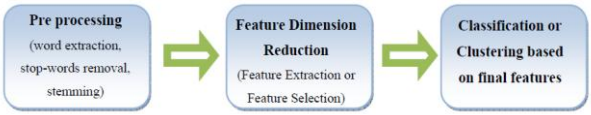

Fig. 1. Steps in Text Classification

The first phase consist of preprocessing the text documents.The second phase of text clustering is feature dimension reduction. Feature extraction and feature selection are two commonly used methods for reducing the dimension of corpus. Feature extraction is the process of extracting new features from the set of all features by means of some functional mapping. Feature selection methods on the other hand select some of the existing terms based on some measures and generate the final feature vector.

Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into Hierarchical and Partitioning methods [2, 3, 4, 5].

Hierarchical clustering method works by grouping data objects into a tree of clusters [6]. These methods can further be classified into agglomerative and divisive
Hierarchical clustering depending on whether the Hierarchical decomposition is formed in a bottom-up or top-down fashion. K-means and its variants [7, 8, 9] are the most well-known partitioning methods [10].
Clustering is a technique which has no predefined class labels but using similarity measures between different objects, it put most similar object in one class and dissimilar in another class. Figure 2 descibes the general steps used in document clustering. Very first words are separated and then weights are applied to each of them.

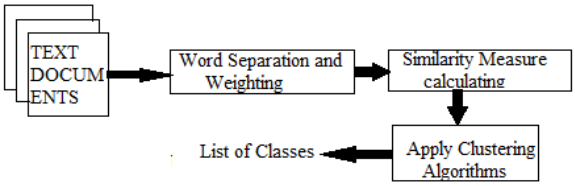

Fig 2. Basic Clustering Steps

II. PROPOSED WORK

After preprocessing of text documents,feature extraction is used to transform the input text documents into a feature set(feature vector).Feature Selection is applied to the feature set to reduce the dimensionality of it. We apply feature selection methods to text clustering task to improve the clustering performance.In this project, we explore the possibility of active feature selection that can influence which instances are used for feature selection by exploiting some characteristics of the data. Our objective is to actively select instances with higher probabilities to be informative in determining feature relevance so as to improve the performance of feature selection without increasing the number of sampled instances. Active sampling used in active feature selection chooses instances in two steps: first, it partitions the data according to some homogeneity criterion; and second, it randomly selects instances from these partitions.In this project, we are applying a combination of NLP, Active feature selection and unsupervised method GA along with clustering thus we would get a better output for text classification with respect to the methods available.We will perform a comparative study on a variety of feature selection methods for text clustering, with other algorithms.Finally, we evaluate the performance of hybrid feature selection(AFS) method based on clustering.

Nature of similarity measure plays a very important role in the success or failure of a clustering method.An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance and Pearson Correlation Coefficient.

A. Preprocessing

The set of documents is given as input to feature extraction which is preprocessed first.Text Preprocessing aims at transforming the text collection into a useful form for the learning algorithms, involving tasks as treatment, cleaning and reduction of the data.

1) Lexical Analysis – Converting byte strings to token.

2) Elimination of Stop Words – Remove stop words like the,and,of,a.

3) Stemming-Remove variant word forms to a single "stem" form like ing,ed,pre,sub.

4) Selection of Index Terms-Term can be individual words or noun phrases.

B. Extraction of feature terms using the NLP

Natural language processing approaches is be applied both to feature extraction and feature reduction phases of the text classification process. Linguistic features can be extracted from texts and used as part of their feature vectors.For example parts of the text that are written in direct speech, use of different types of declinations, length of sentences, proportions of different parts of speech in sentences (such as noun phrases, preposition phrases or verb phrases) can all be detected and used as a feature vector or in addition to word frequency feature vector.

The important features i.e words from the document are extracted using NLP & GA.Parts of Speech(POS) tagging is used for extraction of features.POS tag the document via the standard ngram tagger.It takes a sentence as input,assigns a POS tag to each word in the sentence and produces the tagged text as output.

C. Feature Selection

After extracting features,feature selection is performed and features are classified into clusters based on the similarity.The specific steps of feature selection and clustering are as follows:

Feature selection optimization based on combination GA and bisecting k-means using genetic algorithm to implement global searching, and using bisecting k-means algorithm is used as operator.

The specific algorithm is described as follows:
1) Determine the text features encoding scheme, using binary encoding, the chromosome length is L;
2) Initialization control parameters: N is population size, Pc is crossover probability, P_m is mutation probability, gen is hereditary algebra;
3) Creating an initial population with m individuals;
4) Calculating the fitness of each individual, using Bisecting k-means algorithm to select the parent chromosome of crossover and mutation;
5) In accordance with crossover probability Pc, mutation probability Pm, generate offspring via crossover and mutation on parent population.
6) Repeat steps 4) and 5) , until all the individual no longer changed (or reach the termination conditions).

The Basic Bisecting K-means Algorithm for finding *K* clusters is:
1. Pick a leaf cluster C to split.
2. Use the basic K-means algorithm. (Bisecting step) to split into two sub-clusters C_1 & C_2.
3. Repeat the bisecting step and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is obtained.

III. EVALUATION CRITERIA

Intra-cluster similarity specifies documents within a cluster are similar.Intra-cluster similarity should be high and inter-cluster similarity specifies documents from different clusters are dissimilar.Inter-cluster similarity should be low for efficient clustering. This is an internal criterion for the quality of a clustering. The Clustering results are evaluated using only quantities and features inherited from the dataset.

Internal quality measures: It uses a criterion inherit for the model and/or clustering algorithm. For instance the average similarity of objects in the same cluster.
Let
$C = \{c_i\}$ a clustering with clusters c_i
$K = \{k^{(j)}\}$ a categorization with categories $k^{(j)}$
n number of documents
n_i number of documents in cluster ci
$n^{(j)}$ number of documents in category $k^{(j)}$
$n^{(j)}_i$ number of documents in cluster ci and category $k^{(j)}$
$M = \{n^{(j)}_i\}$ is the The confusion matrix

$sim(c_i , c_i)$ is the cluster intra similarity. It is a measure of how"cohesive" the cluster is.If the centroids are not normalized it is the average similarity of the texts in the cluster.
The intra similarity of a clustering:

$$\Phi_{intra}(C) = \frac{1}{n}\sum_{c_i \in C} n_i \cdot sim(c_i, c_i),$$

which is the average similarity of the texts in the set to all texts in their respective clusters.

Similarily, the average similarity of all texts in each cluster to all the texts in the entire set may be calcuated:

$$\Phi_{inter}(C) = \frac{1}{n}\sum_{c_i \in C} n_i \cdot sim(c_i, C).$$

This measures how separated the clusters are.

The purity of a cluster represents the fraction of the cluster corresponding to the largest class of documents assigned to that cluster, thus the purity of the cluster *j* is defined as:

$$Purity(j) = \frac{1}{n_j}\max_i(n_{ij})$$

The overall purity of the clustering result is a weighted sum of the purity values of the clusters:
In general, the larger the purity value is, the better the clustering result is .

$$Purity = \sum_j \frac{n_j}{n} Purity(j)$$

IV. DATASET

We used our own test data sets consisting of ten text documents collected from various sources.

V. RESULT ANALYSIS

The k-means algorithm is very popular for solving the problem clustering a data set into *k* clusters. First, we compare the clustering accuracy of AFS with k-means, k-means with Active feature selection methods. In [1], supervised and unsupervised feature selection methods

were evaluated in terms of improving the clustering performance by conducting experiments in the case that the class labels of documents are available for the feature selection. As a preprocessing step of text clustering, the AFS feature selection was reported as the best among the unsupervised feature selection methods evaluated in .

TABLE 1. SIMILARITY VALUES OF THE CLUSTERS FOR N=2

Measure	Regular K-Means	Active Feature Selection
Inter Cluster Similarity	32	3

Our experimental results demonstrated that the AFS algorithm performs better than k-means in terms of the accuracy of clustering results.

- Feature selection methods can improve the performance of text clustering as more irrelevant or redundant terms are removed.
- The results suggest that performing a unsupervised feature selection method based on the information of clusters obtained during the clustering process can improve the clustering accuracy.
- Result in Table 1 shows that Inter Cluster similarity of AFS is less than that of K-Means which suggest that it is better than regular K-Means.

VI. CONCLUSION

Clustering is one of the most important tasks in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. In order to solve the high dimensionality and inherent data sparsity problems of feature space, feature selection methods are used. In real case, the class information is unknown, so only unsupervised feature selection methods can be exploited. We have proposed a new text clustering algorithm AFS that performs a unsupervised feature selection during the clustering process.The selected features improve the quality of clustering iteratively, and as the clustering process converges, the clustering result has higher accuracy. AFS has been compared with other clustering and feature selection algorithms, such as k-means. Our experimental results show that AFS performs better than other algorithms in terms of the clustering accuracy for different test data sets.

REFERENCES

[1] T. Liu, S. Liu, Z. Chen, and W. Ma, "An Evaluation on Feature Selection for Text Clustering," Proc. of Int'l Conf. on Machine Learning, 2003.
[2] G. Kowalski, Information Retrieval Systems – Theory and Implementation, Kluwer Academic Publishers, 1997.
[3] D.R. Cutting, D.R. Karger, J.O. Pedersen, and J.W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR '92, Pages 318 – 329, 1992.
[4] O. Zamir, O. Etzioni, O. Madani, R.M. Karp, Fast and Intuitive Clustering of Web Documents, KDD '97, Pages 287-290, 1997.
[5] D. Koller and M. Sahami, Hierarchically classifying documents using very few words, Proceedings of the 14th International Conference on Machine Learning (ML), pp. 170-178, 1997.
[6] G. Salton. Automatic Text Processing. Addison-Wesley, New York, 1989.
[7] M. Steinbach, G. Karypis, and V. Kumar. A Comparison of Document Clustering Techniques. In KDD Workshop on Text Mining, 2000.
[8] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992.
[9] B. Larsen and C. Aone. Fast and Effective Text Mining using Linear-time Document Clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.
[10] D. Arthur and S. Vassilvitskii. k-means++ the advantages of careful seeding. In Symposium on Discrete Algorithms, 2007.
[11] Reuters-21578 Distribution 1.0, available at http://www.daviddlewis.com/resources/testcollections/reuters21578
[12] F. Sebastiani, "Machine Learning in Automated Text Categorization,"ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.
[13] Yanjun Li, Congnan Luo, and Soon M. Chung, "Text Clustering with Feature Selection by Using Statistical Data" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. XX, NO. YY, 2008,pp.1-11
[14] N. Sandhya,Y. Sri Lalitha, V.Sowmy, Dr. K. Anuradha, and Dr. A. Govardhan, "Analysis of Stemming Algorithm for Text Clustering", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011 ,pp. 352-359.