A Cluster Based Approach for Classification of Web Results

Apeksha Khabia[1*] and M. B. Chandak[2]

Received: 7-October-2014; Revised: 3-December-2014; Accepted: 15-December-2014

Abstract

Nowadays significant amount of information from web is present in the form of text, e.g., reviews, forum postings, blogs, news articles, email messages, web pages. It becomes difficult to classify documents in predefined categories as the number of document grows. Clustering is the classification of a data into clusters, so that the data in each cluster share some common trait – often vicinity according to some defined measure. Underlying distribution of data set can somewhat be depicted based on the learned clusters under the guidance of initial data set. Thus, clusters of documents can be employed to train the classifier by using defined features of those clusters. One of the important issues is also to classify the text data from web into different clusters by mining the knowledge. Conforming to that, this paper presents a review on most of document clustering technique and cluster based classification techniques used so far. Also pre-processing on text dataset and document clustering method is explained in brief.

Keywords

Text mining, clustering, classification, IF-IDF.

1. Introduction

One of the popular sources of information is World Wide Web (WWW). Large amount of information on web is present in text format and its ever expanding since the day of its perception. Thus, classification of text document is a classical problem in the area of information retrieval.

A large number of techniques have been developed for text classification, including Naive Bayes (Lewis 1998), Nearest Neighbor (Masand 1992), neural networks (Ng 1997), regression (Yang 1994), rule induction (Apte 1994), and Support Vector Machines (SVM) (Vapnik 1995, Joachims 1998) [1]. Among them SVM has been recognized as one of the most effective text classification methods. Many of these techniques are supervised which requires large number of text documents for training to obtain accuracy in classification. Accuracy of classification decreases as the training data set decreases. Also the creation of compact representations of the feature space and the discovery of the complex relationships that exist between features or pattern, documents and classes is the important for text classification. Thus, clustering based classification approach can be used for classification that uses less training data to achieve high classification accuracy and reduction in dimensionality of feature space.

Clustering based classification of text data is of great importance. Its goal is to automatically classify the text documents into different clusters and then exploit them to train the classifier. Among the large amount of text information available in electronic format, only 2% to 5% words of text corpus are used for text analysis and other words such as stop words, white spaces, header, footer etc. are not used for frequent pattern analysis and clustering the documents. Thus, lot of text pre-processing which is task of text mining, is required before text analysis and extracting knowledge from these text data. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output [2]. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling (i.e., learning relations between named entities) [3]. The main idea is to perform text clustering followed by

[*] Author for correspondence

Apeksha Khabia, Computer Science and Engineering Department, SRCOEM, Nagpur, India.
M. B. Chandak, Computer Science and Engineering Department, SRCOEM, Nagpur, India.

classification with trained clusters with selected features. The latter part of this paper explains the approaches for text clustering: literature review, followed by cluster based classification in brief.

2. Approaches for Text Clustering: Literature Overview

Data mining- A process of analyzing data from different perspectives and summarizing it into useful information is known as data mining. Data mining process allows users to understand the substance of relationships present between data. It helps to derive patterns and trends that are hidden among the data. In data mining the main goal is to extract the information from a data set and transform it into an understandable structure which can be used further. Text data mining deals with deriving information form text data. Classification and clustering are some of the techniques used in text data mining.

Classification- Classification of text documents is one of the most common topics, in the field of information retrieval and machine learning. It is most important technique in the field of text data mining. Assigning text documents to one or more class or categories is the main task in document classification. Classification can be done manually or algorithmically. Classification tasks can be divided into three types: Supervised (where external information is used for correct classification), unsupervised (where classification is done without using external information), semi-supervised (where parts of documents use external information.)

Clustering- This is also one of the method for text data mining. In clustering method we make cluster of text documents that are somewhat similar in characteristics. The ultimate aim of the clustering is to form a grouping of similar documents. More often clustering is confused with classification, but there is some difference between these two. In classification the objects are assigned to pre-defined classes, whereas in clustering the clusters are formed. Here classes and clusters could be treated as synonym.

The creation of compact representations of feature space and discovery of substantial relationship that exists between features and classes is necessary in the process of text classification. In this context Clustering has been the alternative representation scheme while classification of text documents in predefined categories from last several years. Clustering helps in feature compression and extraction, to reduce dimensionality of feature vector by joining similar features into clusters.

There are many approaches proposed for clustering by various authors. In [4] clustering is applied on both training and testing dataset. Further the knowledge obtained from these clusters is used to enhance the classification process by exploiting association between index terms and documents. Also in [5] information bottleneck method is applied to find clusters of word which keeps the information about document categories. The lower dimensional feature space obtained from these clusters has been used for classification by naïve bays classifier. Further in [6], information bottleneck method is applied to generate a document representation in a word cluster space instead of word document space where words are viewed as distributions over document categories. Then an information theoretic fast devise algorithm is proposed that uses word clusters for text classification instead of simple words.

Words/terms are clustered by using two dimensional clustering algorithms to classify text documents in [7]. Problem of data sparseness is avoided by clustering features along with clustering training dataset. Also in [8] clustering based classification approach for minimal labelled data is proposed. Clustering on labelled data gives important hint for latent class variables to label the unlabelled data and, thus help to boost the classification step in semi supervised learning. Later in [9] clustering based classification for minimal labelled data with supervised learning is proposed. Under the guidance of labelled data, both labelled and unlabelled data are clustered. The large amount of training data needed for supervised learning is obtained by expanding the training data iteratively with a self-training style clustering strategy.

It would be useful to apply modern classification techniques on available large datasets. But these techniques could not be used directly as they are computationally expensive. So to reduce the datasets to smaller representative sets, variety of clustering techniques have been proposed in [10].

Clustering based classification on feature vector of term document matrix which is of high dimensions and very sparse. This term document matrix is used for training in classification step. Conventionally the frequency of occurrences of terms is contained in term document matrix. But in the same class, the values of the term frequency vary from document to document. Thus, infraclass features are not preserved. Hence, an interval representation of term document matrix for each document is proposed in [11] by using maximum and minimum values of term frequency vectors for the documents mean and standard deviations. This unconventional symbolic data analysis provides methods for effective representations preserving infraclass variations. Single Linkage, Average Linkage, Complete Linkage, K-Means and Fuzzy C-Means clustering algorithms are used for clustering of term feature vector in [11].

3. Concept Overview: Clustered based text classification

Conventionally clustering based classification algorithms consists of basic two steps:

1. Clustering Step: in clustering step, training data set is clustered into number of clusters, so that similar documents categorize into same cluster
2. Classification Step: In classification step, classifiers are trained by using the above formed clusters

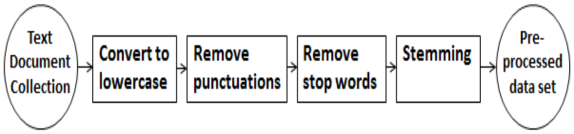

Fig 1: Pre-processing steps in text mining

Clustering of text document is one of text mining tasks. Text mining deals with unstructured data. Text documents fail to get the imposed structure of traditional database, though it out speaks a very wide range of information. Thus, it is important to represent this unstructured data into structured form, so that appropriate patterns and features can be retrieved from this text information.

3.1 Pre-processing of text dataset

The natural language pre-processing operations on text documents collection such as converting to lower case, removing punctuations and stop words, stemming and white space removal (shown in fig. 1) are required for obtaining the structured form of text data. These operations act as pre-processing task. Stop words occur most often in the text documents but they cannot make any sense in the documents. Stop words such as a, an, *the*, *is*, *at*, *which*, *on etc.* are filtered out in the pre-processing of natural language data (text). Stemming in information retrieval is used to describe the process in which infected or derived words are reduced to their base or root form.

Each pre-processed document is treated as bag of words (set of all words with frequency of words appearing in that document). The term-document matrix is formed that describes the frequency of terms that occur in a collection of documents in vector space model. There are various schemes for determining the value that each entry in the matrix should take. One such scheme is TF – IDF (term frequency – inverse document frequency). They are useful in the field of natural language processing. TF– IDF is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining [3]. For each term TF – IDF values are calculated and only terms that have TF – IDF value more than specific threshold are included into the vector space model. Other terms are simply neglected. This is done for dimensionality reduction of feature space.

3.2 Document Clustering

Clustering divides a set of documents into groups such that documents within same group are similar to each other. Documents are grouped together based on some similarity measure into different clusters. Similarity measure means similarity of content. Content based similarity is based on comparison of textual content of documents. Each document has a set of terms and associated frequencies, which help in clustering of documents. Similarity between all pairs of clusters is computed to form a similarity matrix. Documents can be clustered in many ways like hierarchical and K- means technique. Documents can be clustered into hierarchical structure suitable for browsing which suffers efficiency problems. Documents can also be clustered with k-means algorithm and its variants which are more efficient but less accurate. For using k-means clustering, documents should be represented in numeric format.

3.3 Cluster Classification

As clustering results can characterize the Basis for distribution of the whole data set documents, clustering is helpful to aid supervised classification of documents. Thus, clusters can be used to extract useful features and subsequently to augment training data set to improve the performance of classification.

The supervised learning of a classifier can be done using the clustered data set with sufficient features obtained so far. The benefit for integrating the clustering method in classification is that clustering methods are more robust to the bias caused by the initial sparse data [8]. Thus, clustering can be effective when data is sparse.

4. Conclusion and Future Work

Classification problems on text data mainly focus on feature space and relationship between features and classes. This paper presented a brief review on clustering based classification techniques. The central theme in many of these is providing dimensionality reduction to improve text document classification. Clustering helps in reduction of the number of redundant features, which subsequently help in reducing the dimensions.

A very important goal is to achieve high quality information from text available on web. This high quality information helps in clustering and classification. Clustering on text data usually requires- First, parsing that converts unstructured to structured text. Second, text pre-processing operations is to be performed on collection of structured data to obtain pre-processed data. Third, pattern and feature extraction and also similarity measure calculations on text data by mining the knowledge. Fourth, efficient technique for clustering is to be applied to form the clusters with similar documents. K-means clustering algorithm can be used because k-means clustering can be used as feature learning step for supervised classification. Classification requires training of classifier with the obtained clustered result set of text documents. The basic approach is to train a k-means clustering representation first, using input training data. Supervised training with large training data set helps to increase the classification accuracy. With this review we can state that there are main issues of providing large training data set and feature selection procedure to increase the efficiency of classifier. So for feature selection and learning clustering of text documents can be used, which requires small amount of training data. Clustering process is itself feature learning step. In future experiments can be done using k-medoids clustering method for learning of features as it is more robust than k-means. K-medoids helps to minimize the dissimilarities between clusters of documents. There is also scope for concept based text classification as future research.

References

[1] Yang, Y. & Liu, X., “A Re-examination of Text Categorization Methods”, 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, August 15-19, SIGR, 1999, 42-49.

[2] G.K.Gupta, “Introduction to Data Mining with Case Studies”, PHI, 2006.

[3] Deepak Agnihotri, Kesari Verma, Priyanka Tripathi, “Pattern and Cluster Mining on Text Data”, Fourth International Conference on Communication Systems and Network Technologies, Bhopal, April 7-9, IEEE, 2014, 428-432.

[4] Kyriakopoulou A and Kalamboukis T, “Text Classification using Clustering”, Proceedings of ECMLPKDD Discovery Challenge Workshop, 2006.

[5] Slonim N and Tishby, “The power of word clustering for text classification”, Proceedings of the European Colloquium on IR Research (ECIR), 2001.

[6] Dhillon I, Mallela, S and Kumar R, “Enhanced word clustering for hierarchical text classification”, Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, SIGR, 2002, 191–200.

[7] Takamura H and Matsumoto Y, “Two-dimensional clustering for text categorization‖”, Proceedings of the Sixth Conference on Natural Language Learning (COLING - 02), Taiwan, 2002, 29-35.

[8] Hua-Jun Zeng, Xuan-Hui Wang, Zheng Chen, Wei-Ying Ma, “CBC: Clustering Based Text Classification Requiring Minimal Labeled Data”, Third IEEE International Conference on Data Mining, November 19-22, IEEE, 2003, 443-450.

[9] Xue Zhang, Wang-xin Xiao, “Clustering based two-stage text classification requiring minimal training data”, International Conference on Systems and Informatics (ICSAI), Yantai, May 19-20, IEEE, 2012, 2233-2237.

[10] Evans R, Pfahringer B, Holmes G, “Clustering for classification”, Seventh International conference on information technology in Asia (CITA 11), Kuching, Sarawak, July 12-13, IEEE, 2011, 1-8.

[11] B S Harish, S V Aruna Kumar, S Manjunath, “Classifying Text Documents using Unconventional Representation”, International Conference on Big Data and Smart Computing (BIGCOMP), Bangkok, January 15-17, IEEE, 2014, 210-216.

Apeksha Khabia has received her B. E. degree in Computer Science and Engineering from Shri Sant Gajanan Maharaj College of Engineering, Shegaon in 2011. She is pursuing Masters in Technology in Computer Science from Shri Ramdeobaba College of Engineering and Management, Nagpur. Her reseach interest include Text Mining and Natural Language Processing.
Email: apeksha.khabia@gmail.com.

Dr. M. B. Chandak, has received his Ph.D degree from RTM-Nagpur University, Nagpur. He is presently working as Professor and Head of Computer Science and Engineering Department at Shri Ramdeobaba College of Engineering and Management, Nagpur. He has total 21 years of academic expereience, with research interest in Natural Langauge Processing, Advance networking and Big Data Analytics.