IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 7, July 2015.
www.ijiset.com
ISSN 2348 – 7968

A System to Provide Efficient Recommendation Based On Side Information Clustering

Ms. Nikita P.Katariya[1], Prof. M. S. Chaudhari[2]
[1] Dept. of Computer Science & Engg, Priyadarshini Bhagwati College of Engineering Nagpur, India.
[2] Dept. of Computer Science & Engg, Priyadarshini Bhagwati College of Engineering Nagpur, India.

Abstract

Text Databases are rapidly growing due to the increasing amount of information available in various electronic forms. User need to access relevant information across multiple documents. In many text databases, side-information such as document origin information, the links in the document, annotations or other non-textual attributes are available. Therefore, a principled way is needed to perform the mining process, so as to maximize the advantages from using this side information. Thus in this paper we design a method which combines text and side information present inside the text documents for clustering purpose and tried to achieve the goal of a good text clustering scheme i.e. to minimize intra-cluster distances between documents, while maximizing inter-cluster distances using an appropriate distance measure between documents. The experimental results also show that clustering with side information improves the quality of searching information in the text document.

Keywords**:** *side information, clustering, text pre-processing, text mining***.**

1. Introduction

Text mining [4] refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.

Text clustering is the task of automatically organizing text into meaning full cluster or group, In other words, the texts in one cluster share the same topic, and the texts in different clusters represent different topics. There are several number of technique launched for clustering documents since there is rapid growth in the field of internet and computational technologies, the field of text mining have a rapid growth, so that simple text clustering to more demanding task such as production of granular taxonomies, sentiment analysis, and document summarization for the scope of devolving higher quality information from text.

The problem of text clustering is generally defined as “Given a set of texts or documents, would like to partition them into a predetermined or an automatically derived number of clusters, such that the texts assigned to each cluster are more similar to each other than the texts assigned to different clusters”. A tremendous amount of work has been done in recent years on text clustering. However this work is done on pure text.

In many text mining applications, side information [1] is available along with the text documents. Such information may be of different kinds such as document provenance information, the links in the document, user-access behavior from web logs, or other non textual attribute which are embedded into the text document. The common form of side information is the text or data which is present inside the text documents as annotation. Annotation means the explanation or footnote present in the text document. For example in the sentence “We are going for a party (farewell) tonight”, the side information is farewell. In this paper we will present text clustering with side information and experimental result on a text document to illustrate the effectiveness and efficiency of the approach.

Our goal is to show that the advantages of using side-information extend beyond a pure clustering task, and can provide competitive advantages for searching information in the text document. This paper is organized as follows. In the next section, we will present text preprocessing task which is essential in any text mining and text clustering application. We will also present an algorithm for the clustering process in section 3. In Section 4, we will present the experimental results which show

ISSN 2348 – 7968

how clustering with annotations i.e. side information improves efficiency of system. Section 5 contains the conclusions and summary.

2. Text Preprocessing

Text pre-processing[5][6], the task of converting a raw text file, essentially a sequence of digital bits, into a well-defined sequence of linguistically-meaningful units: at the lowest level characters representing the individual graphemes in a language's written system, words consisting of one or more characters, and sentences consisting of one or more words. Pre-processing is an important task applied before any text mining technique on the documents collected from different sources. The text preprocessing in our application performs natural language processing and also finds any annotations present in the text document. Pre-processing is a procedure which can be divided mainly into four text operations (or transformations):

1. Lexical Analysis of the Text - Lexical analysis is the process of converting a stream of characters into a stream of words. Thus, one of the major objectives of the lexical analysis phase is the identification of the words in the text. In lexical analyzer normally removes punctuation marks entirely and converts all the text to either lower or upper case.

2. Elimination of Stopwords - In fact, a word which occurs in 80% of the documents in the collection is useless for purposes of retrieval. Such words are frequently referred to as stopwords and are normally filtered out as potential index terms. Articles, prepositions, and conjunctions are natural candidates for a list of stopwords.

3. Stemming – Many times the user specifies a word in a query but only a variant of this word is present in a relevant document. This problem can be partially overcome with the substitution of the words by their respective stems. A stem is the portion of a word which is left after the removal of its affixes (i.e., prefixes and suffixes). Stems are thought to be useful for improving retrieval performance because they reduce variants of the same root word to a common concept

4. Index Terms Selection - Distinct automatic approaches for selecting index terms can be used. A good approach is the identification of noun groups. Since it is common to combine two or three nouns in a single component (e.g., computer science), it makes sense to cluster nouns which appear nearby in the text into a single indexing component (or concept). A noun group is a set of nouns whose syntactic distance in the text does not exceed a predefined threshold.

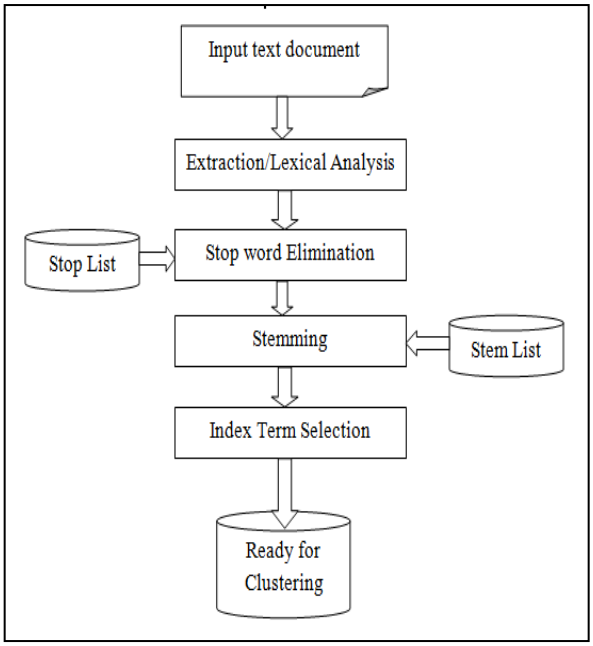

Fig.1 Text pre-processing steps

3. Text Clustering

In this section we present a method of clustering with and without side information. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances using an appropriate distance measure between documents. Many techniques for text clustering are present. We use a method called as bisecting k-means [8] algorithm for text clustering. The bisecting k-means algorithm uses text given by text preprocessing as input. This algorithm firstly performs clustering normally i.e. without using any annotation or side information, then create clusters of annotation and then combines the result of normal clustering and side information clustering to enhance the efficiency of system. The outline of bisecting k-means algorithm is as follows

ISSN 2348 – 7968

Algorithm Bisecting K-Means
Input: K: Number of clusters, D: Top N documents obtained by vector space similarity
Output: K clusters

```
put all the N documents in a single cluster C
        for i=1 to K-1 do
          for j=1 to ITER do
                Use K-means to split C into two
                sub-clusters, C1 and C2
                If(intra-cluster similarity(C1) > intra-
                cluster similarity(C2) )
                    make C1 as permanent
                    C = C2
            else
                 make C2 as permanent
                C = C1
                 end if
          end for
      end for
end Bisecting K-Means.
```

The algorithm starts by putting all the texts or words in a single cluster. It partitions the original cluster into two clusters by using K-Means i.e. K = 2. It makes the cluster which has highest intra cluster similarity as permanent and recursively split the other cluster into two more clusters using K-means with K=2 and continue this until the desired number of clusters are created.

4. Experimental Result

In this section we compare performance of searching algorithm over both a pure text-mining method and a natural alternative which uses both text and side information. For experimental purpose we use a text document which contains information about Linux system and server as sample document. First we apply text preprocessing on sample document. Then on the preprocessed document bisecting k-means algorithm is applied which is discussed in section III, which performs pure text clustering and side information clustering. On both the clusters we apply searching algorithm to search some text. The results show that searching in the clusters with side information is faster than pure text clusters. The effectiveness results for the proposed approach with increasing number of clusters for the sample text document are illustrated in Fig. 2. The number of clusters is illustrated on the *X*-axis, whereas the time required to search (in nanoseconds) with respect to pure and side information clustering methods is illustrated on the *Y*-axis.

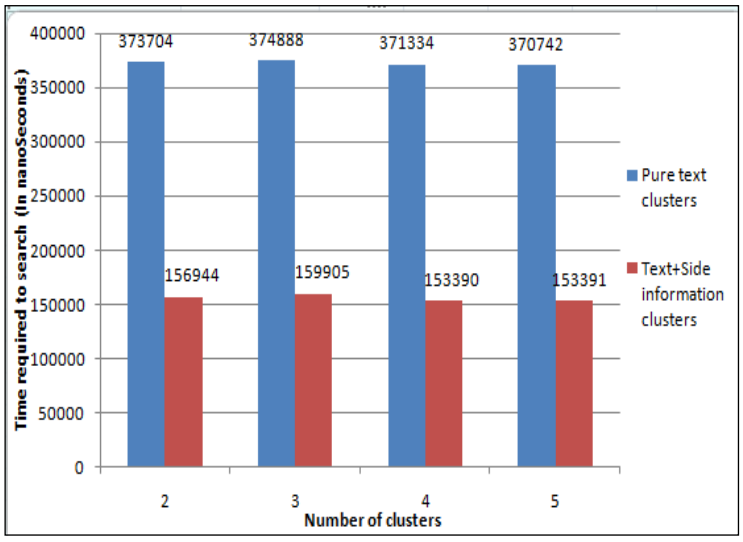

Fig: 2 Experimental results on sample text document

The aim of any clustering approach is to maximize inter cluster difference and minimize intra cluster differences. In plain words, objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters. Clustering with annotation i.e. side information helps to achieve this baseline. The experimental results show that cluster with side information maximizes inter cluster differences by using appropriate distance measures.

5. Conclusion

Text data clustering arises in the context of many application domains. Many forms of text databases contain a large amount of side-information or meta-information, which may be used in order to improve the clustering process. In this paper, we presented methods for clustering and searching text data with the use of side-information. The text data clustering with side information gives a recommendation search engine to improve the clustering process. We present results on sample text document illustrating the effectiveness of our approach. The results show that the use of side-information can greatly enhance the quality of text clustering and searching, while maintaining a high level of efficiency.

ISSN 2348 – 7968

References

[1] Charu C. Aggarwal, Fellow, IEEE Yuchen Zhao, and Philip S. Yu, Fellow, IEEE, "On the Use of Side Information for Mining Text Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 6, JUNE 2014. Pp. 1415-1429.

[2] Mining Text Data", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. 26, NO. 6, JUNE 2014. pp. 1415-1429.

[3] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", *International Journal of Computational Engineering Research (ijceronline.com)*, Vol. 2 Issue. 5, September 2012,pp.1443-1446

[4] C. C. Aggarwal and C.-X. Zhai, "Mining Text Data", New York, NY, USA: Springer, 2012.

[5] Zdenek Ceska and Chris Fox, "The Influence of Text Pre-processing on Plagiarism Detection", International Conference RANLP 2009 - Borovets, Bulgaria, pages 55–59

[6] C.Ramasubramanian and R.Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm", *International Journal of Advanced Research in Computer and Communication Engi*neering, Vol. 2, Issue 12, December 2013

[7] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.

[8] R.Indhumathi, and Dr.S.Sathiyabama, "Efficient time reduction using principal component analysis with bisecting k means algorithm", international journal of engineering science and Technology, Vol. 5 No. 06S Jun 2013, pp. 26-29.

[9] B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R., "Comparative Study of K-means and Bisecting k-means Techniques in Wordnet Based Document Clustering", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-6, August 2012,pp.229-234.