Large Scale Text Clustering Method Study Based on MapReduce

Zhanquan Sun\textsuperscript{1,*}, Feng Li\textsuperscript{2}, Yanling Zhao\textsuperscript{1}, and Lifeng Song\textsuperscript{3}

\textsuperscript{1} Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinan, 250014, China
\textsuperscript{2} Department of History, College of Liberal Arts, Shanghai University, Shanghai, 200436, China
\textsuperscript{3} Information communication section, Shandong Provincial Public Security Bureau, Jinan, Shandong, 250010, China

\texttt{sunzhq@sdas.org}

\begin{abstract}
Text clustering is an important research topic in data mining. Many text clustering methods have been proposed and obtained satisfactory results. Information Bottleneck algorithm, which is based on information loss, can measure complicated relationship between variables. It is taken as one of the most informative text clustering methods and has been applied widely in practical. With the development of information technology, the scale of text becomes larger and larger. Classical information bottleneck based clustering method will be out of work to process large-scale dataset because of expensive computational cost. For dealing with large scale text clustering problem, a novel clustering method based on MapReduce is proposed. In the method, dataset is divided into sub datasets and deployed to different computational nodes. Each computational node will only process sub dataset. The computational cost can be reduced markedly. The efficiency of the method is illustrated with a practical text clustering problem.

\textbf{Keywords:} Text clustering, Large Scale, MapReduce, Information Bottleneck, Feature selection
\end{abstract}

\section{Introduction}

Text Clustering is an unsupervised technique, which groups a collection of documents into a set of clusters, where documents within a cluster share a high level of homogeneity while different clusters exhibit a high level of heterogeneity of information [1]. It has become one of important contents in text mining. In the last decades, many methods have been introduced for text clustering, such as K-means, Clique algorithm, Self Organizing Map (SOM), Information Bottleneck (IB) algorithm and so on[2-4]. Information Bottleneck algorithm, which is based on information loss, can measure complicated relationship between variables. It is taken as one of the most informative

\* Corresponding author.
text clustering methods and has been applied widely in practice. With the development of information technology, the scale of text dataset becomes larger and larger. Classical information bottleneck based clustering method will be out of work to process large-scale dataset because of expensive computational cost. Efficient parallel algorithms and implementation techniques are the key to meeting the scalability and performance requirements entailed in such large scale data mining analyses. Many parallel algorithms are implemented using different parallelization techniques such as threads, MPI, MapReduce, and mash-up or workflow technologies yielding different performance and usability characteristics [5]. MapReduce is a cloud technology developed from the data analysis model of the information retrieval field. The MapReduce architecture in Hadoop doesn’t support iterative Map and Reduce tasks, which is required in many data mining algorithms. Iterative MapReduce architecture software is developed, such as Twister and Spark. Iterative MapReduce supports not only non-iterative MapReduce applications but also an iterative MapReduce programming model [6-7]. Some clustering methods based on MapReduce were proposed, such as k-means, EM, Dirichlet Process Clustering and so on [8]. But the complicated clustering methods are not realized in MapReduce model. In this paper, a novel parallel text clustering method based on parallel IB is proposed to deal with large scale text clustering problems.

The general procedure of text processing is summarized as feature extraction, text representation, clustering process and cluster interpretation. Most textual information is available in the form of natural language. Some text feature selection methods have been developed, such as Document Frequency (DF), Term Contribution (TC), Term variance quality (TVQ) and Term Variance (TV) [9-10]. These unsupervised methods can filter unimportant features efficiently. In this paper, DF and TC are used to select text feature for clustering.

After extracting features, the document collection should be represented using a suitable numerical model in order to be processed by a clustering algorithm. To the best of our knowledge, Vector Space Model (VSM) is the most popular method in representing documents. Tf-idf is taken as the most commonly used feature to represent a text document. In this paper, Tf-idf is used as the feature of a document. It is used to generate document VSM. Parallel IB based on MapReduce is used to cluster the text documents. A paractical text clutering example is analyzed with the proposed method in the end.

2 Feature Selection

Commonly used feature selection methods DF and TC are introduced as follows. They are used to filter unimportant features in clustering.

2.1 Document Frequency (DF)

Document frequency is the number of documents in which a term occurs in a dataset. It is the simplest criterion for term selection and easily scales to a large dataset with linear computation complexity. A basic assumption of this method is that terms appear in minority documents are not important or will not influence the clustering efficiency. It is a simple but effective feature selection method for text categorization.