**ABSTRACT**

With the rapid development of the Internet, the data on the network is growing at a very fast speed and data explosion has become one of the characteristics of the Internet age. If the data which we receive are arranged in logical and structured manner, then the power of information multiplies many times. But unfortunately over 90% of the data is available in unstructured format. Searching the relevant pieces of information in unstructured data is very difficult. So several techniques such as summarization, classification, clustering, information extraction and visualization are available for the same which comes under the category of text mining. Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents. Text mining, also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values. Text clustering which is an important part of text mining is not only used for digging text data independently, but also can be combined with other tasks of text mining, such as text classification, document summaries.

People find it problematic and very tough to extract information from very large number of data sources. Large-scale data need an efficient large-scale data analysis technology. The traditional sequential programming model is a serious lack of scalability, therefore, it can’t satisfy the large-scale data processing demand for computing resources and storage resources. Distributed computing technology represented by MapReduce can greatly improve the efficiency of data-intensive algorithms and make full use of the ability of computing clusters that based on the commodity hardware because of its good scalability.

The emergence of MapReduce distributed computing frameworks greatly reduces the threshold of parallel computing, and the MapReduce programming model has become the best choice of large-scare data processing technology because of its excellent architecture design. To solve the poor scalability of traditional hierarchical clustering when dealing with large-scale corpuses, this paper proposes a parallel text hierarchical clustering algorithm based on MapReduce. As the traditional hierarchical clustering algorithm is designed for sequential programming model, we need to consider the differences between sequential programming model and distributed programming model and make full use of the characteristics of distributed parallel computing platform when parallelize the hierarchical clustering.

There has been some great research of paralleling text clustering based on MapReduce. Professor Bose proposed parallel data mining algorithm based on incremental online MapReduce, which studies have shown that by means of the MapReduce framework can greatly accelerate the speed of the algorithm for large-scale data and can ensure real-time. And Mcnabb proposed optimization algorithm MRPSO MapReduce-based parallel particle swarm, which has good reliability, scalability, and robustness. Besides, some researchers has done some work about paralleling K-Means clustering algorithm based on MapReduce and their result shows that paralleling K-Means clustering algorithm based on MapReduce is feasible and can get great performance. The research of Zhang found that by improving the selection of initial cluster centers, we can optimize the performance MapReduce-based K-Means clustering algorithm. Mao improved the selection of Canopy of Canopy-K-Means algorithm with the "min-max principle", and success in the MapReduce algorithm parallelization computing framework. In order to improve the K-means clustering algorithm processing efficiency for large-scale text corpus, Yi proposed a longitudinal data partitioning strategies related to the content and propose a combination of K-means clustering algorithm data partitioning algorithm. The data partitioning algorithm with the traditional horizontal data partitioning is different, need to combine the characteristics of the input text content to consider data partitioning, which can effectively reduce misclassification, help to improve the accuracy of clustering. Also verify that the longitudinal partition strategy based on content-related than the transverse partitioning strategy has better clustering results. The research of Yu presented a hierarchical clustering based on parallel MapReduce algorithm, so that the hierarchical clustering algorithm can also be applied to a parallel text clustering.

In the other hand, there are some researches that focused on clustering algorithms on the other parallel programming models. Olson proposed a parallel hierarchical clustering algorithm based on Parallel Random Access Machine(PRAM), parallel networks and butterfly tree model and analyzed its time complexity. Rajasekaran proposed a parallel SRAM-based hierarchical clustering algorithm, the algorithm is the premise has n2/log n PRAM allow simultaneous read and write, the algorithm time complexity is O(log n). In addition, some other studies focused on the improvement and expansion issues MapReduce distributed computing framework. For the MapReduce programming model lacks built-in support for an iterative improvement program, Bu found a solution what is adding caching mechanism in the task scheduler loop perception, and it made MapReduce more effectively support an iterative process.

The following is a brief summary of the main contents of my work:

* Build Spark platform for my research. Spark is an open source cluster computing framework originally developed in the AMPLab at University of California, Berkeley but was later donated to the Apache Software Foundation where it remains today. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's multi-stage in-memory primitives provides performance up to 100 times faster for certain applications.
* Study master idea, the theoretical basis, the steps, the advantages, the disadvantages and the use of scenarios of the algorithm that has higher frequencies in the actual production. Analyzing problems that we will encounter during processing large data and proposing solutions.
* Study and propose relevant text clustering algorithm parallelization scheme, the proposed scheme to be able to solve the problem with traditional text clustering algorithm when dealing with a large-scale text data. Also, specific coding implementation.
* Assess the performance of the proposed scheme, the advantages and disadvantages of this issue.

**Keywords**: MapReduce, Text Clustering, Parallel Computing