ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

HADOOP MAPREDUCE FOR TEXT CLUSTERING

Sangeetha.V, Kaviya.R*, Dr.R.Subhashini
UG Students, Department of IT, Sathyabama University, Chennai – 600119.
Research Head and Faculty of Computing, Sathyabama University, Chennai – 600119.
Email: kaviyar128@gmail.com

Received on 20-04-2016 *Accepted on 24-05-2016*

Abstract

The project introduce map reduce approach for clustering the text documents using vector space model. The project implements data mining technique named clustering to enhance the efficiency of information retrieval process. Document clustering (or text clustering) is the application of clusterresearch to textual documents. It has applications in automatic document organization, topic extraction and fast data retrieval or filtering.

Keywords: Big Data, MapReduce, Clustering, MapTask, and ReduceTask.

1. Introduction

The project proposes the faster information retrieval system in internet search engine. The project introduce map reduce approach for clustering the text documents using vector space model. The project implements data mining technique named clustering to enhance the efficiency of information retrieval process. Document clustering (or text clustering) is the application of clusterresearch to textual documents. It has applications in automatic document organization, topic extraction and fast data retrieval or filtering. Current sequential text classification approaches applied to large-scale of text documents. It's generally requiring a large number of training inputs to accurately classify large volume of text documents leading to more processing time. The problem of this research is how to build a text clustering function. Its approach for large volume text that reduces the time and efficiency of retrieval. To develop a faster information retrieval system in internet search engine. The project will be a great advantage for the users to retrievethe information in the web quickly. It enhances the efficiency of the information retrieval system. Text clustering becomes mandatory step for search engines to group the same text documents for faster information

retrieval .K-means algorithm is also used to cluster the documents. Vector space model is an algebraic model for representing text documents. It is used in information filtering, data retrieval, indexing and relevancy rankings.

II. Literature Survey

In a document retrieval, or other pattern corresponding environment where stored entities (documents) are distinguished with each other or with arriving patterns (search requests), it appears that the best indexing (property) space is one where each entity falls as far away from the others as possible; in these circumstances the value of an indexing system may be conveyable as a function of the density of the object space; in particular, retrieval performance may match inversely with space density. An approach depends on space density computations are used to choose an optimum indexing lexicon for a collection of documents. Typical simplification results are shown, demonstrating the usefulness of the model [3].

Data clustering has been received substantial attention in many applications, such as data mining, document retrieval, and image partition and pattern classification. The extending volumes of information emerging by the improvement of technology, makes clustering of very large scale of data a demanding task. In order to deal with the problem, many researchers try to develop efficient parallel clustering algorithms. Here we put forward a parallel *k*-means clustering algorithm depend on MapReduce, which is a simple yet strong parallel programming technique. The experimental outputs reveal that the proposed algorithm can scale well and capably process large datasets on commodity hardware [4].

MapReduce is a programming model and a related implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to create a set of intermediate key/value pairs, and a reduce function that merges all intermediate values related with the same intermediate key. Many real world tasks are expressible in this model, as mentioned here.

Programs written in this functional style are automatically parallelized and implemented on a huge cluster of commodity devices. The run-time system takes care of the details of partitioning the input data, scheduling the program's implementation over a set of machines, handling machine failures, and managing the required inter-machine communication. This provides programmers without any experience with side by side and distributed

systems to easily use the resources of a large distributed system [6].

This tutorial is motivated by the clear need of many organizations, companies, and researchers to perform with big data volumes efficiently. Examples include web analytics applications, scientific applications, and communal networks. A well-known data processing engine for big data is Hadoop MapReduce. Early versions of Hadoop MapReduce endure from severe performance problems. Today, this is becoming history. There are many methodologies that can be applied with Hadoop MapReduce jobs to boost performance by orders of magnitude. In this tutorial we teach such techniques.

1st, we will briefly familiarize the audience with Hadoop MapReduce and motivate its use for big data processing. Then, we will pivot on different data management techniques, going from job optimization to physical data firm like data layouts and indexes. Throughout this tutorial, we will mention the similarities and dissimilarity between Hadoop MapReduce and Parallel DBMS. Moreover, we will locate out unresolved research problems and open issues [5].

This paper presents two approaches to semantic search by incorporating Linked Data annotations of documents into a Generalized Vector Space Model.

One model steals taxonomic relationships among entities in documents and queries, while the other model computes term weights depend on semantic relationships within a document. We publish an evaluation dataset with annotated documents and queries as well as user-rated relevance assessments. The evaluation on this dataset shows significant improvements of both models over traditional keyword based search [7].

III. Proposed Structure

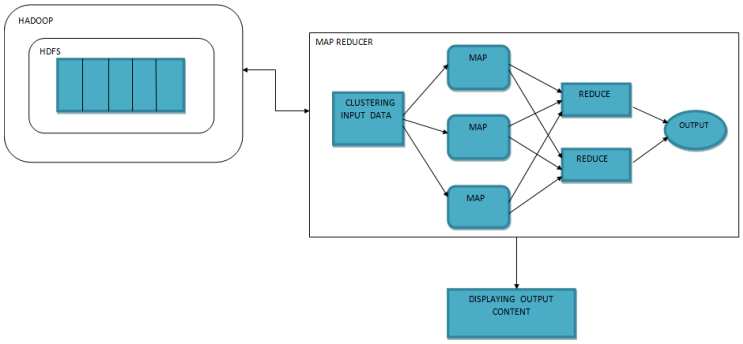
Canopy clustering algorithm can be used as initial step to get initial centroids and then lies this as input to map reduce k-means algorithm. The text corpus can be separated by applying map reduce text clustering based on frequent item sets.

The project implements HADOOP to store and process the data in parallel and distributed environment. Vector space model which is implemented by k-means algorithm is a well-known methodology for clustering the text documents.

The efficiency of this system can be increased by increasing the set of text (Text corpus) along with the number of nodes in the cluster. Faster information retrieval. Since most of the data available in the internet is in the form of

unstructured data. Text clustering plays a major role in the search engine to group the similar text documents. Larger set of data can be processed and placed. Process structured, unstructured and semi structured info.

- Clustering
- Map reduce



A. Clustering

Cluster analysis or clustering is the task of joining a set of elements in such a way that objects in the similar group are more similar to each other than to those in other groups.

Cluster research itself is not one specific algorithm, but the usual task to be solved. It can be achieved by various procedures that differ significantly in their notion of what initiate a cluster and how to efficiently find them. Well-known notions of clusters include groups with small distances amid the cluster members, thick areas of the data space, intervals or specific statistical distributions. Clustering can therefore be work out as a problem. The appropriate clustering algorithm and parameter settings (adding values such as the depth function to apply, a density threshold or the number of expected clusters) based on the individual data set and intended use of the outputs. Cluster analysis as such is not an automatic work, but an iterative process of knowledge discovery or interrupt multi-objective optimization that involves trial and failure. It will usually be necessary to change data preprocessing and model parameters until the output achieves the desired properties. In addition the term clustering, there are a number of terms with same meanings, including automatic categories, numerical taxonomy, bryology (from Greek "grape") and typological research. The subtle differences are usually in the usage of the results: while in data mining, the final groups are the matter of interest, in automatic categories the resulting judicial power is of interest. This generally becomes misunderstandings between researchers reaching from the fields of data mining and machine understanding,

since they use the same terms and usually the same algorithms, but have different goals.

B Map Reduce

Hadoop Map Reduce is a software framework for effortlessly writing applications which process big amounts of data in-parallel on huge clusters (thousands of nodes) of commodity hardware in a dependable, fault-tolerant manner.

The term MapReduce really refers to the following 2 various jobs that Hadoop programs perform:

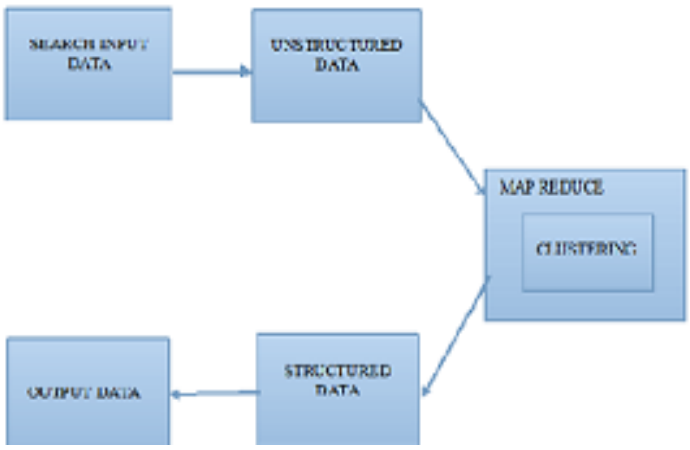
- **The Map Task:** This is the 1st consider, which takes input data and transfer it into a set of data, where individual sets are fragmented down into tuples (key/value pairs).
- **The Reduce Task:** This task considers the output from a map task as input and joins those data tuples into a smaller set of tuples. The decrement task is always performed after the map task.

Generally both the input and the output are placed in a file-system. The framework takes care of programming tasks, observing them and re-implements the failed tasks.

The MapReduce framework composed of a single master **JobTracker** and one slave **TaskTracker** / cluster-node. The master is in charge for resource management, tracking resource consumption/availability and programming the jobs integral tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves Task Tracker execute the tasks as directed by the master and provide job-status data to the master periodically.

The JobTracker is a one point of failure for the Hadoop MapReduce service which means if JobTracker goes decrement, all running jobs are halted.

IV. Background & Related Works:



Big Data

Big data is a term for data sets that are so huge or complex that customary data processing applications are insufficient. Challenges include research, capture, data duration, search, sharing, storage, convert, visualization, and querying and information privacy.

The term usually refers simply to the use of foretell analytics or certain other advanced mechanism to extract value from data, and never to a particular size of data set. Accuracy in big data may become more confident decision making, and better conclusion can result in greater operational efficiency, cost reduction and decreased risk. Analysis of data sets can find new union to "spot business trends, stop diseases, crime and so on.

"Scientists, business execution, practitioners of medicine, advertising and governments like regularly meet difficulties with large data sets in areas adding Internet search, finance and informatics. Data sets are increasing rapidly in part since they are increasingly grouped by cheap and numerous information-sensing mobile instruments, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and WSNs.

The world's technological per-capita capacity to place information has roughly doubled every 40 months since the 1980s; as of 2012, daily 2.5 extra bytes (2.5×10^{18}) of data are created. One question for huge enterprises is deciding who should own big data initiatives that affect the whole organization.

RDBMs and desktop statistics and visualization packages usually have difficulty handling big data. The job instead needs "massively parallel software running on 10’s, 100’s, or even thousands of servers". What is consider "big data" varies depend on the capabilities of the users and their tools, and enlarging capabilities make big data a moving earmark.

"For some firm, facing 100’s of gigabytes of data for the 1st time may trigger a need to revise data management options. For others, it may take 10’s or 100’s of terabytes before data size becomes a significant consideration.

V. Conclusion

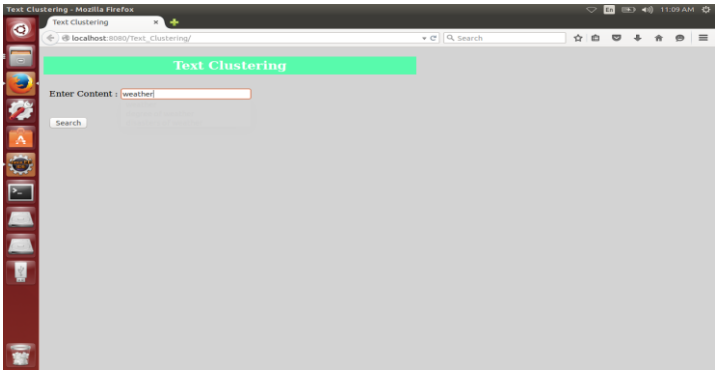
Information retrieval mechanisms are extensively popular in most of the search engines to efficiently organize and retrieve information systems. Most of the data in network is in the layout of unclear and semi structured. Currently clustering techniques are used to arrange and group the similar data objects to retrieve search results faster. Custom

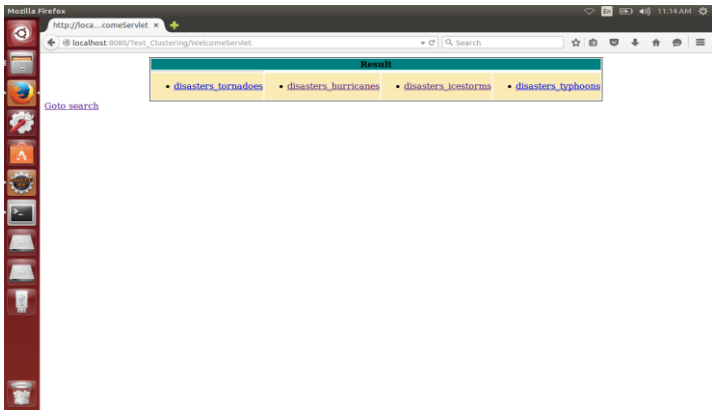
way of clustering text documents is vector space replica, in which tf-idf is used for k-means algorithm with supportive similarity measure. As the info is abundantly increasing day by day, elastic resources are required to store and compute.

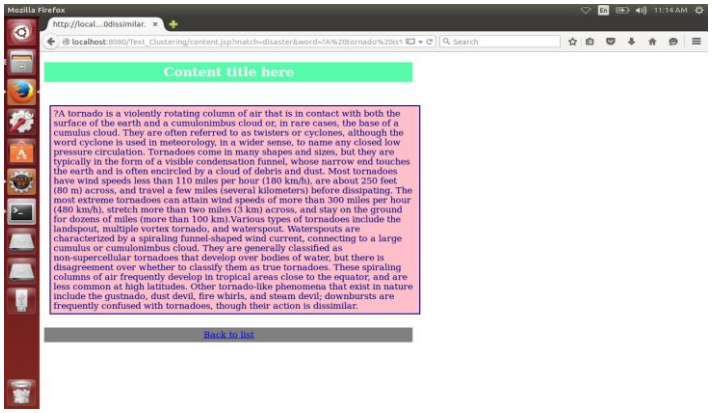
Hadoop framework helps to place and calculate big data in parallel and distributed platform with the help of HDFS and MapReduce. The outputs reached by implementing map reduce k-means algorithm on both single and multi-node cluster shows that the achievement of the procedure increases as the text corpus increases in multi node cluster. In the current method of map reduce k-means algorithm, the opening centroids are chooses randomly.

VI. Output

We are giving an input in search engine using query, the output is retrieved from MapReduce and unstructured data to be send to clustering mechanism and give a fast information retrieval results.







VII. Reference

1. Anna Huang," Similarity measures of Text document clustering",NZCSRSC 2008, Christchurch, New Zealand, April 2008.
2. Apache Lucene Hadoop[EB/OL]. http://hadoop.apache.org/.
3. G. Salton , A. Wong , C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, version.18n.11, pages.613-620, Nov. 1975.
4. GeorgeTsatsaronis and Vicky Panagiotopoulou. A generalized vector space model for text retrieval based on semantic relatedness. Proceedings of the EACL 2009 student research workshop, pages 70-78, April 2009.
5. J Dittrich, JA Quiané-Ruiz . Efficient big data processing in Hadoop MapReduce. Proceedings of the VLDB Endowment,2012 - dl.acm.org, Volume 5 Issue 12, August 2012 ,Volume 5 Issue 12, August 2012.
6. Jeffrey, D. and G. Sanjay,. MapReduce: simplified data processing on large clusters. Commun. ACM, 51(1): pages 107-113 2008.
7. Sanjay, G., G. Howard, and L. Shun-Tak, The Google file system,in Proceedings of the nineteenth ACM symposium on Operating systems principles. ACM: Bolton Lan,ding, NY, USA 2003.
8. Weizhong Zhao, Huifang Ma, Qing He. Parallel *K*Means Clustering Based on MapReduce. Processdings of First international.

Corresponding Author:

Kaviya.R,

Email: *kaviyar128@gmail.com*