ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

# Big Data Mining from Social Networking Services using Spectral Clustering Algorithm

Azizkhan F Pathan[1], Dr. Chetana Prakash[2]

Assistant Professor, Dept. of CS&E, Jain Institute of Technology, Davangere, India[1]

Professor & Head, Dept. of MCA, Bapuji Institute of Engineering and Technology, Davangere, India[2]

**ABSTRACT:** Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time. Those data on the Internet exist in vast scale and grow rapidly, so it is urgently required in technology to mine high-value information from the mass data. This paper introduces an efficient parallel spectral clustering algorithm. The experimental results show that the proposed parallel spectral clustering algorithm is suitable for applying in mass data mining.

**KEYWORDS**: Big Data, HACE, data-driven model

## I. INTRODUCTION

Enormous amounts of data are generated every minute. Some sources of data, such as those found on the Internet are obvious. Social networking sites, search and retrieval engines, media sharing sites, stock trading sites, and news sources continually store enormous amounts of new data throughout the day. For example, a recent study estimated that every minute, Google receives over 2 million queries, e-mail users send over 200 million messages, YouTube users upload 48 hours of video, Facebook users share over 680,000 pieces of content, and Twitter users generate 100,000 tweets. Some sources of data are not as obvious. Consider the vast quantity data collected from sensors in meteorological and climate systems, or patient monitoring systems in hospitals. Data acquisition and control systems, such as those found in cars, airplanes, cell towers, and power plants, all collect unending streams of data. The health care industry is inundated with data from patient records alone. Insurance companies collect data for every claim submitted, fervently working to catch increasing quantities of fraudulent claims.

Regardless of the source of the data, contained within them are nuggets of knowledge that can potentially improve our understanding of the world around us. The challenge before us lies in the development of systems and methods that can extract these nuggets.

We are surely living in an interesting era – the era of big data and cloud computing, full of challenges and opportunities. Organizations have already started to deal with pet byte-scale collections of data; and they are about to face the Exabyte scale of big data and the accompanying benefits and challenges. Big data is believed to play a critical role in the future in all walks of our lives and our societies. For example, governments have now started mining the contents of social media networks and blogs, and online-transactions and other sources of information to identify the need for government facilities, to recognize the suspicious organizational groups, and to predict future events (threats or promises). Additionally, service providers start to track their customers' purchases made through online, instore, and on-phone, and customers' behaviors through recorded streams of online clicks, as well as product reviews and ranking, for improving their marketing efforts, predicting new growth points of profits, and increasing customer satisfaction. The mismatch between the demands of the big data management and the capabilities that current DBMSs can provide has reached the historically high peak. The three Vs (volume, variety, and velocity) of big data each implies one distinct aspect of critical deficiencies of today's DBMSs.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

ranging from a few dozen terabytes to many petabytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale**.**

Big data can be described by its volume, variety and velocity. Volume is nothing but the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. Variety means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data. The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development. Variability is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively. The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data. Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

Big Data has its applications in International development, Manufacturing, Cyber-Physical Models, Media, Technology, Private sector, Retail, Retail Banking, Real Estate, and in Science and Research**.**

## II. RELATED WORK

Scalability is at the core of the expected new technologies to meet the challenges coming along with big data. The simultaneously emerging and fast maturing cloud computing technology delivers the most promising platforms to realize the needed Scalability with demonstrated elasticity and parallelism capacities. Numerous notable attempts have been initiated to exploit massive parallel processing architectures[1]. Google's novel programming model, MapReduce [2], and its distributed file system, GFS (Google File System) [3], represent the early groundbreaking efforts made in this line. From the data mining perspective, mining big data has opened many new challenges and opportunities. Even though big data bears greater value (i.e., hidden knowledge and more valuable insights), it brings tremendous challenges to extract these hidden knowledge and insights from big data since the established process of knowledge discovering and data mining from conventional datasets was not designed to and will not work well with big data. The cons of current data mining techniques when applied to big data are cantered on their inadequate Scalability and parallelism. In general, existing data mining techniques encounter great difficulties when they are required to handle the unprecedented heterogeneity, volume, speed, privacy, accuracy, and trust coming along with big data and big data mining. Improving existing techniques by applying massive parallel processing architectures and novel distributed storage systems, and designing innovative mining techniques based on new frame works/platforms with the potential to successfully overcome the aforementioned challenges will change and reshape the future of the data mining technology.

Gigantic volume requires equally great scalability and massive parallelism that are beyond the capability of today's DBMSs; the great variety of data types of big data particularly unfits the restriction of the closed processing architecture of current database systems [5]; the speed/velocity request of big data (especially stream data) processing asks for commensurate real-time efficiency which again is far beyond where current DBMSs could reach. The limited availability of current DBMSs defeats the velocity request of big data from yet another angle (Current DBMSs typically require to first import/load data into their storage systems that enforces a uniform format before any access/processing is allowed. Confronted with the huge volume of big data, the importing/loading stage could take hours, days, or even months. This causes substantially delayed/reduced availability of the DBMSs). To overcome this scalability challenge of big data, several attempts have been made on exploiting massive parallel processing architectures. The first such attempt was made by Google. Google created a programming model named MapReduce [5] that was coupled with (and facilitated by) the GFS (Google File System [6]), a distributed file system where the data can be easily partitioned over thousands of nodes in a cluster. Later, Yahoo and other big companies created an Apache open-source version of Google's MapReduce framework, called Hadoop MapReduce. It uses the Hadoop Distributed File System (HDFS) – an open source version of the Google's GFS. The MapReduce framework allows users to define two functions, map and reduce, to process large number data entries in parallel [7]. More specifically, in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

MapReduce, the input is divided into a large set of key-value pairs first; then the map function is called and forked into many instances concurrently processing on the large key-value pairs. After all data entries are processed, a new set of key-value pairs are produced, and then the reduce function is called to group/merge the produced values based on common keys. In order to match/support the MapReduce computing model, Google developed the BigTable – a distributed storage system designed for managing structured data.

BigTable can scale well to a very large size: petabytes of data across thousands of commodity severs [8]. In the same spirit, Amazon created Dynamo [9], which is also a key-value pair storage system. The Apache open-source community acted quickly again, created HBase – an opensource version of Google's BigTable built on top of HDFS and Cassandra – an open-source version of Amazon's Dynamo. Apache Hive [10] is an open source data warehouse system built on top of Hadoop for querying and analysing files stored in HDFS using a simple query language called HiveQL. Hadoop is not alone; it has other competitor platforms. All these platforms lack many niceties existing in DBMSs. Some of the competitors improved on existing platforms (mostly on Hadoop), and others came up with a fresh system design. However, most of these platforms are still in their infancy. For example, BDAS, the Berkeley Data Analytics Stack [11], is an open-source data analytics stack developed at the UC Berkeley AMPLab for computing and analyzing complex data. It includes the following main components: Spark, Shark,, and Mesos. Spark is a high-speed cluster computing system that performs computations in memory and can outperform Hadoop by up to 100x. Shark is a large-scale data analysis system for Spark that provides a unified engine running SQL queries, compatible with Apache Hive. Shark can answer SQL queries up to 100x faster than Hive, and run iterative machine learning algorithms up to 100x faster than Hadoop, and can recover from failed mid-queries within seconds [12]. Mesos is a cluster manager that can run Hadoop, Spark and other frameworks on a dynamically shared pool of compute nodes. ASTERIX [13] is data intensive storage and computing platform. Some notable drawbacks of Hadoop and other similar platforms, e.g., single system performance, difficulties of future maintenance, inefficiency in pulling data up to queries and the unawareness of record boundaries, are properly overcome in ASTERIX by exploring runtime models inspired by parallel database system execution engines. In ASTERIX, the open software stack is layered in a different way that it sets the data records at the bottom layer, facilitating a higher-level language API at the top.

While the majority of the big data management and processing platforms have been (or are being) developed to meet business needs, SciDB is an open source data management and analytics (DMAS) software system for data-intensive scientific applications like radio astronomy, earth remote sensing and environment observation and modelling. The difference between SciDB and other platforms is that SciDB is designed based on the concept of array DBMS (i.e., raster data) where big data is represented as arrays of objects in unidimensional or multidimensional spaces. SciDB is designed to support integration with high-level imperative languages, algorithms, and very large scales of data [4].

## III. DATA MINING PROCESS

Generally, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining process mainly consists of 5 major steps as shown in figure 3.1:

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

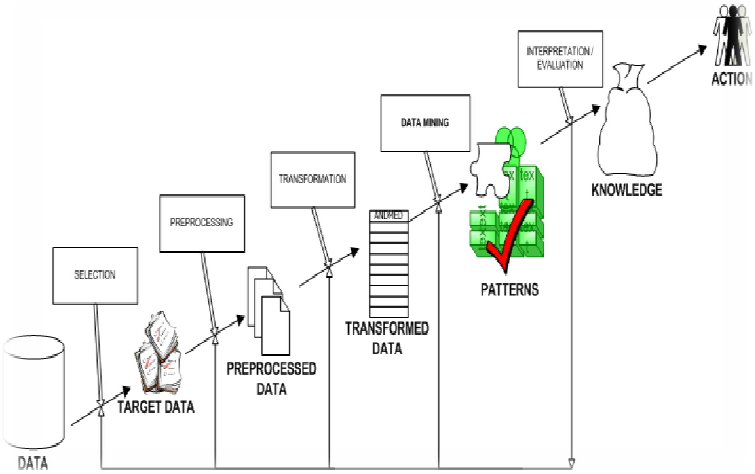


**Figure 3.1 Data Mining Process**

The goals of big data mining techniques go beyond fetching the requested information or even uncovering some hidden relationships and patterns between numeral parameters. Analysing fast and massive stream data may lead to new valuable insights and theoretical concepts. Comparing with the results derived from mining the conventional datasets, unveiling the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge and insights in the target domain.

## IV. SYSTEM ARCHITECTURE

As shown in figure 4.1, the relevant information can be mined using the word from big data. Here we are using twitter as the API, and JASON PARSER is the one which accesses the data from the server. Once the search is successful, the data collected has been displayed according to the USER ID, NAME, LOCATION and the RETWEET COUNT on the screen. But there is a huge set of data that is displayed on the screen. So to get the appropriate data we are going to apply spectral clustering technique. It is one of the techniques in Data mining where we will get the refined data according to LOCATION, HASH TAG and RETWEET COUNT based information.

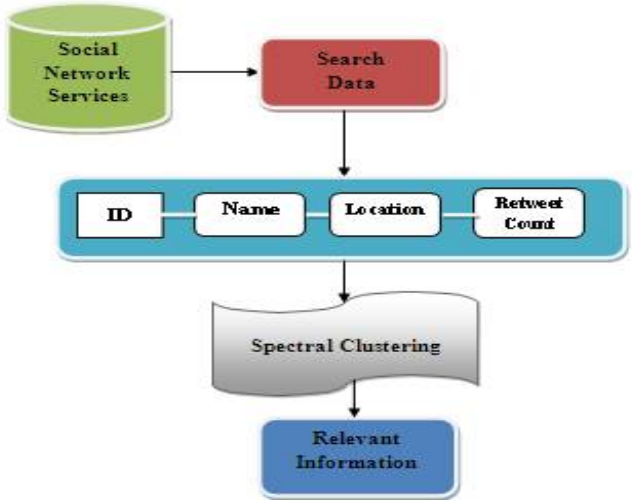


**Figure 4.1 System Architecture**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

## V. SPECTRAL CLUSTERING ALGORITHM

The standard serial spectral clustering algorithm steps are as follows:
(1) By computation, obtain the similar matrix $S \in R^{n \times n}$ and then sparse it.
(2) Construct diagonal matrix D.
(3) Compute the standard Laplace matrix L.
(4) Compute k minimum eigenvectors of matrix L, and compose matrix $Z \in R^{n \times k}$ which contains k minimum eigenvectors and are regarded as the columns of the matrix Z.
(5) Standardize it as $Y \in R^{n \times k}$
(6) Use K-Means algorithm to cluster the data point $yi \in R^{k}(i=1,\ldots,n)$ into k clusters.

Because the Hadoop MapReduce can provide outstanding distributed computing framework, we realize our parallel spectral clustering algorithm in the Hadoop MapReduce.

Firstly, we put the data point $x_1,\ldots, x_n$ in HBase chart, which can be accessed by each machine, and the line key (row key) of each data point xi is set as the subscript $i \varepsilon \{1,\ldots, n\}$ of the data point. Then we use a Reduce function to automatically distribute the similar values between the calculated data points. For each data point $x_i$ with identification i, Reduce function will only clear those whose subscripts are equal to or bigger than i with the data point of $x_j(j = i,\ldots, n)$ and the similar value of $x_i$. We can call it "the similar value calculation of subscript i". In this way, the similar value between each pair of data points can be calculated only once. The apparent "similar value calculation of subscript i" and "similar value calculation of other subscripts" are independent from each other. Therefore, if we distribute different subscripts to different machines, then "similar value calculation of subscript i" can be operated in distributed environment.

Especially, "similar value calculation of subscript i" needs to calculate the similar value $\{< x_i, x_i >, < x_i, x_{i+1} >,\ldots, < x_i, x_n >\}$ of $n - i + 1$ data point pairs. That is to say, the first subscript 1 needs to compute similar value of n data point pairs, and the last subscript n only needs to compute the similar value of a data point, that itself is $< x_n, x_n >$. In order to balance the calculation of similar value, we put the "similar value calculation of subscript 1" and "similar value calculation of subscript n" together, and "similar value calculation of subscript 2" and "similar value calculation of subscript $n - 1$" together, and so on (see Figure 1). When the calculation of similar values is completed, put them back on HBase table and they will be used to calculate the Laplace matrix in later steps. The process of parallel construction of similar matrix can be shown in Algorithm 1.

**Algorithm 1 parallelized constructing the reduce function in similarity matrix**

Input: <key, value>, key is the subscript index of data point, and value is supposed as null.
Output:<key', value'>=<key,value>

```
1. index=key, another_Index=n-key+1
2. for i in{index,another_Index}
        i_content=get Content From HBase(i):
        for j=i to n do
                j_content=get Content From HBase(j);
                sim=compute Similarity(i_content,j_content);
                store Similarity(i, j, sim) into HBase table;
        End for
    End for
3. Output<key,null>
4. End.
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

## VI. RESULTS AND DISCUSSION

We can search the word in big data. Here we are using twitter as the API, and JASON PARSER is the one which accesses the data from the server. As shown in the figure 6.1, the word has been searched using the big data.

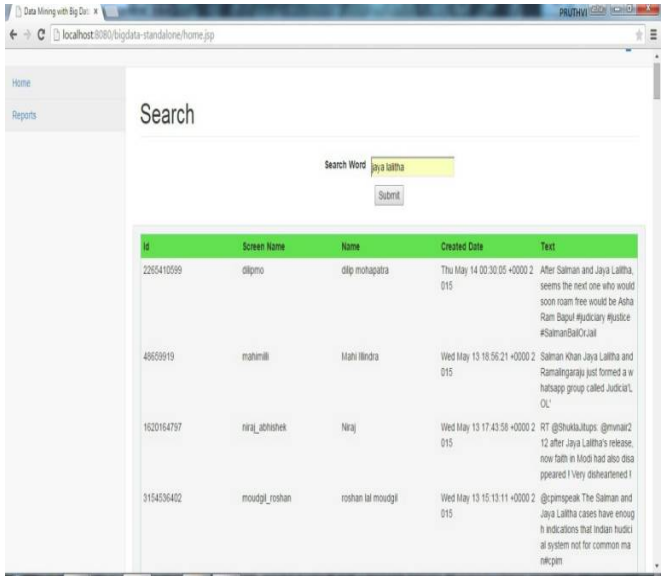


**Figure 6.1 Search Word**

After the search has been finished the data has been collected according to the USER ID, NAME, LOCATION and the RETWEET COUNT and it will be displayed on the screen as shown in the figure 6.2.

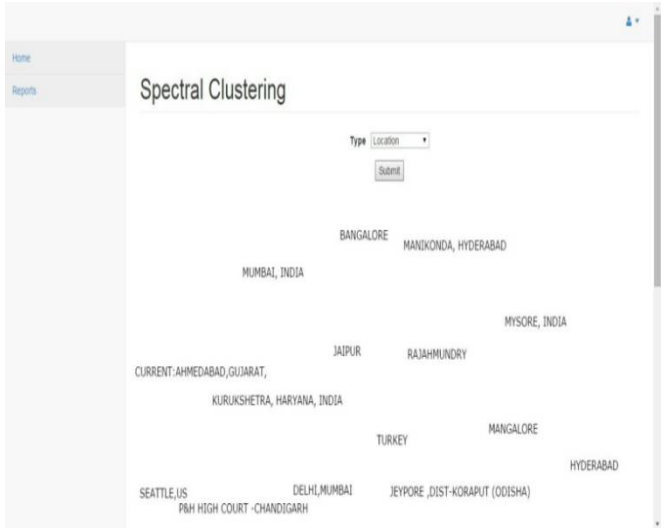


**Figure 6.2 Location Based Search**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

But there is a huge sets of data that is displayed on the screen. So in order to get the appropriate data we are going to apply spectral clustering technique. It is the one of the technique in Data mining where we will get the refined data according to LOCATION, HASH TAG and RETWEET COUNT based information as shown in figures 6.3 and 6.4.

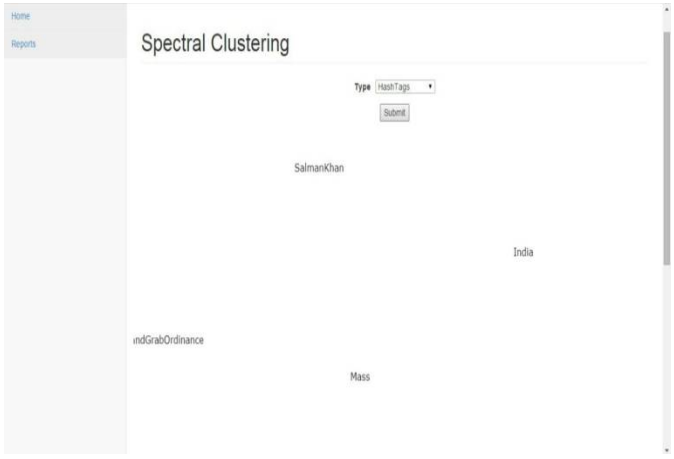


**Figure 6.3 Hash Tag Based Search**

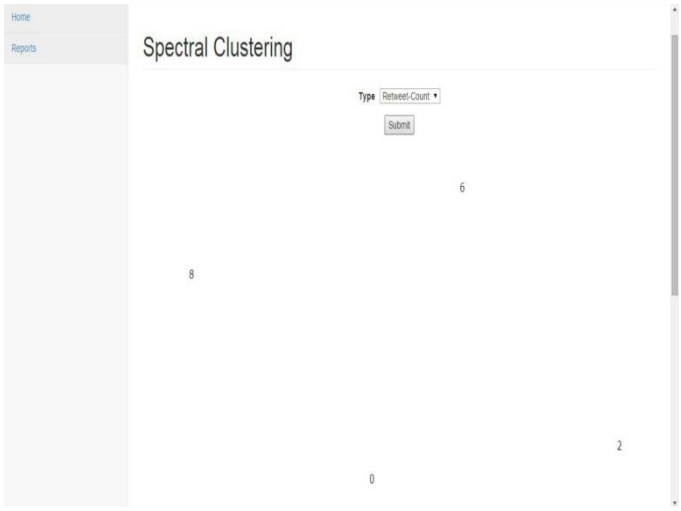


**Figure 6.4 Retweet Count Based Search**

## VII. CONCLUSION

The Big Data is regarded as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time. Those data on the Internet exist in vast scale and grow rapidly, so it is urgently required in technology to mine high-value information from the mass data. As a kind of unsupervised learning method, clustering algorithm is a technique commonly used in data statistics and analysis which contains data mining, machine learning, pattern recognition, image analysis, and many other areas. The traditional serial clustering algorithm has two problems and it is difficult to meet the needs of practical applications: the first one is that the speed of clustering is not fast enough and the efficiency is not high; the other one is that in the face of mass data, subject to the limits of memory capacity, it often cannot run effectively. This paper introduces an efficient

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

parallel spectral clustering algorithm. The strategy of parallel spectral clustering algorithm is to compute similar matrix and sparse according to data points segmentation; when computing eigenvectors, store the Laplacian matrix on the distributed file system HDFS, use distributed Lanczos to compute and get the eigenvectors by parallel computation; at last, in terms of the transposed matrix of eigenvectors, adopt the improved parallel K-Means cluster to obtain the clustering results. Through adopting different parallel strategies about each step of the algorithm, the whole algorithm gets linear growth in speed. The experimental results show that the proposed parallel spectral clustering algorithm is suitable for applying in mass data mining.

## REFERENCES

1. Berkovich S., Liao D., "On Clusterization of big data Streams", 3rd International Conference on Computing for Geospatial Research and Applications, article no. 26. ACM Press, New York, 2012.
2. Beyer M.A., Laney D., "The Importance of 'Big Data': A Definition", Gartner, 2012.
3. Madden S., "From Databases to big data", IEEE Internet Computing 16(3), 4–6, 2012.
4. Shmueli G., Patel N.R., Bruce P.C., "Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XL Miner", 2nd edn. ,Wiley & Sons, Hoboken, 2010.
5. Ghoting A., Kambadur P., Pednault E., Kannan R., "NIMBLE: a Toolkit for the Implementation of Parallel Data Mining and Machine Learning Algorithms on MapReduce", 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, pp.334–342, 2011.
6. Low Y., Bickson D., Gonzalez J., Guestrin C., Kyrola A., Hellerstein J.M., "Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud", VLDB Endowment 5(8), 71–727, 2012.
7. Borkar V.R., Carey M.J., Li C., "Big data Platforms: What's Next?", ACM Crossroads 19(1), 44–49, 2012.
8. Sun Y., Han J., Yan X., Yu P.S., "Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach", VLDB Endowment 5(12), 2022–2023, 2012.
9. Tene O., Polonetsky J., "Privacy in the Age of big data: A Time for Big Decisions", Stanford Law Review Online 64, 63–69, 2012.
10. Gartner, "Gartner Says Big Data Will Drive $28 Billion of IT Spending in 2012", October 17, 2012
11. NewVantage Partners: Big Data Executive Survey(2013), http://newvantage.com/wpcontent /uploads /2013/02/ NVP-Big- Data-Survey-2013-Summary-Report.pdf
12. Xin R.S., Rosen J., Zaharia M., Franklin M., Shenker S., Stoica I., "Shark: SQL and Rich Analytics at Scale", ACM SIGMOD Conference, 2013.
13. Agrawal D., Bernstein P., Bertino E., "Challenges and Opportunities With big data A Community White Paper Developed by Leading Researchers Across the United States", 2012.

## BIOGRAPHY

**Azizkhan F Pathan** is a Research Assistant in the Computer Science and Engineering Department, Bapuji Institute of Engineering and Technology, Visvesvaraya Technological University. He is currently working as an Assistant Professor in Jain Institute of Technology, Davangere, Karnataka, India. He received Master of Technology in Computer Science and Engineering in 2014 and Bachelor of Engineering degree in 2011 from BIET, Davangere, Karnataka, India. His research interests are Big Data, Cloud Computing, and Data Mining etc.

**Dr. Chetana Prakash** is a Research Guide in the Computer Science and Engineering Department, Bapuji Institute of Engineering and Technology, Visvesvaraya Technological University. Currently she is working as Professor & HOD of MCA Department at Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India. She received Ph.D. in Computer Science and Engineering in 2013 from International Institute of Information Technology (IIIT), Hyderabad and received M.S degree in 1996 from BITS – PILANI, India. She got her Bachelor of Engineering degree in 1990 from UBDTCE, Davangere, Karnataka, India. She has 24 years of teaching experience. Her research interests are Computer Networks, Big Data, Cloud Computing, and Data Mining etc.