Advanced Computational Intelligence: An International Journal (ACII), Vol.3, No.4, October 2016

# BIG DATA SUMMARIZATION: FRAMEWORK, CHALLENGES AND POSSIBLE SOLUTIONS

Shilpa G. Kolte[1] and Jagdish W. Bakal[2]

[1]Research Scholar, University of Mumbai, India
[2]Principal & Professor in Computer Engineering,
Shivajirao S. Jondhale College of Engineering, Mumbai, India.

***ABSTRACT***

*In this paper, we first briefly review the concept of big data, including its definition, features, and value. We then present background technology for big data summarization brings to us. The objective of this paper is to discuss the big data summarization framework, challenges and possible solutions as well as methods of evaluation for big data summarization. Finally, we conclude the paper with a discussion of open problems and future directions..*

***KEYWORDS***

*Big data, Big data summarization, Data Generalization, Semantic term Identification*

## 1. INTRODUCTION

In recent years, with the rapid development of the Internet, network service has become one of the most frequently used computer applications. Search engine, webmail, and social network services are presently crucial data intensive applications. Because increasingly more people use web services, processing a big amount of data efficiently can be a significant problem. The very word "Big" indicates size. Big data have rapidly developed into a hotspot that attracts great attention to academia, industry, and even governments around the world [1-4].

Big data is of great value, which is beyond all doubt. From the perspective of the information industry, big data is a strong impetus to the next generation of IT industry, which is essentially built on the third platform, mainly referring to big data, cloud computing, mobile Internet, and social business. IDC predicted that by 2020 the market size of the third IT platform will reach US$ 5.3 trillion; and from 2013 to 2020, 90% of the growth in the IT industry would be driven by the third IT platform. From the socio-economic point of view, big data is the core connotation and critical support of the so-called second economy, a concept proposed by the American economist W.B. Arthur in 2011[5], which refers to the economic activities running on processors, connectors, sensors, and executors. It is estimated that at 2030 the size of the second economy will approach that of the first economy (namely, the traditional physical economy). The main support of the second economy is big data, as it is an inexhaustible and constantly enriching resource. In the future, by virtue of big data, the competence under the second economy will no longer be that of labor productivity but of knowledge productivity.

Compare with traditional datasets, big data typically includes masses of unstructured data that need more real-time analysis. In addition, big data also bring about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and also incurs new challenges, e.g. how effectively organized and manage such datasets. Recently, industries become interested in the high potential of big data, and many government agencies

announced major plans to accelerate big data research and applications [6]. Nowadays, big data related to the service of Internet companies grow rapidly. For example, Google processes data of hundreds of Petabyte (PB), Facebook generates log data of over 10 PB per month, Baidu, a Chinese company, processes data of 10 PB per day, and Taobao, a subsidiary of Alibaba, generates data of tens of Terabyte (TB) for online trading per day.

The features of big data can be characterized by 5V, namely, huge Volume, high Velocity, high Variety, low Veracity, and high Value. The main difficulty in coping with big data does not only lie in its huge volume, as we may alleviate to some extent this issue by reasonably expanding or extending our computing systems. Actually, the real challenges center on the diversified data types (Variety), timely response requirements (Velocity), and uncertainties in the data (Veracity). Because of the diversified data types, an application often needs to deal with not only traditional structured data, but also semi-structured or unstructured data (including text, images, video, and voice). Timely responses are also challenging because there may not be enough resources to collect, store, and process the big data within a reasonable amount of time.

Emplacing on the value addition that big data analysis is brings with it for identifying the problems and its solution to education level. It is imperative that educational institution start paying attention to both collection and analysis of data that has been collected over the past years. A big data analysis could help institutes, educator and policy makers get a better handle of not just how to teach but how students like to learn and other related issues plying in the mind of educator and policy makers like right from getting better attendance at low income group institution and controlling school dropout rates. Analysis of large data set can be given is insight on the students and their ability to learn as compared to many factors such as gender, socio-economic group, ethnicity and even location and culture. Big data analysis could help create the right incentives and educators, students, parents and other stakeholders to develop requisite skill in students.

The reminder of the paper is organized as follows. Section 2 presents background technology for big data summarization, Section 3 explore Big data summarization Framework, challenges and possible solution, Section 4 explains evaluation techniques for big data summarization and Section 5 concludes the paper.

## 2. TECHNOLOGY FOR BIG DATA

Big data is a new concept for handling massive data therefore the architectural description of this technology is very new. Apache Hadoop [7] is an open source framework of storing and processing large datasets using clusters of commodity hardware. Hadoop is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant. The various components of a Hadoop Stack are illustrated in Figure 1. The Hadoop platform contains the following important components.

### 2.1. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

HDFS is a fault tolerant and self-healing distributed file system designed to turn a cluster of industry standard servers into a massively scalable pool of storage. Developed specifically for large-scale data processing workloads where scalability, flexibility and throughput are critical, HDFS accepts data in any format regardless of schema, optimizes for high bandwidth streaming, and scales to proven deployments of 100PB and beyond. Data in HDFS is replicated across multiple nodes for the compute performance and for the reliable data protection [9].

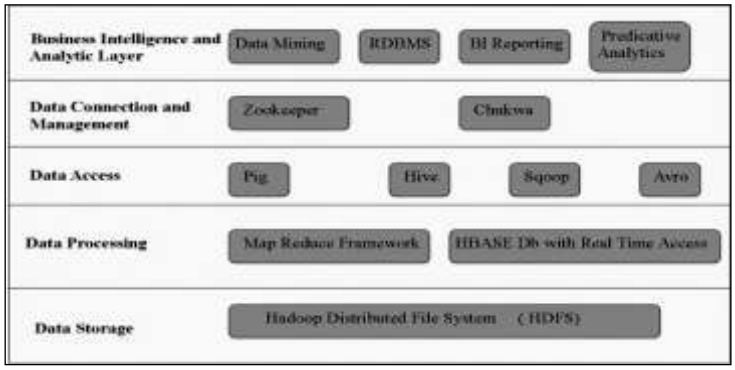


Figure 1. Hadoop Framework

## 2.2. DATA PROCESSING

In Hadoop framework following MapReduce and Hbase component are used for data processing: MapReduce is a parallel processing framework which is massively scalable, parallel processing frame work that works in tandem with HDFS. MapReduce processes exceedingly large amounts of data without being affected by traditional bottlenecks like network bandwidth by taking advantage of this data proximity. A MapReduce job usually consists of three phases—map, copy and reduce. The input data is split into chunks of 64MB size (by default). In the map phase, a user defined function operates on every chunk of input data producing intermediate key-value pairs which are stored on local disk. One map process is invoked to process one chunk of input data. In the copy phase, the intermediate key-value pairs are transferred to the location where a reduce process would operate on the intermediate data. In reduce phase, a user defined reduce function operates on the intermediate key value pairs and generates the output. One reduce process is invoked to process a range of keys [9].

HBase [10] is open source, distributed and Non relational database system implemented in Java. It runs above the layer of HDFS. It can serve the input and output for the Map Reduce in well mannered structure. It is a NoSQL database that runs on top of Hadoop as a distributed and scalable big data store. HBase can leverage the distributed processing paradigm of the HDFS. HBase allows you to query for individual records as well as derive aggregate analytic reports on a massive amount of data.

## 2.3. DATA ACCESS

A certain set of data access are currently being developed for MapReduce called as wrapper. These wrappers can provide a better control over the MapReduce code and aid in the source code development. The following wrappers are being widely used in combination with MapReduce.
Pig[11] is high-level platform where the Map Reduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high level language. Pig is a high level platform for creating MapReduce programs used with Hadoop and the language we use for this platform is called PIG Latin. The pig was designed to make Hadoop more user-friendly and approachable by power-users and non developers. PIG is an interactive execution environment supporting Pig Latin language.

Hive [12] is application developed for data warehouse that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing

conclusion, and analysis for respective queries. Hive looks very much like traditional database code with SQL access. Hive is read-based and therefore not appropriate for transaction processing that typically involves a high percentage of write operations.

Sqoop[13] is a command-line interface application that provides platform which is used for converting data from relational databases and Hadoop or vice versa. Apache Sqoop is a tool designed for transferring bulk data between Apache Hadoop and structured data stores such as relational databases or data warehouses. It can be used to populate tables in Hive and Hbase. Sqoop uses a connector based architecture which supports plugins that provide connectivity to external systems. Sqoop includes connectors for databases such as MySQL, Postgre SQL, Oracle, SQL Server and DB2 and generic JDBC connector. Dataset being transferred is sliced up into different partitions and a map-only job is launched with individual mappers responsible for transferring a slice of this dataset. Sqoop uses the database metadata to infer data types.

Avro[14] is a system that provides functionality of data serialization and service of data exchange. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records. Avro relies on schemas. When Avro data is read, the schema used when writing it is always present. This facilitates to use with dynamic, scripting languages, since data, together with its schema, is fully self-describing. When Avro data is stored in a file, its schema is stored with it, so that files may be processed later by any program.

### 2.4. DATA CONNECTION AND MANAGEMENT

Zookeeper and Chukwa are used for data connection and management to process and analyze the massive amount of logs. Zookeeper [15] is an open source Apache project that provides a centralized infrastructure and services that enable synchronization across a cluster. ZooKeeper maintains common objects needed in large cluster environments. Examples of these objects include configuration information, hierarchical naming space, and so on. Applications can leverage these services to coordinate distributed processing across large clusters. Chukwa[16] is an open source data collection system for monitoring large distributed systems. It is built on top of the Hadoop Distributed File System (HDFS) and Map/Reduce framework and inherits Hadoop's scalability and robustness. It also includes a flexible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.

### 2.5 BUSINESS INTELLIGENCE AND ANALYTICAL LAYER

The basic function of this layer is to analysed and predict patterns using SQL, data mining, and BI tools.

## 3. BIG DATA SUMMARIZATION FRAMEWORK, CHALLENGE, AND POSSIBLE SOLUTION

Big data not only refers to datasets that are large in size, but also covers datasets that are complex in structures, high dimensional, distributed, and heterogeneous. An effective framework when working with big data is through data summaries, such as data integration, data redundancy, and data compression, etc. Instead of operating on complex and large raw data directly, these tools enable the execution of various data analytics tasks through appropriate and carefully constructed summaries, which improve their efficiency and scalability.

In this section, we describe a framework of big data summarization is shown in the figure 2. We explore challenges and potential solutions to every specific step in the framework. The big data summarization is performed in four major stages. The challenges and solutions to each stage are described as follows:

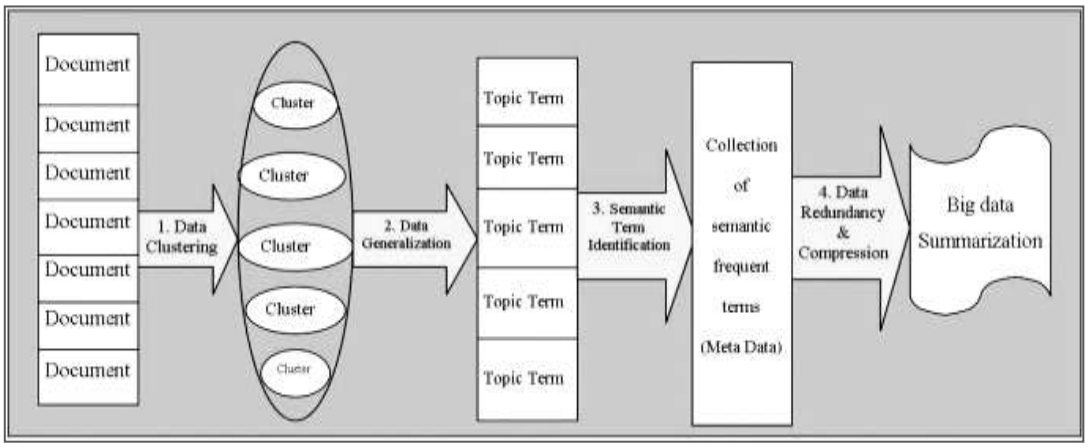


Figure 2. Framework of big data summarization

## 3.1 DATA CLUSTERING

The first stage is the document clustering stage where clustering technique is applied to the multi document data collection to create the document clusters. The purpose of this stage is to group the similar documents for making it ready for summarization and ensures that all the similar set of documents participates in a group of summarization process.

***Challenges:*** The issue of clustering or grouping n objects into K groups arises in many scientific applications, such as biology, education, genetics, criminology, etc [17], [18], [19], but obtaining an optimal grouping is not always an easy proposition. Investigating all possible partitions of a dataset is computationally infeasible for all but the most trivial cases, as the number of ways that a dataset can be partitioned into K groups increases exponentially with both n and K. A number of approaches [20], [21], [22], [23], [24], [25], [26], [27], [28] have been developed in this regard, but the simplicity and computational speed of the K-means algorithm [29], [30] has made it a popular choice. However, the algorithm needs initializing values which greatly influence its terminating optimal solution.

***Solutions:*** A promising solution to Big Data clustering is to provide a detailed assessment of the performance of many commonly-used and well-performing initializing methods over datasets of many dimensions, numbers of observations, groups and clustering complexity. Performance is evaluated both in terms of the ability to best recover the true groupings as well as the achieved minima of the objective function. Results from extensive investigations into factors that influence the performance of each algorithm are also provided.

## 3.2 DATA GENERALIZATION

Data generalization is process that abstracts a large set task relevant data from low conceptual to high conceptual level for better prediction and summarization. The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, and similarity and relevance judgments.

***Challenges:*** The first challenge is database Enrichment for the purpose of effective data generalization. The second challenge is the generalization process, if automatized, must be controlled by some decision system. The system usually makes a decision based on the set of attributes describing data being generalized. The challenge of selecting the right attributes to control generalization process may be as important as creation of decision system using those attributes.

***Solutions:*** To efficiently data generalization we need to choose or proposed suitable modeling technique is applied to each individual document cluster to generate the cluster topics and terms belonging to each cluster topic. A statistics-based approach for clustering documents and for extracting cluster topics can proposed. Relevant (meaningful) Expressions (REs) automatically extracted from corpora are used as clustering base features. These features are transformed and its number is strongly reduced in order to obtain a small set of document classification features. This is achieved on the basis of Principal Components Analysis. Model-Based Clustering Analysis finds the best number of clusters. Then, the most important REs are extracted from each cluster.

### 3.3 SEMANTIC TERM IDENTIFICATION (META DATA)

Efficient storage and retrieval of information is a growing problem in Big Data, particularly since very large-scale quantities of data such as text, image, video, and audio are being collected and made available across various domains, e.g. social networks, security systems, shopping and marketing systems, defense systems, fraud detection, and cyber traffic monitoring. Semantic term identification presents the data in a more efficient manner and makes it useful for a source of knowledge discovery and comprehension, for example by making search engines work more quickly and efficiently.

**Challenges:** Data representations play an important role in the indexing of data, for example by allowing data points/instances with relatively similar representations to be stored closer to one another in memory, aiding in efficient information retrieval. However, the high-level abstract data representations need to be meaningful and demonstrate relational and semantic association in order to actually confer a good semantic understanding and comprehension of the input.

**Solutions:** To enable the discovered patterns/results to be useful domain knowledge, the data analysts must provide process how global frequent terms are generated from the collection of multiple documents. For frequent terms generation of the multiple documents in each cluster, propose new process/algorithm based on the principle that data (words) that are used in the same contexts tend to have similar meanings. Design novel semantic indexing algorithm which has ability to extract the conceptual content of a body of text by establishing associations to those terms that occur to similar contexts.

### 3.4 DATA REDUNDANCY AND COMPRESSION

Data redundancy and compression aims to reduce redundant information in data to save the consumption of resources such as storage space or I/O bandwidth, and therefore this technique has very important application in the areas of data summarization. Apart from the space savings, data compression provides another benefit that compressed data could be stored and transmitted faster, thereby improving the performance of I/O intensive workloads. However, compression is a compute-intensive operation and imposes an extra processing cost and resource contention on compute node.

**Challenges:** Data compression algorithms involve trade-offs among various factors, including the degree of compression, the amount of distortion introduced (using in lossy data compression), and the computational resources required to compress and decompress the data.

**Solution:** To facilitate the approximate duplicate data is identified and removed from the summary report and final summary data is generated, the Lossless data compression algorithm is provided. Lossless data compression algorithms usually exploit statistical redundancy to represent data more concisely without losing in-formation, and can reconstruct the original data exactly from the compressed data. There are two key metrics – compression ratio and (de)compression speed that are used to evaluate the performance of lossless data compression algorithms. Different compression algorithm has different compression ratio. Even the same compression algorithm could have varied compression ratio of different original data formats. In general, the compression ratio of text format data is higher than binary format, and the compression ratio of repetitive data is higher than random data.

## 4. EVALUATION OF DATA

Data summarization process is evaluated using performance parameters namely, Compression Ratio (CR) and Retention Ratio (RR), and precision, recall and F-score.

### 4.1 COMPRESSION AND RETENTION RATIO

The Compression Ratio (CR) is the ratio of size of the summarized data document to the total size of the original data documents. Retention Ratio (RR) is the ratio of the information available in the summarized document to the information available in the original data collections. The expressions of calculating the CR and RR are given in the equation (1) and equation (2) respectively.

$$CR = \frac{|d|}{|D|} \tag{1}$$

$$RR = \frac{Info(d)}{Info(D)} \tag{2}$$

Where *d* represents the size of the summarized data is document and *D* is the total size of the original data collection. *Info(d)* represents the information available in the summarized data document and *Info*(*D*) is the information present in the original data collection.

### 4.2 PRECISION, RECALL AND F-SCORE

The main evaluation metrics of co-selection is precision, recall and F-score. Precision (P) is the number of data occurring in both system and ideal summaries divided by the number of sentences in the system summary. Recall (R) is the number of data occurring in both system and ideal summaries divided by the number of sentences in the ideal summary. F-score is a composite measure that combines precision and recall. The basic way how to compute the F-score is to count a harmonic average of precision and recall:

$$F = \frac{2.P.R}{P+R} \tag{3}$$

## 5. CONCLUSION

In this paper, we have proposed an overall framework of big data summarization which combines data clustering, data generalization, semantic data indexing and data compression. The proposed framework definitely performs better than existing systems. We have explored challenges and possible solution for each steps of data summarization which alleviates scalability and reduce data redundancy and data compression for better big data summarization. Also we have discussed big data summarization technology and the evaluation technique using compression Ratio (CR) and retention Ratio (RR) with F-score.

Future work in this direction can be providing the novel clustering algorithm and semantic data indexing and data compression techniques for big data summarization over the MapReduce framework.

## ACKNOWLEDGEMENTS

We would like to thank management of Sarswati College of Engineering, Kharghar, Navi Mumbai, India and Principal, Terna Institute of Technology, Nerul, Navi Mumbai, India for providing the infrastructure to carry out the proposed research work

## REFERENCES

[1] V. Mayer-Schonberger, K. Cukier (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt.
[2] R. Thomson, C. Lebiere, S. Bennati (2014)cHuman, model and machine: a complementary approach to big data, in: Proceedings of the 2014 Workshop on Human Centered Big Data Research, HCBDR '14.
[3] A. Cuzzocrea (2014) Privacy and security of big data: current challenges and future research perspectives, in: Proceedings of the First International Workshop on Privacy and Security of Big Data, PSBD '14.
[4] XiaolongJina, Benjamin W, Xueqi Chenga, and YuanzhuoWang (2015) Significance and Challenges of Big Data Research.
[5] W.B. Arthur, (2011) http://www.images-et-reseaux.com/sites/ /medias/blog/2011/12/the-2nd-economy.pdf.
[6] Federal government (2012). http://www.whitehouse.gov/sites/big data
[7] Borthakur, D. (2007) The hadoop distributed file system: Architecture and design. Hadoop Project Website (Available online at - https://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf).
[8] Borthakur D (2008) HDFS architecture guide. HADOOP APACHE PROJECT.
[9] K V N Rajesh (2009)Big Data Analytics: Applications and Benefits –
[10] J. C. Brian Dolan and J. Cohen (2009) MAD Skills: New Analysis Practices for Big Data," in Very Large DAta Bases(VLDB) 09. Lyon, France: ACM.
[11] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins (2008) SIGMOD international conference on Management of data, pp. 1099-1110. ACM.
[12] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, & Raghotham Murthy (2009) Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment 2, no. 2 .
[13] http://www.tutorialspoint.com/sqoop/sqoop.pdf
[14] http://avro.apache.org/avro.pdf
[15] J. Baker, C. Bond, J. Corbett, J. Furman, Lloyd, and V. Yushprakh (2011) Megastore: providing scalable, highly available storage for interactive services. In Proceedings of Conference on Innovative Data Systems Research.
[16] J. Boulon, A. Konwinski, R. Qi, A. Rabkin, E. Yang, and M. Yang. Chukwa (2008) A large-scale monitoring system. In First Workshop on Cloud Computing and its Applications (CCA '08), Chicago.

[17] J. Quackenbush (2001) Computational genetics: Computational analysis of microarry data," Nature Reviews Genetics, vol. 2, pp. 418–427.
[18] P. S. Bradley, K. P. Bennett, and A. Demiriz (2000) Constrained k-means clustering, Microsoft Research, Tech.
[19] W. Y. Arms and C. R. Arms (1978) Cluster analysis used on social science journal citations, Journal of Documentation, vol. 34.
[20] J. A. Hartigan, (1985) Statistical theory in clustering," Journal of Classification, vol. 2, pp. 63–76.
[21] D. B. Ramey (1985) Nonparametric clustering techniques, in Encyclopedia of Statistical Science. New York: Wiley, 1985, vol. 6, pp. 318–319.
[22] L. Kaufman and P. J. Rousseuw (1990) Finding Groups in Data. New York: John Wiley and Sons.
[23] B. S. Everitt, S. Landau, and M. Leesem (2001) Cluster Analysis (4th ed.). New York: Hodder Arnold.
[24] G. J. McLachlan and K. E. Basford (1988) Mixture Models: Inference and Applications to Clustering. New York: Marcel Dekker.
[25] C. Fraley and A. E. Raftery (2002) Model-based clustering, discriminant analysis, and density estimation, Journal of the American Statistical Association, vol. 97, pp. 611–631.
[26] J. R. Kettenring (2006) The practice of cluster analysis, Journal of classification, vol. 23, pp. 3–30.
[27] J. A. Hartigan and M. A. Wong (1979) A k-means clustering algorithm, Applied Statistics, vol. 28, pp. 100–108.
[28] J. MacQueen (1967) Some methods of classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297.
[29] R. Maitra (2009) Initializing partition-optimization algorithms, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 6, pp. 144–157.
[30] D. Steinley (2006) K-means clustering: A half-century synthesis, The British Psychological Society, vol. 59, pp. 1–34.

## Authors

Ms. Shilpa G. Kolte is a PhD student at University of Mumbai, India. Currently, she is working in big data mining. She received his Graduate degree in Computer Engineering from Dr. BAMU, Aurangabad, India in 2001and Post Graduate in Computer Engineering from University of Mumbai, India in 2008.

Mr. Jagdish W. Bakal is working as a Professor in Computer Engineering and Principal at Shivajirao S. Jondhale College of Engineering, Dombivli, Mumbai, India. He received his PhD from Bharati Vidyapeeth, Pune, India in 2009. He has published more than 90 research papers in the field of mobile networks, big data mining, and hybrid computing systems in the reputed journals and conferences.