A New Approach to Most Advanced Technique for Data Mining of the Big Data

Pooja Sharma
Computer Science. Dept. of Computer Science, MUIT University, Lucknow

Dr. B.D. K. Patro
Associate Professor, Dept. of Computer Science, MUIT University, Lucknow

DOI Number: 10.14704/nq.2022.20.9.NQ44529 **Neuro Quantology 2022; 20(9):4598-4603**

Introduction

Big data is collection of very large-scale data which is generated due to usage of social networking sites, business transactions, scientific experiments, sensor networks and sharing of resources etc. Big data is dataset that has high volume and complex variety whichare difficult to handle for traditional softwares. Gartner defined 3Vs-Velocity,Variety, and Volume for Big Data**(**Chen, Mao and Liu, 2014). Velocity is the rate of data flow for any process,the data rate is very fast for Big data, i.e. data sharing is at a rapid rate.Data increases at a very high pace such as streaming data generated in weather reports and stock prices. Variety is collection of different types of data – structured, semi-structured and unstructured.Structured data is the type which is generated by simple mode such as business transactions. The format used for this type of data is structured, e.g. relational data. Semi-structured data is structured and unstructured format such as Internet log files. Unstructured data is a type which has no structured format such as images and videos.Volume is large-scale data which is beyond the storage and processing capabilities of existing systems.The reduced cost of computation results in easier collection and storage of data (Bajari et al., 2019). According to International Data Corporation (IDC), data volume was 1.8ZB, which was nine time increase in five years (Chen, Mao and Liu, 2014). Some researchers also proposed veracity and value to characterize Big data in 5Vs(Lomotey and Deters, 2014). Veracity is the noise in data which is not in capability of traditional softwares to analyze. Data integrity is a concern for decision making due to availability of noise. For example, Social networks posts which are in the unstructured format are not reliable. Veracity characteristic is of the highest concern for data processing (Uddin and Gupta, 2014).Value is the addition in the decision-making process which is also one of the most essential Big data characteristics.The need to leverage volume, velocity and variety results into novel techniques of data storage and visualization (Mikalef et al., 2018).

Big data is an active research area in recent years and a lot of innovations and developments are being carried out by researchers and practitioners. Big data is defined as, "datasets which could not be captured, managed and processed by general computers within an acceptable scope" by Apache Hadoop in 2010. In 2011, Mckinsey described in the report that Big data is the next active topic for innovation and development. Industrial Development Corporation(IDC) and EMC Corporationconfirm that data generated in 2020 will be 44 times greater than in 2009.Big data is of interest in academics and industry due to its applications in data mining,information retrieval, social network analysis,opinion mining, and sentiment analysis, etc. Database was invented in 1970 which could store and process the data. As data volume increased, "share- nothing" architecture was defined which worked in parallel to store and process the data. With the development of social networking, sensors and users interactions, there was a need for much better techniques to deal with such large-scale data. Novel techniques were invented to deal with large-scale data

which was termed as "Big Data". Application areas of Big data are – Manufacturing, Healthcare, Education, Media, and Business Management. Social Network analysis, Sentiment analysis, Business decision making, Social recommendation, Machine learning, Cloud computing, and IoT are the related techniques where big data is strongly correlated. Big data is also applied in ICT domain for online learning (Huda et al., 2018).

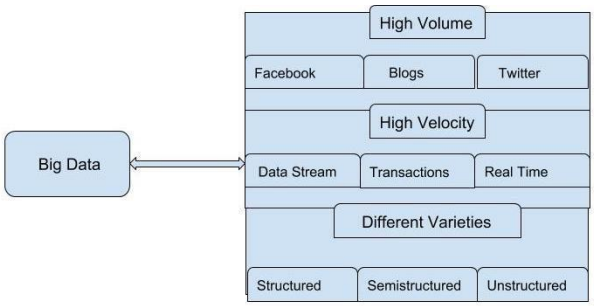

Figure 1.1: 3Vs of Big data(Chen, Mao and Liu, 2014)

In Figure 1.1, high volume, high variety and high velocity of Big data are depicted.These Big data characteristics are available in every technique which generates large-scale data.

Big data analysis fields are as follows **(**Chen, Mao and Liu, 2014)**:**

(i) Structured data

(ii) Text data

(iii) Web data

(iv) Multimedia data

(v) Network data

(vi) Mobile data

Big data value chain is – generation, acquisition, storage and analysis of data as defined by **(**Chen, Mao and Liu, 2014). Data acquisition, data analysis, data storage, and data usage are essential phases of Big Data mining(Strohbach et al., 2016), but Big data mining and analysis requires novel and improved techniques so that mixed types of data can be processed efficiently and effectively.Cluster, factor, regression, crowd-sourcing, and classification analysis are the main methods used in Big data **(**Chen, Mao and Liu, 2014).

Review of Literature

(Chen, Mao and Liu, 2014)	
Description/Research contributions	• In this paper, state-of-the-art approaches of Big data are studied. Techniques which are used with Big data such as Machine learning and Cloud computing are analyzed in detail. • Value chain of Big data – generation of data, data acquisition, storage of data and analysis of data, are explained. • Several definitions of Big data are given from Apache Hadoop, Gartner, NIST, Mckinsey, Gartner. It is mentioned that companies such as Google, Facebook, Amazon etc. have started Big data project. • Storage mechanism for Big data is demonstrated which covers GFS (Google File System) and HDFS (Hadoop Distributed File Systems).
Limitations/Future	• Higher diversity and large-scale data requires big

Need for Study

Big data is large-scale data generated due to business transactions, sensor networks and social networking sites interactions. Large-scale data cannot be stored and processed by traditional tools and techniques. There is need of novel approaches which can deal with large-scale data with high efficiency. In this thesis, storage techniques of Big data are studied extensively. It is observed that traditional

relational data storage is not sufficient enough for storage of large-scale heterogeneous data. NoSQL data stores are used for Big data storage. Column-oriented, Graph- based, Key-value and Document-based storage techniques are the main storage techniques. There are approximately 120 real solutions of Big data storage. Several research works have been carried out for users to select the most appropriate data storage techniques. The limitation of these existing research works is comparing these storage techniques based on real solution which is not relevant for users. In this thesis, main storage architectures are compared based on their advantages and disadvantages. It is proposed that if user selects storage technique, then it is easy to select real solution amongst the storage technique.

Traditional data mining algorithms extract patterns from data and provide information that can be used for analysis. K-means, K-prototype, SVM and Naïve Bayes algorithms are studied in this thesis. The limitations of these algorithms are that these algorithms can be efficient only for small-scale data. When these algorithms are deployed for Big data, response time and throughput degrades significantly. There is need of efficient approach which can process heterogenous data in less response time. Limitation of K-means is that it can work well for numerical data. Data which are commonly used are mixed- numerical as well as categorical data. K-prototype is used for processing numerical and categorical data. In this thesis, K-prototype is implemented on MapReduce to process Big data. Intelligent splitter is proposed in this thesis, which divides numerical and categorical data before sending data to Mappers and Reducers. Experiment analysis proves that response time and accuracy is improved by using proposed approach.

Research Gaps

Several research works have been studied and analyzed in this thesis. The main research gaps found in existing works are as follows:

1. NoSQL data storage for Big data are compared based on real solutions. Comparative studies of existing literature do not provide clear applications of these solutions. There is need of better comparative study which should be focused on broad categories i.e. column-oriented, document-based, key-value and graph-based.
2. Traditional data mining algorithms are suitable for numerical data. Semi-structured, unstructured and categorical data cannot be processed efficiently with conventional algorithms.
3. Large-scale data cannot be analyzed and processed efficiently by conventional mining algorithms. There is need for novel Big data techniques which can address scalability issue. Moreover, Big data technologies such as Hadoop, MapReduce, Pregel, Giraph and Tensorflow should be used to implement improved algorithm.
4. Sparsity, cold start and scalability are issues in social recommendation which should be addressed.
5. Big social graph cannot be processed on centralized systems. Partitioning is necessary for processing subgraphs on clusters. Degraded locality is also important disadvantage of existing work.
6. Prediction of ratings in social recommendation needs improvement. Deep learning models can be used to address this issue.

Activation functions in deep learning models are not identified which is best suitable for deep learning based recommendation.

Objectives of Study

1. Study of various existing storage structures of big data- Column oriented, graph, Document based and Key Value
2. Comparative analysis of various conventional data mining algorithms
3. To propose and design an efficient technique for mining of Big data
4. Testing and validation of the proposed methodology

Research Methodology

In this research work, several research works are studied and analyzed to observe the limitations in traditional data mining and social recommendation. Traditional data mining algorithms are compared based on response time. K-prototype algorithm is

implemented on MapReduce to process numerical as well as categorical data. Furthermore, social recommendation is improved by using hyperedge and transitive closure. Trust-based partitioning is used for large-scale social recommendation. Moreover, deep learning is applied to improve recommendation accuracy. These improvements require suitable research methodology which is described in next subsections. Research methodology is categorized based on objectives of study.

Research methodology for objective 1

Objective 1 is to study and compare Big data storage techniques – column-oriented, document- based, graph-based and key-value. The first step of our research methodology is to study existing literature work and observe their advantages and limitations. The conclusion after comparing several research works is that researchers have utilized distributed approach for storing Big data. NewSQL and NoSQL are data storage techniques which are proposed by researchers. The limitations of existing literature work are comparative study and analysis is based on real solutions of NoSQL data storage techniques which form the basis of our research methodology. The features such as format, storage, flexibility, scalability and complexity are used in our research work to compare storage techniques. The reason for selecting this research methodology is that readers can easily compare broad categories and select particular storage techniques based on their specific application. The next step of research methodology is to select most significant real solutions out of 120 solutions. The reason for selecting these solutions is that advantages and limitations of these solutions for particular application are concluded. The conclusion from research methodology of this objective is that graph-based storage is best suitable for social big data and will be used in research methodology for proposed approach.

Research methodology for objective 2

Objective 2 is to compare and analyze conventional data mining algorithms. The first step used for research methodology for this objective is to study several research papers which cover traditional data mining algorithms. It is observed that researchers have mentioned that issues for these algorithms are scalability and sparsity. These limitations form the basis of research methodology of our work. The next step for research methodology is to select data mining algorithms which cover traditional data mining features. K-means, Support Vector Machine and Naïve Bayes algorithms are selected as these algorithms cover classification and clustering. The next step of research methodology is to collect data so that these algorithms can be analyzed. Standard datasets Iris (W1), College (W2) and Labour (W2), Weather (W3) and Supermarket (W3) available in Weka library are used. The reason for selecting these datasets is that different classes are available for classification and similar data samples are available for clustering. The next step of research methodology is to analyze data. Quantitative approach i.e. CPU time taken by these algorithms are used for analyzing and comparing these algorithms. Moreover, it is concluded from this step that K-means cannot be implemented for categorical data efficiently. This limitation forms the basis of next step of our research methodology i.e. implementing K- prototype algorithm to process numerical as well as categorical data. In K-prototype algorithm, Euclidean distance is used for numerical data and Hamming distance is used for categorical data. Chess (W4) dataset is used for implementation. The reason for selecting this dataset is that numerical as well as categorical data are available in this dataset.

Research methodology for objective 3

Objective 3 is to propose technique for efficient mining of Big data. The first step of research methodology for this objective is to understand the limitations of traditional data mining. It is concluded from research methodology for previous objective that scalability and sparsity is the main issues which needs to be addressed. Moreover, traditional techniques work efficiently for numerical data. These limitations form the basis for research methodology. The next step is to address these issues by using improved Big mining techniques and technologies.K-prototype algorithm is utilized to process numerical as well as categorical data. Chess (W4) dataset is used for implementation. The reason for selecting this dataset is that numerical as well as categorical data are available in this dataset. Furthermore,

K-Prototype is also implemented on MapReduce to observe response time when it is deployed on multiple clusters. Intelligent splitter is proposed which splits numerical and categorical data before sending data to MapReduce. This step of research methodology addressed numerical data only and scalability issue. Sparsity is the main issue in social Big data. In social recommendation, it is observed that many users do not provide ratings to products which results in sparse matrix. This forms the basis of next step to address sparsity issue. Graph-based approach is used as it is best suitable for social Big data. Qualitative approach i.e. trust amongst users is selected for research methodology. If users trust each other, there is more probability for same ratings for a product. Transitive closure and hyperedge are proposed in this step to improve trust amongst users. Figure 3.1 demonstrates flow of proposed approach research methodology. The next step is to collect dataset which contains trust and ratings values. Epinions and FilmTrust datasets are selected. The reason for selecting this dataset is that sparse entries for trust and ratings exist in this dataset. Mean Absolute Error and Root Mean Square Error are selected as metric for quantitative data analysis. Moreover, to address scalability LiveJournal datasets is used. The reason for selecting this dataset is that it contains large numbers of links amongst users as nodes. These nodes are represented in the form of graphs and trust values amongst them are represented as edges. Large-scale graph is partitioned using trust-based partitioning instead of random partitioning to improve scalability. The next step of research methodology for this objective is to further improve recommendation accuracy by using deep learning. Autoencoder is used for improving recommendation as it reduces large- scale dimensions.

Research methodology for objective 4

Objective 4 is to validate the efficiency of proposed approach. The first step of research methodology for this objective is to select technologies which are most suitable for proposed approach. K-Prototype is to implement on distributed nodes to improve response time. MapReduce is used to implement this algorithm which can run in parallel using mapper and reducer files. The next step is to validate proposed approach. Chess (W4) dataset is used for implementation. Response time is calculated for this algorithm on multiple nodes to validate that response time is reduced. The next step is to collect dataset which contains trust and ratings values. Epinions (W5) and Film Trust (W6) datasets are selected. The reason for selecting these datasets is that sparse entries for trust and ratings exist in this dataset. The next step is to select technologies for proposed hyperedge approach for social recommendation. SNAP (W8) library is used as it can manipulate graph easily. MAE and RMSE evaluation metrics are used for validating social recommendation accuracy and compared with state-of-the-art approaches. Live Journal (W7) dataset, Pregel and Giraph (W9) are used for implementing large-scale graph partitioning approach. Locality is used as evaluation metrics in this step of research methodology to validate better partitioning approach. The next step is to validate deep learning based recommendation. Tensor Flow (W10) is used for implementing recommender system using Auto Encoder model. Proposed shared layer approach is compared with existing approaches to validate that recommendation accuracy is improved.

Limitations of study

1. In this research work, NoSQL data storage techniques are compared based on broad categories- column-oriented, document-based, graph-based and key-value. There are approximately 120 real solutions of NoSQL data storage. In this research work, approximately 10 real solutions are covered for study. Moreover, standard query language for NoSQL data storage is not available. Different real solutions use their specific query language. The limitation of this study is that query languages could not be standardized.

2. Traditional data mining algorithms are compared based on response time. K-means, SVM and Naïve Bayes are analyzed in this research work. Other classification, clustering and association rule mining algorithms could not be compared. Moreover, evaluation metrics such as precision, recall, F-measure etc. could not be used as evaluation metrics in this research work.

3. Social recommendation based on trust is

proposed in this research work. Moreover, large-scale social graphs are partitioned using trust-based approach. Other recommendation techniques such as group recommendation could not be analyzed.

4. Deep learning is applied on social recommendation to improve recommendation accuracy. Auto Encoder is deployed on user-item and user-user trust ratings which extracts hidden information from sparse matrices. Other deep learning models such as CNN, RNN etc. could not be deployed. Moreover, activation function such as tanh and ReLu could not be used in this research work.

Future Directions

- NoSQL data storage query languages are not standardized. Different data solutions use their own query languages. Extensive research is required to propose standard query language for NoSQL.
- A lot of further improvements are required in mining unstructured data with novel Big data technologies.
- Social recommendation should be analyzed on different large-scale datasets to evaluate accuracy.
- Precision, recall etc. can be used in addition to MAE and RMSE evaluation metrics.
- Other deep learning models such as multilayer perceptron, deep belief network etc. can be deployed to analyze Big data analytics improvement.
- Other activation functions such as tanh and ReLu can be deployed to evaluate recommendation accuracy.

Conclusion

This research work is concluded with future directions to provide details about possibility for further enhancement. The advantages, limitations, experiment findings and analysis are concluded by the use of categorizing this research work into NoSQL data storage, traditional data mining, Big data mining, trust improvement and improved recommendation accuracy by using proposed approach.

Reference: -

1. R. Agrawai, M. Mehta, J. Shafer, R. Srikant, A. Arning, and T. Bollinger. The Quest Data Mining System, In Proceedings of the 2nd international Conference Knowledge Discovery and Data Mining, (KDD), 1996.
2. R, Agrawai and G. Psaila. Active Data Mining. In Proceedings of the 1st International Conference Knowledge Discovery in Data Mining, AAAI, pages (3-8), 1995.
3. R, Agrawai and R. Srikant. Fast algorithm for mining association rules. In Proceedings of the 20th Internationa! Conference on Very Large Data Bases, Santiago de Chile, Chile, pages (487-499), 1994.