© INNOVATIVE RESEARCH THOUGHTS
Refereed | Peer Reviewed | Indexed
ISSN : 2454 – 308X | Volume : 09 , Issue : 01 | January - March 2023

Comparative Analysis of Traditional Mining with Big Data Mining

Dr. Sushil Kumar Sharma[1], Dr. Naveen Verma[2]
Dr. B.R. Ambedkar Govt. College, Kaithal
E-mail: shilu_online@rediffmail.com, drnaveengovtclg@gmail.com

Abstract

In this research paper, Big data storage, traditional data mining, Big data mining, Social recommendation, Scalable social recommendation and social recommendation improvement using deep learning are analyzed with key stress on comparaing Traditional Mining with Big Data Mining. The issues such as scalability, sparsity, cold start, data variety etc. are addressed by proposed approaches.

Introduction

In this research work, broad categories of NoSQL data storage i.e. Column-oriented, Document-based, Graph-based and Key-value are compared. Several research works have compared real solutions of NoSQL data storage such as Cassandra, MongoDB, Neo4J or DynamoDB. Real solutions are compared based on very narrow features. In this research work, broad NoSQL storage categories are compared based on various characteristics such as storage, applicationarea, advantages, flexibility, performance and scalability. The reason for selecting these real solutions is to compare based on scalability, storage and applications. It is observed that query languages for NoSQL are not standardized due to different storage structures of NoSQL data storage techniques. Different data storage techniques use its own query languages.

Overview of Techniques

- Big data storage techniques are studied and analyzed from existing literature work.Several researchers have concluded that distributed storage, NewSQL and NoSQL data storage techniques are used for Big data storage.
- NoSQL data storage techniques- Column-oriented, Document-Based, Graph-Based and

Key-value are studied and compared in this research work.

- The broad categories of NoSQL data storage are compared which is different from existing research work in which only real solutions are compared with narrow features.
- The most commonly used real solution such as MongoDB, CouchDB, BigTable, HBase etc. are selected to compare these solutions based on features such as scalability, applications etc.
- Traditional data storage techniques and Big data storage techniques are also compared. It is concluded that Big data storage techniques outperforms traditional techniques in terms of scalability, Object relational mapping, schema etc.
- It is also concluded that standard query language is not available for NoSQL data storage. There is need for standardizing query language for Big data storage.

Analysis of traditional data mining techniques

In this research work, ten most popular data mining algorithms – C4.5, K-means, SVM, Apriori, EM, PageRank, Ada Boost, kNN, Naïve Bayes and CART are studied. K-means, SVM and Naïve Bayes are algorithms which are selected for comparative analysis. These three algorithms are compared based on various features such as distance measures, application area, limitations and geometric plane. K-means is implemented using Weka library and deployed on Iris, College and Labour datasets. It is analyzed that as the numbers of instances in datasets increase, response time increases linearly. SVM and Naïve Bayes algorithms are compared based on response time. These algorithms are implemented in Weka using Iris, Weather and SuperMarket datasets available in Weka library. It is analyzed that Naïve bayes response time is less as compared to SVM.

Big data mining

In this research work, it is observed that K-means can perform well for numerical data. Big data is collection of numerical as well as categorical data. K-means is not efficient for categorical dataas it is based on distance measure using geometric space. K-prototype algorithm is implemented in the research work, which uses Euclidean distance for numerical and Hamming distance for categorical data. Intelligent splitter is proposed in this research work which splits numerical and

categorical data before sending data to Mapper and Reducer. Approximately linear speedup is achieved using proposed approach.

The conclusion for this category of research work is as follows:

- K-Prototype algorithm is implemented on MapReduce which overcomes the limitationsof K-means which can work efficiently for numerical data. It is concluded that K-Prototype response time on Mapreduce reduces significantly when deployed on multiple clusters. Speedup is calculated as 1 for K-Prototype on 1 cluster, 2.8 for 3 clusters and 4.6for 5 clusters. Linear speedup is not achieved on multiple clusters. The reason is time is devoted to communication cost amongst clusters.
- Traditional data mining and Big data mining are compared based on several features suchas scalability, technologies, structure of data etc. It is concluded that due to unstructured, categorical and large-dimensional data generated due to social networking sites, business transactions etc., there is need for Big data technologies and techniques to extract pattern from dataset.

Trust improvement in Social Recommendation

In this research work, sparsity, cold-start and scalability issues of social recommendation are addressed. Sparsity and cold start are removed by using proposed approach IPG (Influence Product Graph). Direct trust as well as indirect trust is used in this thesis. Hyperedge and transitive closure are used for proposed approach which enhances trust values amongst users. This proposed approach is implemented on Epinions and FilmTrust datasets. Trust is improved by 1.091% and 0.81% for Epinions and FilmTrust dataset respectively. Mean Absloute Error andRoot Mean Square Error are improved significantly by using proposed approach.

Conclusion

- Several research works are studied and analyzed for social recommendation. Researchers have concluded that sparsity, cold start and scalability are the issues in recommendation.
- In this research work, direct as well as indirect trust values are utilized for improving recommendation accuracy.

- Hyper edge and transitive closure are used to improve trust values which improve recommendation accuracy.
- Epinions and FilmTrust datasets are used for experiment analysis. These datasets are selected because trust and ratings values are in synchronization and also entries are sparse.
- Mean Absolute Error and Root Mean Square Error are used for evaluating difference in predicted ratings and actual ratings.
- Experiment analysis proves that MAE and RMSE are improved by using proposedapproach.

Future Scope

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.

References

[1] L. Duan and Y. Xiong, "Big data analytics and business analytics," *Journal of Management Analytics*, vol. 2, no. 1, pp. 1–21, 2015.

[2] H. N. Rothberg and G. S. Erickson, "Big data systems: knowledge transfer or intelligence insights?," *Journal of Knowledge Management*, vol. 21, no. 1, pp. 92-112, 2017.

[3] G. Chetty and M. Yamin, "A distributed smart fusion framework based on hard and soft sensors," *International Journal of Information Technology*, vol. 9, no. 1, pp. 19– 31, 2017.

[4] D. Che, M. Safran, and Z. Peng, "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities," in *the Proceedings of International Conference* on *Database Systems for Advanced Applications*, 2013, pp. 1–15.

[5] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[6] S. Schneeweiss, "Learning from Big Health Care Data," *Perspective*, vol. 363, no. 1, pp. 1–

3, 2010.

[7] U. Kang and C. Faloutsos, "Big Graph Mining : Algorithms and Discoveries," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 29–36, 2013.

[8] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143-148, 2016.

[9] S. Moens, E. Aksehirli, and B. Goethals, "Frequent Itemset Mining for Big Data," in *the Proceedings of IEEE International Conference on Big Data*, 2013, pp. 111–118.