

Real World Application of Big Data in Data Mining Tools

J Uma Mahesh¹ Madan Mohan R²

¹Student of M.Tech ²Assistant Professor (PhD)

^{1,2}Department of Computer Science & Engineering

¹Bharat Institute of Engineering & Technology ²Ibrahimpatnam, Hyderabad

Abstract—The main aim of this paper is to make a study on the notion Big data and its application in data mining tools like R, Weka, Rapidminer, Knime, Mahout and etc. We are awash in a flood of data today. In a broad range of application areas, data is being collected at unmatched scale. Decisions that previously were based on surmise, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. The paper mainly focuses different types of data mining tools and its usage in big data in knowledge discovery.

Key words: Big data, Data Mining, Mining Tools, and Data streams, Statistics, summarization

I. INTRODUCTION

Big data is “Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is “big data.”

Massive volume of both structured and unstructured data from various sources such as social data, machine generated data, traditional enterprise which is so large that it is difficult to process with traditional database and software techniques. Big Data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

A. Big Data Spans Five Dimensions:

1) Volume:

- Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.
- Turn 12 terabytes of Tweets created each day into improved product sentiment analysis Convert 350 billion annual meter readings to better predict power consumption.

2) Velocity:

- Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
- Scrutinize 5 million trade events created each day to identify potential fraud.
- Analyze 500 million daily call detail records in real-time to predict customer churn faster.
- The latest I have heard is 10 nano seconds delay is too much.

3) Variety:

- Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.
- Monitor 100’s of live video feeds from surveillance cameras to target points of interest.
- Exploit the 80% data growth in images, video and documents to improve customer.

4) Value:

- Importance of Analysis, which was previously limited by technology.

5) Visibility:

- Access from disparate geographic location. There are different types of data such as relational, structural, textual, semi structured, graph data, streaming data etc can be included in big data. These data can be used for summarization and Statistics in Data warehouse and OLAP, Indexing, Searching, and Querying, Keyword based searching, Pattern matching (XML/RDF), Knowledge discovery in Data Mining and Statistical Modeling.

– BIG DATA is not just HADOOP



II. CHALLENGES AND USE CASES OF BIG DATA

The challenges in handling big data includes in technology. The technology needs new architecture, algorithms, techniques for its implementation. It also requires technical skills .So experts are needed for this new technology to deal with big data. The correction and correlation of data makes more complexity.

The Major Use cases of Big Data in the real world data:

A. Big Data Exploration:

For a variety of reasons, data exploration is an important path to gaining business value from all kinds of data, from traditional enterprise data sources to big data and streaming machine data to take a decision.

B. View Of The Customer:

Analytics: Extend customer views by Listen, learn, and execute to effectively use big data in customer analytics. Gain a full understanding of customers—what makes them tick, why they buy, how they prefer to shop, why they switch, and others.

C. Big Data Security Intelligence:

Given the very large datasets that contribute to a Big Data implementations, there is a virtual certainty that either protected information or critical Intellectual Property (IP) will be present. This information is distributed throughout the Big Data implementation as needed – with the result that the entire data storage layer needs security protection. and sources of under-leveraged data to significantly improve intelligence, security and law enforcement sight.

D. Operations Analysis:

By combination of Big Data and advanced analytics in Exploration and Development activities, managers and experts can perform strategic and operational decision-making. The areas where the analytics tools associated with Big Data exploration include: Analyze a variety of machine and operational data for improved business results.. By using big data for operations analysis, organizations can gain real-time visibility into operations, customer experience, transactions and behavior.

E. Data Warehouse:

Advancement: Data warehouse stores data with four terms subject oriented, Integrated, Time Variant and Non Volatile. Optimize your data warehouse to enable new types of analysis. Use big data technologies to set up a staging area or landing zone for your new data before determining what data should be moved to the data warehouse. Remove infrequently accessed or aged data from warehouse and application databases using information integration software and tools.

III. APPLICATION OF BIG DATA IN DATA MINING

In data mining a number of different data repositories can be involved. Data mining is s a tool or technique used to extract the knowledge or Extraction of implicit, previously unknown and unexpected, potentially extremely useful information from data from data repositories. The challenges and techniques of mining may differ for each of the repository systems.

Advanced databases or information repositories require sophisticated facilities to efficiently store retrieve and update large amounts of complex data. They also provide fertile grounds to raise many challenging research and implementation issue for data mining

For data mining in object relational database system, techniques need to be developed for handling complex object Structures, generalization, specialization class hierarchies, property inheritance and methods and procedures. Data mining techniques can be used to find the characteristics of object evaluation or the trend of changes For objects in the database. Such information can be useful decision making and strategy planning. For example market data can be mined to uncover trends that could help to retail strategies.

Geographic databases have also numerous applications ranging from forestry and ecology planning to provide public service information regarding the location of cables, pipes or sewage system. They are also useful for vehicle navigation. Spatiotemporal database that change with time is also a big data in which information can be mined. Streams of data flow in and out of an observation pattern dynamically. They may be huge infinite volume, dynamically changing in nature. Usually multi level, multidimensional on-line analysis and mining should be performed on stream data. Even if the web pages are fancy and informative to readers, they can be highly unstructured and lack pattern. Data mining can often provide additional help to the web search services which include big data.

Data mining are used to specify the kind of patterns to be found in data mining task. The tasks can be classified as predictive and descriptive.

A. Different Types Of Data Mining System:

There are different types of data mining system which can be used with big data. The main techniques used with data mining are as follows.

1) Classification & Prediction:

Classification is the process of finding a model or technique used to classify unknown values with known values called class labels by constructing a Decision Tree.

Prediction is a Technique used to predict unknown values or missing values with few known values. Classification and Prediction Techniques are same but their models are different.

2) Evolution Analysis:

Evolution analysis is used with time series data of previous years. Regularities in such time series data is used to predict future trends in Retail market prices, contributing to decision making regarding retail market prices.

3) Outlier Analysis:

Outlier analysis may be a random error. Detected using statistical tests that assume a distribution or probability model for the data or using distance measures where objects that are a substantial distance from any other cluster are considered outliers.

4) Cluster Analysis:

Cluster Analysis, is technique used to group similar class of objects and to remove dissimilar class of objects. It follows the principle Maximizing Intra class Similarity and Minimizing Inter Class Similarity There is no class labels in the training data sets. The labels are generating using these techniques. The objects in a cluster are grouped based on their similarity. Then rules are formed from the clusters .The major clustering methods includes portioning methods, hierarchical methods, density based methods, model based methods and constraint based clustering method.

IV. BIG DATA IN DATA MINING TOOLS

A. R And Big Data:

Useful features of R:

- Effective programming language
- Relational database support
- Data analytics
- Data visualization

- Extension through the vast library of R packages

Note: Apache Hadoop is an open source Java framework for processing and querying vast amounts of data i.e Big Data

I would also say that sometimes the data resides on the HDFS (in various formats). Since a lot of data analysts are very productive in R, it is natural to use R to compute with the data stored through Hadoop-related tools. As mentioned earlier, the strengths of R lie in its ability to analyze data using a rich library of packages but fall short when it comes to working on very large datasets. The strength of Hadoop on the other hand is to store and process very large amounts of data in the TB and even PB range. Such vast datasets cannot be processed in memory as the RAM of each machine cannot hold such large datasets. The options would be to run analysis on limited chunks also known as sampling or to correspond the analytical power of R with the storage and processing power of Hadoop

The R language is well established, and typically used for statistics and predictive analytics. Despite this, some organizations have been reluctant to use R in production applications because it is memory-bound. Data sets are now so large -- sometimes exceeding tens of gigabytes and hundreds of millions of rows -- that scalability and performance become issues, particularly for mission-critical applications with precise deadlines. Revolution Analytics has extended R to work with terabyte-class data sets through RevoScaleR(tm), an add-on package specifically designed for use with large data sets. It doesn't require expensive or specialized hardware.

B. Rapidminer And Big Data:

Radoop Is Now Part Of Rapidminer. Big Data Analytics Made Easy By Radoop.

1) Radoop:

Advanced Big Data Analytics. Big Data is a worthless without analyzing; visualizing and making sense of it. RADOOP not just provides ETL, analytics and visualization in a single package but uniquely offers predictive analytics. From now on, big data analytics is just more than just reporting the past, it is predicting the future.

C. Weka And Big Data:

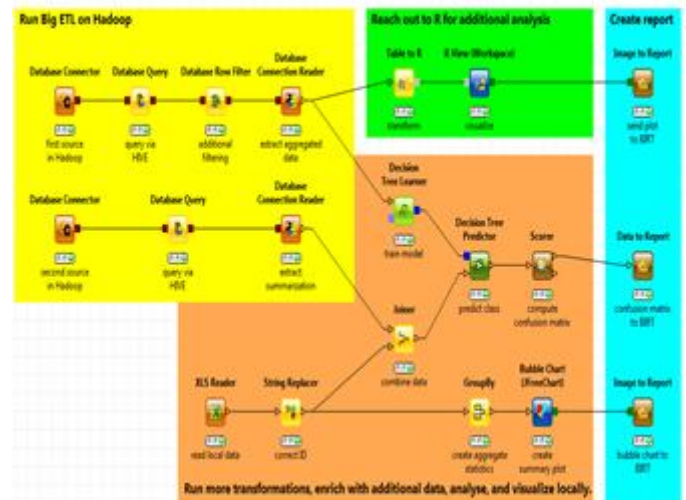
Recent versions of Weka 3.7 also provide access to new packages for distributed data mining. The first new package is called distributedWekaBase. It provides base "map" and "reduce" tasks that are not tied to any specific distributed platform. The second, called distributedWekaHadoop, provides Hadoop-specific wrappers and jobs for these base tasks. In the future, there could be other wrappers.

D. Knime And Big Data:

Big Data can be handled within a normal KNIME workflow. The current set of KNIME database nodes can already be used to perform Big ETL using Hadoop and HIVE and combine that part of the workflow seamlessly with the remaining set of powerful data processing and analysis nodes available in KNIME and through the KNIME community. Downstream, it is even possible to reach out to the R integration and tap into the vast amount of advanced statistical analysis and visualization available there. Upcoming releases will allow to model data processing on

Hadoop even more intuitively and will also allow to run distributed learning algorithms on Hadoop.

Big Data and KNIME - combining the best of many worlds.



E. Mahout And Big Data:

Apache™ Mahout is a library of scalable machine-learning algorithms, implemented on top of Apache Hadoop® and using the Map Reduce paradigm. Machine learning is a discipline of artificial intelligence focused on enabling machines to learn without being explicitly programmed, and it is commonly used to improve future performance based on previous outcomes.

Once big data is stored on the Hadoop Distributed File System (HDFS), Mahout provides the data science tools to automatically find meaningful patterns in those big data sets. The Apache Mahout project aims to make it faster and easier to turn big data into big information.

What Mahout Does

Mahout supports four main data science use cases:

1) Collaborative Filtering:

mines user behavior and makes product recommendations (e.g. Amazon recommendations)

2) Clustering:

takes items in a particular class (such as web pages or newspaper articles) and organizes them into naturally occurring groups, such that items belonging to the same group are similar to each other.

3) Classification:

learns from existing categorizations and then assigns unclassified items to the best category.

4) Frequent Item Set Mining:

analyzes items in a group (e.g. items in a shopping cart or terms in a query session) and then identifies which items typically appear together.

V. RESULTS

Big Data are used to be included for finding the user behavior, for identifying the market and research trends, for increasing the innovations and technology, for retaining the customers, for performing the operations efficiently. Flood of data coming from many sources must be handled effectively using data mining tools with data mining techniques. It provides more market value and methodical for the upcoming generation. Big data has a

wide & variety of application and influence in the field of data mining.

VI. CONCLUSION

To execute Data Mining tools and techniques, we can use big data notion in the real world. Big data creates much interest, presents more opportunities for research and reference in the public sector as well in technical progress. The challenges in data analyzing can be overcome by capturing the techniques in big data along with data mining techniques.

VII. ACKNOWLEDGEMENT

I express my sincere gratitude to God Almighty for all his blessings showered upon me for the completion of this work. I am heartily thankful to my supervisor, R. Madana Mohana, whose encouragement, guidance, supervision, assistance and support from the initial to the final level enabled me to complete the work.

REFERENCES

- [1] Big Data Analytics with R & RAPIDMINER
- [2] <http://www.cs.waikato.ac.nz/ml/weka/bigdata.html>
- [3] <http://www.revolutionanalytics.com/whitepaper/big-data-analysis-revolution-r-enterprise>
- [4] Rapid miner and hadoop." Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011). 2011.
- [5] A. K. Choudhary, J. A. Harding and M. K. Tiwari, "Data Mining in Manufacturing: A Review Based on the Kind of Knowledge", Journal of Intelligent Manufacturing, Volume 20, Number 5, 501-521, 2008.
- [6] survey of Recent Research Progress and Issues in Big Data.
- [7] Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. Wadsworth, Belmont. 1984. Classification and Regression Trees.