Applying NoSQL Database in Data Mining Models for Educational Big Data

Sumon Deb, Md. Ismail Jabiullah, Abdus Sattar

Abstract: Data mining models of education system is limited to relatively small scope of researches and analysis. In the information age, it is the vital issue to retrieve and processing available data in real time. The education system based on relational database has lack capabilities of processing real time data for its large volume and unstructured format. On the other side, NoSQL repositories have strong architecture, it can process big data efficiently in real time by using appropriate model. NoSQL is a non-relational database management system, which stand for "Not Only SQL" or "Not SQL", it can store a wide variety of data, including document, key-value, columnar and graph formats. Relational Database Management System uses SQL syntax to store, retrieve and mining necessary data. Instead, NoSQL database system circumscribe a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data. In this paper, we proposed an improved model for data mining by NoSQL repositories and transformation of data from unstructured form to structured form using NoSQL technologies. Double view of big data collected in the database use data mining technology to retrieve required data. Appropriate process of query a database allows to effective retrieval of imported data for analytical systems. Data can be available for analysis instantly after submitted by users. The proposed model can be suitable for any institutes or organizations, which have a large amount of data.

Keywords: NoSQL, MongoDB, Data Mining, Big Data, Education System.

I. INTRODUCTION

Big Data is now very common word and it has become popular in scientific community for its great impact. There is no universal accepted definition for big data. It can be describe by 5V model, which are Velocity, Variety, Volume, Value and Veracity [1]. Velocity is the speed of shifting digital content from data set to data stream. Variety is different kind of nature of data such as texture data, multimedia data etc. Volume describes the size of data, which can be terabyte to petabyte. Value determines by the cost of data and noise renovation of data stream is called veracity.

Generally, Relational Database Management System is very popular and common in this time, but it has become unfit for work day by day. The main drawback of relational warehouse is limitation of its storage and data mining capability. On the other side, there is no particular and significant structure of big data. For this, it is a big trouble to handle big data using RDBMS. In this case, NoSQL has various databases such as graph, object, multimodal etc. Moreover, the main advantage of NoSQL is various type of data storage capability such structured, semi structured and unstructured data. The repository can store real data in its storage like as file, image, attachment etc.

In educational world, institutions gather a large amount of structured and non-structured data about students, teachers and other employees as well as course materials. Which can be than used to improve educational systems, as related personnel such as teachers can depends not only their own knowledge but also acquire knowledge from others experiences. Nowadays, it is the big problem to perform data mining in such large storage. There are various kinds of data and the major problem is summarization of these data.

In this paper, we proposed a data-mining model using NoSQL, which can perform to generate analytical reports efficiently. In practical world, we need structured data in everywhere. The model has given an accumulated structured data from NoSQL repositories. The model is designed by using MongoDB (NoSQL Database) for its powerful data engine.

Thus, the proposed data-mining model can provide better output in short time. It can be fruitful for educational personnel such students, teachers and so on.

II. LITERATURE REVIEW

In information age, educational data field, this is essential to work on it. Data mining process that produces knowledge from dataset, as system needs and disclose hidden patterns. To improve education field, many researcher tried to establish better system by the virtue of data mining technology. There are huge research work on data mining technique with structured and unstructured data. Many researchers work with data mining for relational database. Till now, it is not popular to practice with NoSQL.

Marcin Mazurek proposed a model titled to combine both RDBMS and NoSQL technologies in an analytical system [2]. A B M Moniruzzaman and Dr. Syed Akhter Hossain has given a discussion to provide classification, evaluation and characteristics of NoSQL databases for Big Data Analysis [3]. Aryan Bansel, Horacio Gonz’alez–Velez, Adriana E. Chis tried to provide data standardization and classification stages with efficient mapping and they discuss different between cloud-based and NoSQL data stores [4].

Benymol Jose and Sajimon Abraham tried to solve the problem of inserting unstructured data.
Applying NoSQL Database in Data Mining Models for Educational Big Data

They discussed flexibility of NoSQL to various sorts of data, structured data, semi structured data or unstructured data [5].

F.N. Chernilin, M.M. Rovnyagin, A.V. Guminskaia, O.V. Myltsyn, V.M. Kinash, A.V. Kuzmin and A.P. Orlov have given discussion to accelerate NoSQL-system and amplify their functionality [6].

The mentioned papers are helpful for my research work. In these papers, they worked with Relational Database Management System (RDBMS) and NoSQL. Moreover, they tried to provide acknowledgement and depth information of structured and unstructured database to analysis Big Data. In this paper, we proposed a model for data mining with NoSQL repositories efficiently. Applying the model, a better output could be achieved for proper analysis of educational big data. It will be more useful for educational data analysis.

III. RESEARCH METHODOLOGY

A. Research Instrumentation

Data mining techniques and concepts are uses in various perspectives. Educational System, Customer Relationship Management, Healthcare, Banking, Fraud Detection, Market Analysis, Real Estate, Criminal Investigation, Manufacturing, Engineering, Education are the most common Data Mining applications. Educational data mining is now one of the most demanding issue. There are so many techniques are used for educational data mining such as Decision Tree, Naïve Baye, Regression, Nural Networks, K-Nearest Neighbors and so on. These techniques provide us various knowledge to discover such as association rule, classification, regression and clustering. These techniques can be useful us to retrieve data using NoSQL.

As research input, system need collected data from various university and some other educational institutions. To develop this system, the experimental model is designed by Python programming language. The collected data is stored in JSON format. Dataset is produced by using MongoDB. Then the model produce expected output by implementing the data mining technique on the generated dataset using “Matplotlib” library of Python. Here are also used “Numpy”, “Seabon” and “Bokeh” library of python in implementing time. The models are assesses by calculating accuracy through sklearn preprocessing.

B. Data Collection Procedure

Firstly, it is mandatory to identify the Big Data and determine the origins from which come. Suppose that we dealing with a varsity, which is very similar to any other universities. After collecting data of one university, it need try to collect data from other educational institutions. Main differences start to emerge at the level of elementary and secondary schools. Then this data will be separated on various groups based on its origins, as shown in Figure 1. The sources of collected data can be different ways, some of these ways have shown in bellow:

- Data from social network system

Figure 1. Big Data Collect from various Sources

C. Usage Scenario and Design Principle

The system targets various groups. This system will be useful to researchers, teachers, students and system of education. We discussed various use-case scenarios for the system:

1. Education Sector
   a. Actors: Teachers, Students, Stuff.
   b. Input: Students Information, Teacher’s Information, Students attendance results, others education information.
   c. Output: Big data analytical report, structure dataset, aggregated statistics for educational system, visualized pattern.
   d. Tools: Desktop application.

2. Data distribution from large amount of both structured data and unstructured data.
   a. Actors: Database administrator.
   b. Input: Stored data from various repositories.
   c. Output: Extracted dataset of structured and unstructured data, algorithms.
   d. Tools: NoSQL repositories such as MongoDB, RDBMS.

3. Data Analysis.
   a. Actors: Data scientist.
   b. Input: Dataset of structured and unstructured data.
   c. Output: Data analytical reports, visualization patterns, algorithms.
   d. Tools: Statistical packages of programmatic language like Python or R.

4. Predictive Modeling.
   b. Input: Relational repositories, Anomaly detected data.
   c. Output: Graphical Report, Prediction.
   d. Tools: Data mining tools, visualization package of Python.
To achieve the goal of our proposed system, the above usage scenario has been determined and it will be useful for any institutions.

**D. System Architecture**

There are two types of dataset in our proposed system, which are structured and unstructured dataset. The reason is large volume and variety of dataset. There are many institutions, which have various kinds of data in their own format. These have become an unstructured data when all data are combined in a single database system.

**Data Sources**

Data stored by different actors such as teachers, students, and stuffs. They collect data from various sources such as universities, schools; educational institutions and so on about student’s information, teacher’s information, student’s attendance, results etc. They are mainly collected as data. Collected data can be structured or unstructured form.

**NoSQL Repositories**

Global educational raw data is loaded in NoSQL repositories like as MongoDB. Which is able to store both kinds of structured data and unstructured data in Database. In NoSQL database, data are stored in JSON format. For this, the data can be document, key-value, columnar or graph. The NoSQL database can also provide user with relational data.

**Relational Data Warehouse**

It is only able to gather structured data from repositories. The model passes unstructured data to Hadoop for transform into structured data by performing some operations.

**Figure 2. Big Data Collect from various Sources**

In this paper, our prime concern is to retrieve those data from large volume of data efficiently. To do this, data are stored in NoSQL repositories, then classify and cluster these dataset using data mining technology, as shown in Figure 2.

The architecture focuses our proposed model, by which, users can retrieve and mining data efficiently and they get all analytical reports in the system. The architecture is performing with proper algorithm, which has given in next section. The working procedure has shown below:
Applying NoSQL Database in Data Mining Models for Educational Big Data

Figure 3. Flow Chart of Proposed System

It is the final output of our system, which provide analytical output to users in various form that depends on the user’s requirements; such as graphical, tabular, SQL and so on. The generated output, which is structured form, can also be store in relational data warehouse.

E. Proposed Model Algorithm

After find and identify a data source we can begin processing data with them. The algorithm of our propose system is given bellow, as shown in Figure 3.

This figure shows that, first collect data from various sources, then load data into NoSQL database. The collected data can be in both form, which are structured or unstructured format. After loading data into NoSQL database, data processed by data mining techniques. There are various processing techniques such as clustering, classification, regression, Summarization etc.

After processing data, processed data can be provide to Data Scientists or Data Researchers for analysis. Users can also receive processed data in graphical form. After analysis data, system provides big data analytical output. The output can be also stored in data warehouse by processing and formatting data.

The Flow Chart can be described as follows,

Step 1: Collect input data in structured or unstructured form and produce the data into JSON format.

Step 2: Load data in NoSQL repositories. NoSQL database such as MongoDB is preferable.

Step 3: If loaded data has already in structured form then data can be stored directly in relational data warehouse; Otherwise data is sent to Hadooop for transforming into structured form.

Step 4: For analysis data of NoSQL repositories, data would be mining from NoSQL database.

Step 5: Retrieval data can be clustering, classification or regression for analysis.

Step 6: Retrieval data can be also summarization for visualized and graphical report.

Step 7: After classification structured data can be also stored in relational data warehouse.

Step 8: Through proper analyzing of data using clustering, classification, regression and summarization, data analyst can get expected outcome from stored data of NoSQL database.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In general, there are different kinds of data in real life and all over data produced in unstructured form. To evaluate performance, need collected data in JSON format. For performing the model, we have collected data from various universities. These data have some similarity. All those data, we have converted in JSON format as shown in Figure 4.

Figure 4. Collection of Data in JSON Format

To test the system, it is most important to check the outcome. The experimental outcome and explanation shown in Table-I.

<table>
<thead>
<tr>
<th>Name</th>
<th>Student_ID</th>
<th>Passing_Year</th>
<th>Varsity</th>
</tr>
</thead>
</table>

Table-I: Collection of Data in JSON Format

Publishes By:
Blue Eyes Intelligence Engineering & Sciences Publication
Table I: Experimental results and discussions

<table>
<thead>
<tr>
<th>Input Data</th>
<th>Expected Output</th>
<th>Experimental Output</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured data in JSON format.</td>
<td>Structured data in SQL format.</td>
<td></td>
<td>Relational data warehouse such as MySQL or SQL server can gather only structured data. So it is essential to transform unstructured data into structured form. In this model, Hadoop technology is used for data transformation process.</td>
</tr>
</tbody>
</table>

- **Student results of different universities.**
  - Input Data: Student results of different universities.
  - Expected Output: Result comparison between different universities.
  - Experimental Output: Box chart of student results.
  - Explanation: Data analyst need to analyze educational data for comparing or rating different educational institutions based on student results. To produce this analytical report, our system at first classify data from NoSQL database and then summarize the classified data. The produced report can be graphical, tabular or textual form.

- **Student information with passing year of different universities.**
  - Input Data: Student information with passing year of different universities.
  - Expected Output: Year based number of passing students for individual university.
  - Experimental Output: Line chart of passing students.
  - Explanation: To generate the analytical report of the total number of passing students based on year, the system at first classify passing year and then accumulate and cluster the data. Educational personnel can be generated these kind of report by using this model.

- **Student information with their residential area/location.**
  - Input Data: Student information with their residential area/location.
  - Expected Output: Area based number of students, globally comparison of these data.
  - Experimental Output: Pie chart of student results.
  - Explanation: For educational survey, it is most important to know student information of various cities. The system can generate report based on area by classifying area and university of all over student information.

- **Global Student information with passing year.**
  - Input Data: Global Student information with passing year.
  - Expected Output: Ratio analysis of yearly grow up of number of passing students.
  - Experimental Output: Area chart of global passing students.
  - Explanation: Sometimes we need to know, how many students are passed from all or specific number of universities in a particular year. To produce the report, system summarize all student information based on passing year.

- **All passing student information of different universities.**
  - Input Data: All passing student information of different universities.
  - Expected Output: Ratio analysis of total number of passing students and comparison between different universities.
  - Experimental Output: Pie chart of passing students.
  - Explanation: To compare different universities, at first data should be regretted for removing redundant data and then classify and summarize data based on particular university. The generated report can be shown in graphical form for ratio analysis of total number of students of different universities.

V. CONCLUSION

The proposed model architecture obtains from evaluation real life implementation of educational database. In this paper, covered implementation of data repositories based on relational database. Theses research is creating a large amount of data in educational repositories that is not being utilize perfectly. It is possible to get advantage from the large educational dataset, by using data mining processes, which serve as stronger tools for characterization, investigation, analysis and prediction. Nowadays, many people are using data mining techniques in various field but it is limited used in educational perspectives. Moreover, it is more limited for NoSQL repositories. In this proposed model, used NoSQL database, which provide an efficient way to retrieve data from database that contains a large volume of data. Who need to analysis data from such database can be able to retrieve data in a short time to use that information for analysis, characterize, investigation and preparing valuable reports. This model can be helpful for any organizations especially for educational institutions, which play with large volume of data. This model will provide advantages to the both students and teachers also and they can support their institutions by quality assurance of education system. It is more effective to improve our education system.
Applying data mining techniques with NoSQL repositories on educational data. Which discloses some significant region on education system, where analysis with data mining has acquired benefits, like identification performance of different student, comparison between different teachers, students and educational institutions, identify students satisfaction for a specific area, student evaluation, analysis students course registration planning, analysis the enrolment head count and so on. The paper primarily investigated, analyzed and compare student performance with NoSQL database for achieving efficient data mining from various data sources. The study has developed a significant system, which can be apply for analysis student activity. The study of this research can be make easy to analysis student performances, global educational environment evaluation and so on. It is more desirable to improve education system.

REFERENCES

AUTHORS PROFILE

Sumon Deb was born in Dhaka, Bangladesh, in 1995. Currently he is studying Master of Science in Computer Science and Engineering, faculty of Science and Information Technology (FSIT) at Daffodil International University. He completed his Bachelor of Science in Computer Science and Engineering, faculty of Science and Information Technology (FSIT) at Daffodil International University. His research interests include Big Data, Data Mining, NoSQL, Machine Learning, Internet of Things (IoT), Network Security, Information Security, Cryptography, Cyber Security, Cryptocurrency, Steganography, Wireless Network, Mobile Network, Artificial Intelligence, Machine Language, Deep Learning, Software Security, Satellite Network, Image Processing, Software Testing and Neural Networks.

Abdus Sattar was born in Comilla, Bangladesh, in 1983. Currently he is working as Assistant Professor at the Department of Computer Science and Engineering at Daffodil International University, Faculty of Science and Information Technology (FSIT). Previously, He was employed as Assistant Professor in the Department of Computer Science and Engineering (CSE) at Britannia University, Comilla. He received Bachelor of Science in Computer Science and Engineering (CSE) from Ahsanullah University of Science and Technology (AUST), Bangladesh and Master’s Program of Interactive Systems Engineering (ISE) from KTH Royal Institute of Technology, Sweden. His Master’s thesis research areas cover user-centered design methodologies and user experiences. During his Master’s thesis, he was employed as a Research Assistant at Karolinska Institute Science Park, Sweden. His research work carried out in Collaboration with Södertörn University, Stockholm University and Karolinska Institutet Innovations, Sweden. His research interests include Human Computer Interaction (HCI), IoT, Machine Learning, and Interaction Design.

Md. Ismail Jabullah received Ph.D. degree in Computer Science and Engineering from the University of Dhaka, Bangladesh. His Ph.D. topic was in Cryptography and Network Security. He is currently a Professor in the Computer Science and Engineering Department at the Daffodil International University, Dhaka, Bangladesh. He has published 39 research articles in reputed journals and more than 78 research papers in the International Conferences. He authored more than 26 books. His research interests are includes: Network Security, Information Security, Cryptography, Cyber Security, Cryptocurrency, Steganography, Wireless Network, Mobile Network, Artificial Intelligence, Machine Language, Deep Learning, Software Security, Satellite Network, Image Processing, Software Testing and Neural Networks. He serves as a reviewer for various reputed journals and conferences annually.