Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.67 – 75

SURVEY ON STUDENT INFORMATION ANALYSIS

K.V.SATHIYAPRIYA

Final Year M.Tech, Department of Information Technology, Hindustan Institute of Technology and Science (HITS), Chennai, India, E-mail: ksathya.3358@gmail.com

B.V.BAIJU

Assistant Professor, Department of Information Technology, Hindustan Institute of Technology and Science (HITS), Chennai, India, E-mail: bvbaiju@hindustanuniv.ac.in

***ABSTRACT:** Educational institutions are important parts of our society and playing a vital role for growth and development of nation. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today, one of the biggest challenges that student's details handling and face explosive growth of personal and educational data and to use this data to improve the quality of tutor's managerial decisions. The design and implementation of a comprehensive student information system and user interface is to replace the current paper records. College Staffs are able to directly access all aspects of a student's academic progress through a secure, online interface embedded in the college's website. The existing technique consider students study and personal details because of library, hostel and behavior related progress results only handle individual database and individual system. In this project propose to deep learning algorithm provides a simple interface for maintenance of student information. It can be used by educational institutes or colleges to maintain the records of students easily. The creation and management of accurate, up-to-date information regarding a students' academic career is critically important in the university as well as colleges. Student Information System deals with all kind of student details, academic related reports, course details, hostel and other resource related details too. After collecting the details using Classifier the dataset clustered. In clustering, based on the students registration categorize the students as a three dataset like library, hostel and admin for education with circular details of each department. It also facilitate us explore all the activities happening in the college, different reports and queries can be generated based on registration options related to students, batch, course, faculty, exams, semesters, certification and even for the entire departments.*

***KEYWORDS:** Student Information System, student dataset, classification and clustering, deep learning algorithm, Data Mining.*

1. INTRODUCTION

This project focuses on analysing the data generated in an educational setup by the various intra-connected or disparate systems to develop model for improving learning experience and institutional effectiveness. The existing system at present huge amounts of data is being accumulated. Traditional way of mining data is manual but in case of large quantities this task becomes tedious. To overcome this condition Data mining tools have been used. In this existing paper we are using pre-processing tool for the analysis of student class data. By using Data mining techniques, knowledge cloud is mined from the data large in size as well as in dimensionality. Understanding and analyzing the factors for different data access is a complex and incessant process hidden in past and present information congregated from academic performance and students' personal information but not enough good result produce. Previously, the college relied heavily on paper records for this initiative. While paper records are a traditional way of managing student data there are several drawbacks to this method. First, to convey information to the students it should be displayed on the notice board and the student has to visit the notice board to check that information. It takes a very long time to convey the information to the student. Paper records are difficult to manage and track. The physical exertion required to retrieve, alter, and re-file the paper records are all non-value added activities. This system provides a simple interface for the maintenance of student information. It can be used by educational institutes or colleges to maintain the records of students easily. Achieving this objective is difficult using a manual system as the information is scattered, can be redundant and collecting relevant information may be very time consuming. All these problems are solved using deep learning algorithm and classification with clustering for student information manage based and intelligible manner which provides facilities like student individual registration id and student profile dataset creation of student's thus reducing paper work and automating the record generation process in an each department. This paper mainly focuses on the managing the information of the students, faculty, library information, study, hostel related information of the college which is maintained by the administration through various levels of registration number controlling.

2. BIG DATA

Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The challenges include capture, storage, search, sharing, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.

2.1 3V'S OF BIG DATA

VOLUME OF DATA

Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to "petabytes".

VARIETY OF DATA

Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi-structured, audio, video, XML etc.

VELOCITY OF DATA

Velocity refers to the speed of data processing. 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.

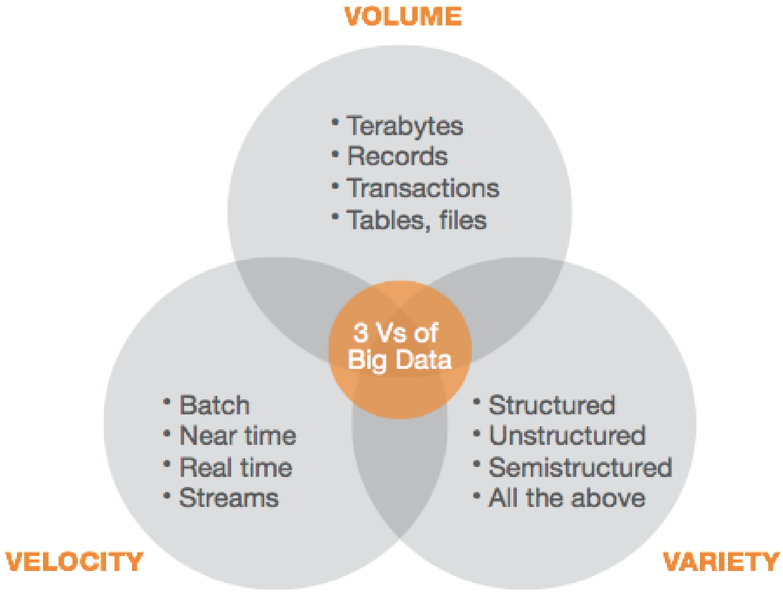

Figure 1 3V's of Big Data

2.2 INFRASTRUCTURE REQUIREMENTS OF BIG DATA

DATA ACQUISITION IN BIG DATA

Even though the data will be in distributed environment, infrastructure must support to carry out very high transaction volumes and also support flexible data structures. To collect and store data, NoSQL are often used in Big Data. NoSQL will not have any fixed schema since it supports high variety of data by capturing all types of data. Keys are used to identify the data point without designing schema with relationship between entities.

DATA ORGANIZATION IN BIG DATA

In the classical term of data warehousing, organizing data is called as data integration. Big Data requires good infrastructure, so that processing and manipulating data in the original storage location can be done easily. It must also supports variety of data formats like structured format, unstructured format etc. Hadoop is a new technology that allows large data volumes to be organized and processed while keeping the data on the original data storage cluster. For example Hadoop Distributed File System (HDFS) in the long term storage system for web logs. These web logs are turned into browsing behavior (sessions) by running MapReduce programs on the cluster and generating aggregated results on the same cluster. These aggregated results are then loaded into a Relational DBMS system.

DATA ANALYSIS IN BIG DATA

Since data is not always moved during organization phase, the analysis may also be done in distributed environment, where some data will stay where it was originally stored and be transparently accessed from a data warehouse. The infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining, on a wider variety of data types stored in diverse systems; sale to extreme data volumes; deliver faster response times driven by changes in behavior; and automate decisions based on analytical models. Most importantly, the infrastructure must be able to integrate analysis on the combination of big data and traditional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to provide new perspectives on old problems. For example, analyzing inventory data from a smart vending machine in combination with the events calendar for the venue in which the vending machine is located, will dictate the optimal product mix and replenishment schedule for the vending machine.

3. DATA MINING

Data mining (the analysis step of the “Knowledge Discovery in Database” process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that results in the discovery of new patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract knowledge from an existing data set and transform it into a human-understandable structure for further use. Besides the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structures, visualization, and online updating.

The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is discovery, commonly defined as “detecting something new”. Even the popular book “Data mining: Practical machine learning tools and techniques with Java” (which covers mostly machine learning material) was originally to named just “Practical machine learning”, and the term “data mining” was only added for marketing reasons. Often the more general terms “(large scale) data analysis”, or “analytics”- or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the user of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

4. MAPREDUCE

Central to the scalability of Apache Hadoop is the distributed processing framework known as MapReduce. MapReduce helps programmers solve data- parallel problems for which the dataset can be sub-divided into small parts and processed independently. MapReduce is an important advance because it allows ordinary developers, not just those skilled in high-performance computing, to use parallel programming constructs without worrying about the complex details of intra-cluster communication, task monitoring, and failure handling. MapReduce simplifies all that.

The system splits the input data-set into multiple chunks, each of which is input as a set of (key, value) pairs and produces a transformed set of (key, value) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to the reduce tasks, which group them into final results. MapReduce uses JobTracker and TaskTracker mechanisms to schedule tasks, monitor them, and restart any that fail.

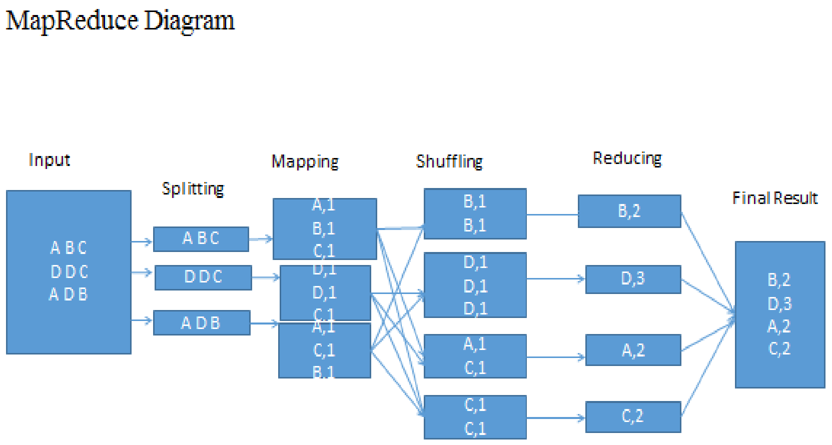

Figure 2 MapReduce Diagram

5. PROPOSED MODULES DESCRIPTION

COLLECT THE STUDENT'S DETAIL DATASET

The student details dataset is being collected, where the detail of the student mentioned, the record have been collected from the library registration, department, hostel register, and etc. sources. Every particular dataset have store their particular information as library has the book transaction details, department has the study performance details, hostel has the room number and other related details. In same way every dataset contain the particular details.

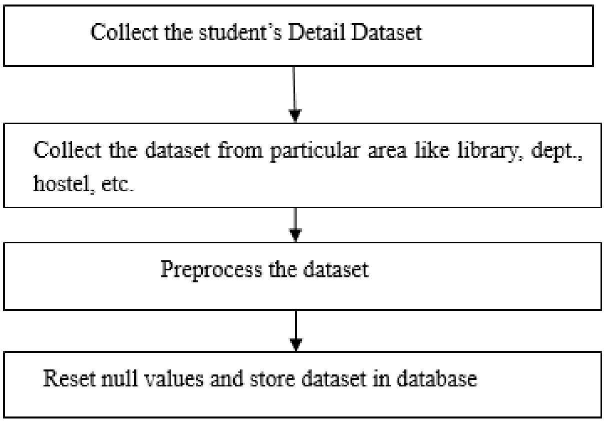
Collect the student's Detail Dataset
Collect the dataset from particular area like library, dept., hostel, etc.
Preprocess the dataset
Reset null values and store dataset in database

Figure 3 Collect the student's Detail Dataset

DATASET TRAINING AND CLASSIFICATION

The dataset training is collecting the information from every particular dataset for every particular registration number, such as Name, hostel allocation information, Library book details, educational performance and etc. the information is being classified based on the allocated information and registration number of the student. Student information has to be mention in the dataset else it will be calculated as a null value.

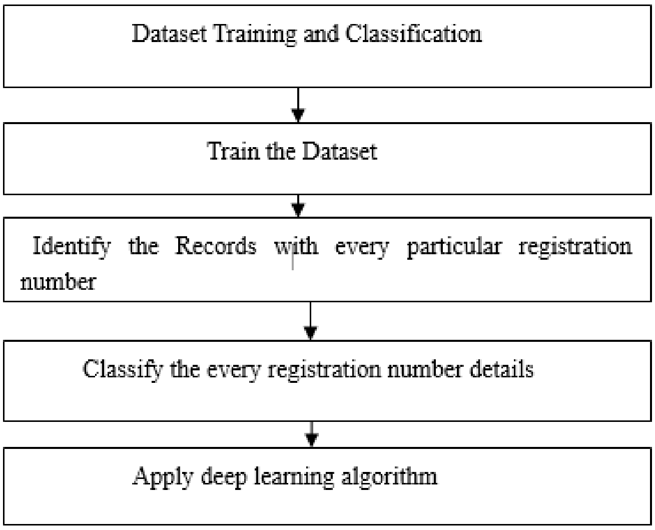
Dataset Training and Classification
Train the Dataset
Identify the Records with every particular registration number
Classify the every registration number details
Apply deep learning algorithm

Figure 4 Dataset Training and Classification

CLUSTERING OF CLASSIFIED DATASET BY DEEP LEARNING

The clustering process is performing the grouping of information for every particular registration number, where it represent the all information being extracted in the training process and finalizing the students details. The grouped dataset information is calculating the overall information within a single record and has to be uploaded in the database.

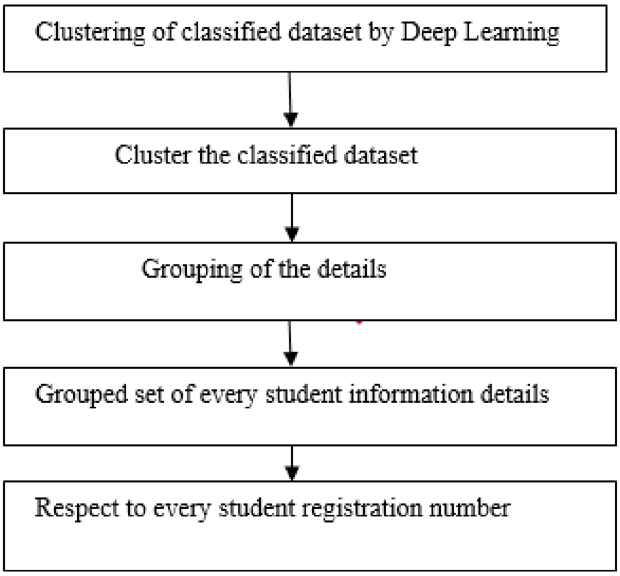
Clustering of classified dataset by Deep Learning
Cluster the classified dataset
Grouping of the details
Grouped set of every student information details
Respect to every student registration number

Figure 5 **Clustering of classified dataset by Deep Learning**

STUDENT'S INFORMATION EXTRACTION WITH REGISTRATION NUMBER

The student's information will be extracted within the particular registration number. The registration number is the input over here and the related information of the particular information will be the output, which represent the complete information of the student.

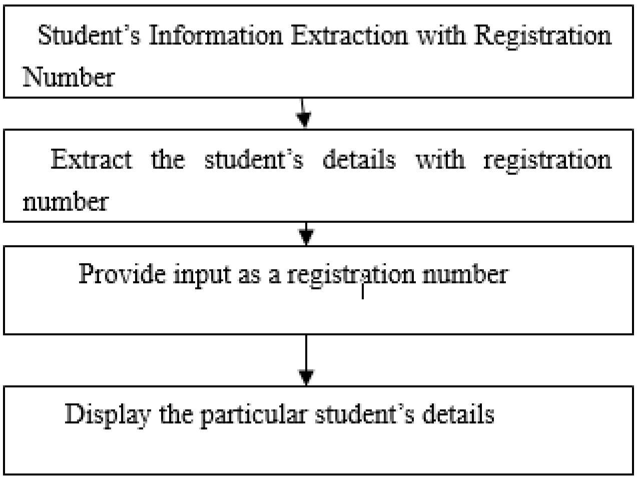
Student's Information Extraction with Registration Number
Extract the student's details with registration number
Provide input as a registration number
Display the particular student's details

Figure 6 **Student's Information Extraction with Registration Number**

6. IMPLEMENTATIONS

APACHE HADOOP

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity of contributors and top-level project being built and used by a global community of contributors and users. It is licensed under the Apache License 2.0.

The Apache Hadoop framework is composed of the following modules:

- Hadoop Common- contains libraries and utilities needed by other Hadoop modules.
- Hadoop Distributed File System (HDFS)- a distributed file-system that stores data on commonly machines, providing very high aggregate bandwidth across the cluster. Hadoop YARN- a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.
- Hadoop MapReduce- a programming model for large scale data processing.

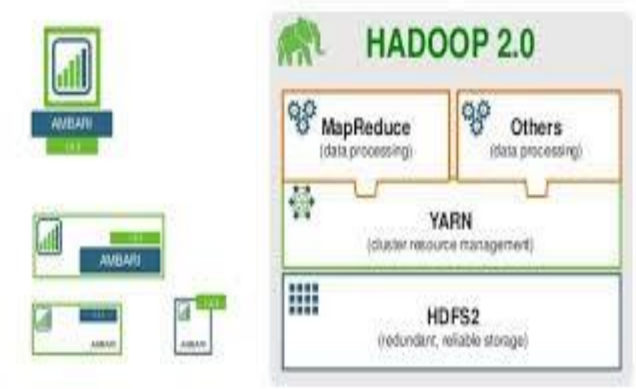

Figure 7 Hadoop Architecture

YARN stands for "Yet another Resource Negotiator" and was added later as part of Hadoop 2.0. YARN takes the resource management capabilities that were in MapReduce to do what it does best, process data. With YARN, we can now run multiple applications in Hadoop, all sharing a common resource management. As of September, 2014, YARN manages only CPU (number of cores) and memory, nut management of other resources such as disk, network and GPU is planned for the future.

7. CONCLUSION

This project is only focused on analyzing the academic information of the students using only influencing factors by deep learning algorithm Classifier and clustering accuracy of the academic information on of the students using a dataset that comprises of all academic, personal information using deep algorithm. This project helps the institution to know the academic status of the students in advance and can concentrate on each student information details. The study can be carried by better prediction or a new algorithm can be developed for better classification and prediction using the high influence attributes would be the future work.

REFERENCES

1. Ahmad Slim, Jarred Kozlick, Gregory L. Heileman, Jeff Wigdahl, Chaouki T. Abdallah. "Network Analysis of University Courses", 2014.
2. Alfred Essa, Hanan Ayad "Improving student success using predictive models and data visualizations", 2012.
3. Baker R.S.J.D., & Yacef K, "The state of educational data mining in 2009: A review and future vision", Journal of Educational Data mining, I,p,g. 3-17, 2009.
4. Bichsel, J., "Learning Analytics in Higher Education: An Annotated Bibliography", 2013.

5. Jai Ruby & K. David, ”A study model on the impact of various indicators in the performance of students in higher education”, IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, pp.750-755, May 2014.
6. Jin Mei-shan1 Qiu Changi-li 2 Li Jing 3. “The Designment of student information management system based on B/S architecture”. 978-1-4577-1415-3/12 2012 IEEE.
7. Michael Fire, Gilad Katz, Yuval Elovici, Bracha Shapira, and Lior Rokach, “Predicting Student Exam’s Scores by Analyzing Social Network Data” 2012, chapter, Active Media Technology, Volume 7669 of the series Lecture Notes in Computer Science pp. 584-595.
8. S.R.Bharamagoudar, Geeta R.B., S.G.Totad “Web Based Student Information Management System”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013m Copyright to IJARCCE www.ijarcce.com 2342.