ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Survey of Data Governance in VESIT Management System

Kunal Dulani[1], Mohit Chanchlani[1], Meeta Chanchlani[1], Manoj Ahuja[1], Richard Joseph[2]

B.E. Students, Dept. of Computer, VESIT, Chembur, Mumbai, India[1]

Assistant Professor, Dept. of Computer., VESIT, Chembur, Mumbai, India[2]

ABSTRACT: Data governance is a control that assures that the data entered by an operations team member or by an automated process meets precise standards, much as a Business norm, definition of data and its integrity constraints in the data model. For this, the document level faculty score is calculated, along with student performance. All score calculations are done in accordance to the criteria specified by National Board of Accreditation (NBA). This project aims at providing a better insight to the result generated by feedback so that institution can strive to improve in those areas.

KEYWORDS: Educational Data Mining, Accreditation.

I. INTRODUCTION

With rise of technological advancement everyday, we move more towards automating the manually managed previous processes. This has led to generation of large volumes of data or knowledge base. Data mining is successfully extracting the useful information from these knowledge bases. Data mining has become critical for any institution to get a better insight on the utility of its resources. The main function of data mining is to apply algorithms and extract patterns in data. There is increasing research interest in data mining and this has led to the emergence of a new field called educational data mining. Using these techniques different types of knowledge can be found such as association rules, classifications and clustering.

In this paper we have proposed a system to replace the current manually managed data of Vivekananda Education Society's Institute of Technology with online system that generates results of student performance in seconds that would have otherwise been required hours of manual computation. Data governance is a control which confirms that the data entered by an operations team member or by an automated process meets precise standards, much as a Business rule, a data definition and data integrity constraints in the data model.

For managing the data of this above mentioned institution we have first modeled the entire data sets to accept only clean data. This is achieved through the medium of google sheets and Google app scripts. Certain rules are defined to accept only data that is consistent with the rules that define the clean data sets. This overcomes the problem of inconsistency in manual data entry wherein each data item entered may be in different format than the previous one and data consistency is subject to circumstantial situations.

Taking the next step on the same guidelines, we have developed rules that check for satisfaction of rules stated by National Board of Accreditation in their Self Assessment Report (SAR). these rules will help develop patterns in the previous datasets and predict the ultimate quality of education imparted by the institute and also the student response and absorption level of the concepts. On having these predictions we will be able to judge the performance of students and the faculty and also predict their approximate performance in the near future. This prediction will help in estimating the college performance outcome after the accreditation review.

The main objective of the paper will be to help the college understand the students strengths and weakness and work accordingly. For this the achievement of Course Outcomes and Program Outcomes will be predicted and students

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

will be helped to work into areas of their strength in situations like choosing their domain of interest. Additionally, we also compare the working of two algorithms ID3 and C4.5 and by the end aim to determine which one would better suit the cause.

II. DATA MINING DEFINITIONS AND TECHNIQUES

A. PREPARATION OF CLEAN DATASETS

Data that is available with the institution is scattered across various locations and uses different file formats. Therefore, firstly, we need to bring all the data to one location and in one format for working with it. Data available with the institution would be in a different order than the expected one. To analyze data we need to sort data and segregate it so that we get data that can be classified and understood better. Google app scripts is used to introduce integrity rules to the entered data.

B. TECHNIQUES USED

I] Model for Data cleaning and storing:

Data will be cleaned as follows:

1. Data scattered across the institute will be collected.
2. Dynamic rule will be formed to determine the authenticity of the data. (For example only phone numbers having ten digits will be accepted.)
3. Another set of rules will be determined dynamically as per the standard used to measure percentage of the satisfaction of rules set by the accreditation committee. (For example students having a pointer above 6 are to be considered as successful this year. Next time if the criteria is raised to 6.2 the rule can be defined dynamically.)
4. Using these dynamic rules the data is segregated into different sheets.
5. This hierarchy is represented as follows:

VESIT

- ELECTRICAL
- COMPUTER
  - ADMIN
    - ACCREDITATION
    - SUPPORT
    - DATABASE ADMIN
    - INFRASTRUCTURE MANAGEMENT
  - STUDENT
    - ATTENDANCE
    - CO CURRICULAR ACTIVITIES
    - COMPETITIVE EXAMS
    - HIGHER STUDIES
    - PLACEMENTS
    - RESULTS
    - SCHOLARSHIPS
  - FACULTY

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

- PUBLISHED BOOKS
- GRANTS
- CERTIFICATIONS
- COURSES CONDUCTED
- COURSES ATTENDED

◆ COURSES

- COURSE OUTCOMES
- PROGRAM OUTCOMES
- MAPPING TABLE
- NO. OF STUDENTS ENROLLED IN COURSE

➔ INFT
➔ INSTRUMENTATION
➔ EXTC

6. Data is stored in these folders now and mined as required using ID3 algorithm.

II] ID3 Algorithm

Iterative Dichotomizer 3 or ID3 algorithm will be used for prediction purposes in this paper. The description of ID3 algorithm can be given as follows:

The ID3 algorithm begins with the original set *S* as the root node. On each iteration of the algorithm, it repeats through each and every attribute of the set *S* that is not used and finds the entropy *H(S),*

(or information gain *IG(S)*) of that attribute. It then selects the attribute which has the lowest entropy (or highest information gain) value. Selected attribute is used to split the set S (e.g. age < 50, age > 50 and age < 100, age < 100) to construct subsets of data. The algorithm continues to recurse on each subset, considering only attributes never selected before.

- Calculate the entropy of every attribute using the data set *S*
- Split the set *S* into subsets with the help of attribute for which entropy is lowest (or, gain is highest)
- Construct the decision tree node having that attribute
- Using remaining attributes continue recursion on the subsets.

Entropy:

Entropy H(S) is the amount of uncertainty the (data) set S has (i.e. entropy specifies the (data) set S).

Entropy = - p(a)*log(p(a)) - p(b)*log(p(b)

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

Where,

S - It is the present data set for which we calculate the entropy (it differs in every iteration of algorithm)

X - Set of the classes included in S

p(x) - It is the proportion in which no. of elements in class X to the amount of elements in the set S

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

When H(S)=0, the set S is said to be perfectly classified (i.e. all the elements in S belong to same class).

In ID3, entropy is calculated for all the attributes that are remaining. The attribute with the smallest entropy is used to split the set S on this iteration. More the entropy, the greater is the potential to improve the classification here.

Information gain:

Information gain IG(A) is the difference in entropy from initial stage till the set S is split using an attribute A. Therefore, it is the amount of uncertainty in S that was lowered after splitting the set S using attribute A.

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where,
H(S) - Entropy of set S
T - The subsets formed by splitting set S using attribute A such that p(t) - The proportion of the no. elements in t to no. of elements in the set S
H(t) - Entropy of subset t.
In ID3, information gain can be calculated (rather than entropy) for the attributes that are remaining. The attribute with the highest information gain is selected to split the set S in this iteration.

III] C4.5 Algorithm

This algorithm works as follows:
Entropy calculated above is used to form decision tree.
Form a few base classes that each case can be fitted to, if no base class is found it creates a new bass class.
Perform pruning to overcome data correction. This is done as C4.5 goes back to the tree once it is formed to remove branches that do not help and replace them with leaf nodes.

IV] Training

Both these algorithms are highly dependent of training sets for their success and hence a diverse training set is applied to generate better results. For this, records of over 1000 students from the previous year are taken whose results are already known. These records are then fed to the algorithms and they form their decision trees using these examples. Each attribute of this training set is is a node of the decision tree. And these decision trees will be used in future to predict the results.

III. METHODOLOGIES USED

1. Google Apps Scripts

Google Apps Script is a scripting language that occupies little memory for application development on the Google Apps platform. Google Apps Scripts is based upon JavaScript 1.6 including some parts of 1.7 and 1.8 and provides subset of ECMAScript 5 API, however, the entire app is executed on Google which would rather have used client machine and memory in client site. According to Google, Apps Script facilitates simple methods to automate most of the tasks across Google products lineup and third party apps and services. Apps Script is also the tool that powers the add-ons for Google Docs, Sheets & Forms.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Benefits of Google App Scripts are as follows:

- As it is based on JavaScript it is simple to learn.
- Debugging the App Scripts are done by debugger that's entirely on cloud and can be accessed in web browser.
- It can be used to make easy tools for an organization for itself.

2. Google Sheets

2.1. About Google Sheets

Google Sheets is a spreadsheet that is included in the web-based software office suite that is offered by Google within the Google Drive services. The suite allows users to create and edit documents online on the go and also collaborating with other users in real-time.

The app is available as web applications, as Google Chrome apps that work offline, and also as mobile apps for Android devices as well as iOS machines. The app offers compatibility for Microsoft Office file formats and other documents. The suite also includes Google Forms, Google Drawings (diagramming software.)

IV. IMPLEMENTATION EXAMPLE

A example set of 30 students is taken and ID3 algorithm is implemented to convert this raw data to a binary tree. This tree then can be used to predict all future instances of performance related results.

The implemented example can be explained as follows:

No.	ITM	ASM	ATT	TW	FG	No.	ITM	ASM	ATT	TW	FG
1	13	B	50	17	P	8	17	A	75	20	P
2	14	B	75	20	P	9	20	A	75	18	P
3	16	A	75	22	P	10	14	B	75	16	P
4	16	A	75	23	P	11	15	A	75	17	P
5	14	B	80	20	P	12	16	A	75	20	P
6	15	B	90	25	P	13	14	A	80	21	P
7	16	A	75	19	P	14	12	B	50	24	F

Table No. 01

Example Scores of Students

The above example set comprises of marks of 14 random students in a particular subject. ID3 Algorithm was run on the example set using Weka tool and a decision was generated after calculating the entropy and information gain.

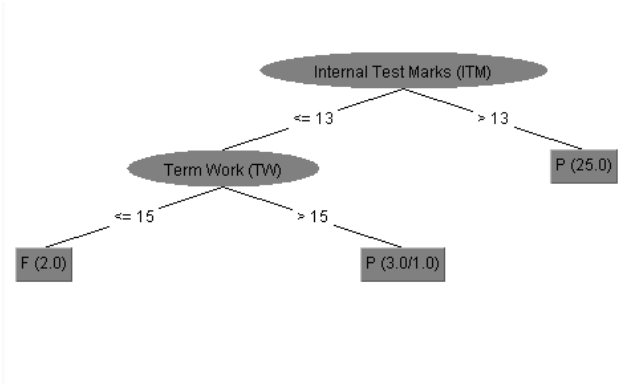
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017



This above example is a small training set and the decision tree generated will be used for predictions in future.

This tree will now be referred to predict the results of current batch of students.

And in the above shown way all aspects related to college grading by the accreditation committee from N.B.A.(National Board of Accreditation) will be converted into rules in real time and trees will be formulated to predict all future similar instances.

V. CONCLUSION

This paper will help the college in supervising predicting all factors that are responsible for grading the college during accreditation based on the previous performance results and rules specified by National Board of Accreditation (NBA). For this, google sheets and google app scripts is utilized and Iterative Dichotomizer 3 (ID3) prediction algorithm is applied. After, the implementation of of this project the college is able to control the grade that will be obtained from the accreditation committee post the college analysis.

VI. DEFINITIONS

NBA - National Board of Accreditation.

SAAR - Self Assessment Report.

ITM - Internal Test Marks.

ASM - Assignment Grade.

ATT - Percentage of Lectures Attended.

TW - Termwork Granted.

FG - Final Result.

REFERENCES

1. Brjjesh Kumar Baradwaj and Saurabh Pal, “Mining Educational Data to Analyze Students Performance”
2. Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer and William F. Punch “Predicting Student Performance: An Application of Data Mining Methods with an Educational Web Based System”
3. Suhem Parack, Zain Zahid and Fatima Merchant “Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns” .
4. Yahya Eru Cakra and Bayu Distiawan Trisedya “Stock price prediction using linear regression based on sentiment analysis”.
5. N. D. Valakunde and M. S. Patwardhan “Multi-aspect and Multi-class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process”.
6. Kranti Ghag and Ketan Shah “Comparative analysis of the techniques for Sentiment Analysis”.
7. Cheng-Fa Tsai and Chun-Yi Sung “DBSCALE: An efficient density based clustering algorithm for Data Mining in large Databases.”

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

8. Majid Al-Ruithe and Elhadj Benkhelifa and Khawar Hameed, "Key Dimensions for Cloud Data Governance."
9. Richard J. Self, "Governance Strategies for the Cloud, Big Data, and Other Technologies in Education."
10. Jannieca Camba, Roanna Ellise David, Ariel Betan, Annette Lagman and Jaime D. L. Caro "Student analytics using support vector machines"
11. Young-Nam Kim, Hye-Yeon Yu and Moon-Hyun Kim, "ID3 algorithm based object discrimination for multi object tracking."
12. Jiabin Deng, JuanLi Hu, Hehua Chi and Juebo Wu "A Study of Teaching Evaluation in Adult Higher Education Based on Decision Tree."

BIOGRAPHY

Kunal Dulani, Mohit Chanchlani, Meeta Chanchlani, Manoj Ahuja are all Final year students of Vivekanad Education Society's Institute of Technology and are currently pursuing their Bachelor of Engineering degree with a specialisation in Computers.

Richard Joseph is an assistant professor at Vivekanad Education Society's Institute of Technology in the department of computer.