A LITERATURE SURVEY: EDUCATIONAL DATA MINING AND WEB DATA MINING

Sarvesh Tendulkar, Swarada Adavadkar
Department of Computer Engineering
MIT College of Engineering Pune, India

ABSTRACT:

Data Mining (DM) is extracting information from large set of data and obtaining meaningful results from it. Data Mining has evolved into diverse fields which requires development of data collection, creation, data management (including data storage retrieval and database transaction processing) and advanced data analysis (data warehousing and mining). This paper explains about two of Data Mining fields : Educational Data Mining (EDM) and Web Data Mining (WDM). Educational Data Mining (EDM) is an emerging field of data mining which applies various data mining algorithms for betterment of students and improving quality of student courseware. Web Data Mining (WDM) is a technique used to crawl through various web resources to collect information which an individual or a company use to promote business. In this survey, different published research papers on methodologies used in EDM and WDM are compared. Also, it gives an insight into techniques used in EDM and WDM.

Keywords: Keyword Indexing, Search engines, document frequency, term weights, data mining, neural network, predicting school failure, job trends analysis

PART I: EDM

[1] INTRODUCTION

Educational Data Mining (EDM) is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses in mining useful patterns and discovering useful knowledge from the educational information systems, such as, admissions systems, registration systems, course management systems (moodle, blackboard, etc…), and any other systems dealing with students at different levels of education, from schools, to colleges and universities.

As interest in EDM continued to increase, EDM researchers established an academic journal in 2009, the Journal of Educational Data Mining, for sharing and disseminating research results. In 2011, EDM researchers established the International Educational Data Mining Society to connect EDM researchers and continue to grow the field.

Goals for EDM in educational field:
Baker and Yacef [1] describes the following four goals of EDM:
1) Predicting student’s future learning behaviour
2) Discovering or improving domain models
3) Studying the effects of educational support
4) Advancing scientific knowledge about learning and learners

[2] EDM METHODS

Classification
It is a two-way technique (training and testing) which maps data into a predefined class. This technique is useful for success analysis with low, medium, high risk students used in [2], student monitoring systems [3], predicting student performance, misuse detection used in [4] etc.

Statistics
It is a technique to identify outlier fields, record using mean, mode etc. and hypothetical testing. This technique is useful to improve the course management system & student response system [5].

Clustering
It is a technique to group similar data into clusters in a way that groups are not predefined. This technique is useful to distinguish learner with their preference in using interactive multimedia system used in [6], Students comprehensive character analysis used in [7] and suitable for collaborative learning.

Prediction
It is a technique which predicts a future state rather than a current state. This technique is useful to predict success rate, drop out used in Dekker et al. [7], and retention management used in [8] of students.

Neural Network
It is a technique to improve the interpretability of the learned network by using extracted rules for learning networks. This technique is useful to determine residency, ethnicity used in [9], to predict academic performance used in [10], accuracy prediction in the branch selection used in and explores learning performance in a TESL based e-learning system.

Association Rule Mining
It is a technique to identify specific relationships among data. This technique is useful to identify students’ failure patterns [11], parameters related to the admission process, migration, contribution of alumni, student assessment, co-relation between different group of students, to guide a search for a better fitting transfer model of student learning etc. Apart from the above methods, two new methods i.e. distillation of data for human judgment and discovery with models to analyze the behavioral impact of students in learning environments.

[3] APPLICATIONS

1. Analysis and Visualization of Data
2. Predicting Student Performance
3. Enrolment Management
4. Grouping Students
5. Predicting Students Profiling
6. Planning and scheduling
7. Organization of Syllabus
8. Detecting Cheating in Online Examination
[5] LITERATURE SURVEY OF EDM TRENDS

<table>
<thead>
<tr>
<th>Author(s) and Year</th>
<th>Data Mining Software(s)</th>
<th>Data Mining Algorithm(s)</th>
<th>Input Dataset</th>
<th>Educational Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutta Borah et al., 2011 [13]</td>
<td>SPSS Clementine 11.1</td>
<td>Classification-Decision Tree</td>
<td>AIEEE 2007</td>
<td>Predict student’s enrolment decision</td>
</tr>
<tr>
<td>Lin S.H., 2012 [8]</td>
<td>WEKA</td>
<td>Decision Tree Algorithm</td>
<td>5943 records of 1st year students of Biola University</td>
<td>Student Retention Management</td>
</tr>
<tr>
<td>Merceron and Yacef. 2005 [14]</td>
<td>Excel, Clementine, Tada-Ed, SODAS</td>
<td>Classification-DT, Clustering, Association Mining</td>
<td>Logic-ITA Student Data</td>
<td>Student/teachers’ performance monitoring system</td>
</tr>
<tr>
<td>Amjad Abu Saa 2016 [15]</td>
<td>RapidMiner, WEKA</td>
<td>Decision Tree - C4.5, CART, CHAID , ID3</td>
<td>survey distributed to students within their daily classes and as an online survey using Google Forms</td>
<td>Students’ Performance Prediction</td>
</tr>
<tr>
<td>Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto, 2013 [12]</td>
<td>WEKA</td>
<td>Decision Tree, Rule Induction, Synthetic Minority Oversampling Technique(SMOTE)</td>
<td>A specific survey, a general survey, The Department of School Services</td>
<td>Predicting the academic status of students</td>
</tr>
</tbody>
</table>

PART II: WDM

[6] INTRODUCTION

The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc.

Web mining refers to the automatic discovery of interesting and useful patterns from the
A LITERATURE SURVEY: EDUCATIONAL DATA MINING AND WEB DATA MINING

data associated with the usage, content, and the linkage structure of Web resources. Web mining lies in between and copes with semi structured data and/or unstructured data. Web mining calls for creative use of data mining and/or text mining techniques and its distinctive approaches. Mining the web data is one of the most challenging tasks for the data mining and data management scholars because there are huge heterogeneous, less structured data available on the web and we can easily get overwhelmed with data.[18] Web data mining can be broadly classified into 3 types:

- Web usage mining
- Web content mining
- Web structure mining

[7] WDM APPLICATIONS

Applications:
1. E-commerce to do personalized marketing
2. Analyze feedback for particular product
3. Government agencies to identify threat and fight terrorism
4. Companies to keep better relation with customers
5. Analyze current market trends for better performance[20]

[9] LITERATURE SURVEY: INDEXING TECHNIQUES

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Author &amp; Year</th>
<th>Indexing Techniques</th>
<th>Methodology</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Changsha ng Zhou, Wei Ding, Na Yang 2006[24]</td>
<td>Double Indexing Mechanism of Search Engine</td>
<td>It has document index as well as word index. The document index is based on the clustering, and ordered by the position in each document. In the retrieval, the search engine first gets the document id of the word in the word index, and then goes to the position of corresponding word in the document index.</td>
<td>Time consuming as the index exists at two levels.</td>
</tr>
<tr>
<td>2</td>
<td>Fabrizio Silvestri, Raffaele Perego and Salvatore Orlando 2004[23]</td>
<td>Assigning Document Identifiers to Enhance Compressibility of Web Search EnginesIndexes</td>
<td>It was the reordering algorithm which partitions the set of documents into k ordered clusters on the basis of similarity measure. According to this algorithm, the biggest document is selected as centroid of the first cluster and n/k1 most similar documents are assigned to this cluster. Then the biggest document is selected and the same process repeats</td>
<td>It is not effective in clustering the most similar documents. The biggest document may not have similarity with any of the documents but still it is taken as the representative of the cluster.</td>
</tr>
<tr>
<td>3</td>
<td>L. Huilin, K.</td>
<td>Efficiently Crawling</td>
<td>Focused search engine comprising of two components; filtering and link forecasting.</td>
<td>Focused search engine does not</td>
</tr>
</tbody>
</table>
Strategy for Focused Searching Engine

The first component filters out the irrelevant documents from previously fetched list of documents and the second component envisage the best matched links form related documents to assign them to crawler for further downloading.

It shows improved effectiveness over the classic word based indexing techniques. Herein a Boolean Information retrieval system adds word semantics to the classic word based indexing. Two of the core tasks of the system, namely the indexing and retrieval components, use a combined word-based and sense-based approach

It is based on the content of the document and on link analysis. It is implemented using user-defined relevant formula, shark-search and Page-Rank.

An index is built on the basis of context of the document rather than on the terms basis using ontology. The context of the documents being collected by the crawler in the repository is being extracted by the indexer using the context repository, thesaurus and ontology repository and then documents are indexed according to their respective context.

This indexing considers the senses of the keyword. Different users type the same query to get the different results according to their interest. It will provide the different senses to the user which will help in generating the more relevant results to the user. The processing of file includes: removal of stop words, stemming, adding of weights and indexing of file.

Representation is dense, so hard to index based on individual dimensions.

It is does not make use of domain ontologies to represent topical maps and link Web pages.

It considered terms only in title of the document with maximum frequency. User has to pass the context along with query keyword which creates problem with naive users.

consider the context of the keywords of the user query.

[10] CONCLUSION

Data Mining technology has emerged to remedy the issues of management and analysis of massive volumes of data. In this paper, a comparison of trends and techniques of EDM and WDM is presented. This work focuses on various trends used in educational data mining and
different technologies used for web data mining. It is observed that the educational outcome varies according to input dataset provided for the given experiment. From above survey it can also be inferred that context based indexing has proved to give more relevant results efficiently.

REFERENCES

[17] P. Mudgil, A.K.Sharma and P.Gupta,” An Improved Indexing Mechanism to Index Web Documents”, 5th International Conference on Computational Intelligence and Communication Networks