WCIT 2010

“Hello World”, Web Mining for E-Learning

Oğuz Mustapaşa\textsuperscript{a}, Adem Karahoca\textsuperscript{a}, Dilek Karahoca\textsuperscript{a,b}, Hüseyin Uzunboylu\textsuperscript{b}

\textsuperscript{a} Department of Software Engineering, University of Bahçeşehir, İstanbul-Turkey
\textsuperscript{b} Department of Computer Education and Instructional Technologies, Near East University, Lefkosa 98010, Northern Cyprus

Abstract

As the internet and mobile applications are getting an important role in our lives, usage of mobile services also took place in educational field since the internet is widespread, which is usually called by the terms “e-learning” or “distance learning”. A known issue on e-learning is all the content’s being online and less face-to-face communication than traditional learning; this brings the problem of chasing student’s success, and advising and managing student’s way of studying. Hence, a recent hot topic, data mining, can be applied on student’s data left on e-learning portals to guide the instructor and advisors to help students’ being more successful. Recent researches done on this topic showed that e-learning combined with data mining can decrease the gap between itself and traditional learning – referred as semantic web mining in general.

Keywords : Semantic Web; Web Mining; E-Learning; Distance Learning; Personalization

1. Introduction

Recent changes on technology and internet changed our life styles and became the center of it. With the new generation telecommunication standard 3G (3\textsuperscript{rd} Generation) and its being widespread, increased the number of mobile applications and thing we get used to do in the most traditional way - such as paper and pencil - started getting mobile. It’s known that with the World Wide Web’s huge movement and improvement in recent years triggered many new trends and one of them is e-learning or distance learning in other words; took a very important and popular place in the field. This new education trend and its effects have been in discussion, pros and cons been defined. These discussions point at a fact that mobile learning; e-learning is facing some problems like lack of face-to-face communication which has an important role in traditional learning. To be able to minimize this gap between e-learning and traditional learning, all advantages of mobile services are in use but there comes a point that instructor needs to follow each student’s way of studying and their success on lectures. As the number of students increase, this gets harder and harder since no face-to-face communication takes place via Course Management Systems generally. We examined two known improvement approaches on students success – index of learning styles and learning and study strategies inventory – and combined them with yet another hot topic data mining; where these two are called as semantic web mining or semantic web when it comes to personalizing the web content in any purpose by mining user’s data left on servers [15].

1.1 Background

In the time, personalization and visualization of data depending on user interest became very popular due to fashion of web technology. This popularity also brought functionality and usability with itself so that W3 became widespread on e-commerce, e-government, and e-learning areas. Considering e-commerce and e-government and
such applications are not hard to implement, application on e-learning was very interesting since that would be something new for education field to teach without any face-to-face interaction in real life [13].

If we think of a traditional education scenario, an educator should track each student’s situation for any lecture and also should try to improve the learning method for the student just to let him/her more successful. When we take this into account for the need of personalization, there had to be some technology which would allow e-learning users to have personalized material in their interest or profile. In the time being, many works and implementations took place for a semantic, intelligent web agent or service for e-learning. As the research topic continued to get into area of interest for many researchers, worldwide organizations standardized the techniques that can be used to build a semantic learning mechanism for educational fields. This goal is achieved by Web Mining (WM), Semantic Web (SW) and a mixture of both as Semantic Web Mining (SWM) so that we can lead the way through personalization of content and personalization of service. The following sections of this paper briefly summarize concepts of Web Mining (WM), Semantic Web (SW), Semantic Web Mining (SWM) and E-Learning [14].

Speaking of a semantic logic on e-learning portal, there must be way to classify or cluster users for providing personalized content depending on valuable information. Considering we are recording each student’s navigation through the system, recent works show that student’s learning style as dependency information for classifying students are widely in use. It is also mentioned that in some cases there had been some practical approaches pointed the weakness of the link between student’s performance and mismatching results to cognitive style instructional methods. However these approaches points these weaknesses, many approaches proved the benefits of using cognitive style instructional methods on developing an adaptive learning environments, say lecture management systems or course management systems supported with a web agent.

In this work we will develop a semantic web agent for students at our university and examine the results on GEP1006 – History of Civilization course which is an online course that students reach materials over an open source e-learning, course management system, named Moodle. While building the semantic structure, we will need to classify the user types in our database so that we will be using cognitive learning styles as a dependency factor for this problem. Also, we will examine student profiles, their navigation through the system and relations with learning materials and search for new relations via data mining techniques.

1.1.1 Web Mining

Web Mining (WM) is the application of data mining on web logs, web contents and web structures. Thus it is the nontrivial process of identifying valid, previously unknown, and potentially useful patterns. As given in the definition, WM has three different types of analysis specs; Web Usage Mining (WUM), Web Content Mining (WCM) and Web Structure Mining (WSM). The specific analysis types of WM for e-learning are WUM and WCM.

Web Usage Mining tries to find out what users are looking for while they are using Web, and WUM also helps to find the patterns for a particular group of people belonging to a region or depending on their interest. Web Content Mining is a kind of text mining application on Web content.

1.1.2 Semantic Web

Semantic Web (SW) derives from W3C director Tim Berners-Lee’s vision of Web as a universal medium for data, information and knowledge exchange. The word semantic web is a product of Web2.0 (second generation web) which makes the web itself to understand and satisfy the user requests and web agents or machines to use the content of web.

1.1.3 Semantic Web Mining and E-Learning

The term Semantic Web Mining is described well by Stumme and et. al. as “Semantic Web Mining aims at combining the two areas Semantic Web and Web Mining. This vision follows our observation that trends converge in both areas: Increasing number of researchers work on improving the results of Web Mining by exploiting
semantic structures in the Web, and make use of Web Mining techniques can be used for mining Semantic Web itself. The wording Semantic Web Mining emphasizes this spectrum of possible interaction between both research areas: It can be read both as Semantic (Web Mining) and as (Semantic Web) Mining.”[8]. Pointing at the definitions given, it’s possible to use web logs for any course available on any course management system or e-learning portal for investigation of semantic information. In a case study on Moodle, case studies for applications of data mining techniques are given.[7, 8] In these studies, the possible techniques for data retrieval and management, educator has to run third party programs manually for information retrieval, for educator are explained briefly. For a semantic and real time system, web services and web agents were announced to be useful. Also, trustworthiness of the data is very important since it can lead the algorithms or mining techniques in wrong or inadequate results. At this point, we can assume that the data we get from student’s answers or information left on CMS is reliable or we can also run data mining algorithms to fetch conflicts on answers to filter them somehow.

2. Materials and Methodology

2.1. Defining Students’ Learning Style

To be able to define students’ learning style we applied Index of Learning Styles Questionnaire on students which is used as an instrument to assess preferences on four defined dimensions; active/reflective, visual/verbal, sequential/global and sensing/intuitive. These are formulated by Richard M. Felder and Linka K. Silverman[9] and the questionnaire is developed by Barbara A. Soloman and Richard M.Felder’s with the name of “Index of Learning Styles Questionnaire”. The aim of this test to get to know each student’s preferences on education and learning. At the end of the test results are written into our database and each learning style classification has a range, (-12, 12); the values that fall in (-12, 0) interval gets a tag “b”; the values that fall in (0, 12) interval gets a tag “a” e.g. -7b or 7a.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Definition</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>act_ref</td>
<td>Active – Reflective</td>
<td>(-12,+12)</td>
</tr>
<tr>
<td>vis_verb</td>
<td>Visual – Verbal</td>
<td>(-12,+12)</td>
</tr>
<tr>
<td>sen_int</td>
<td>Sensing – Intuitive</td>
<td>(-12,+12)</td>
</tr>
<tr>
<td>seq_glo</td>
<td>Sequential - Global</td>
<td>(-12,+12)</td>
</tr>
</tbody>
</table>

As described in the table, cognitive learning styles are divided into four classifiers. The aim of the index of learning styles questionnaire was to discover these preferences of students. In the development period there had to be a way of storing those data into our database and I ended up with a Php page which does the needed calculation. In our application we classified a student on a learning style such as:

\[
x \in \{ \text{act_ref, vis_verb, sen_int, seq_glo } \} \quad \text{and} \quad y \in \{ \text{act, vis, sen, seq} \} \quad \text{and} \quad z \in \{ \text{ref, verb, int, glo} \} \quad | \text{A student “S” is most likely to be } X(y) = 1; \text{ S is most likely to be } X(z) = 3; \text{ and “S” is likely to be between } X(y) \text{ and } X(z) : X(y/z) = 2. \text{ Since we have 3 different classification of learning styles, in the way we cluster our students, we can have a maximum number of 27 possible different ls_code that can describe a student.}
\]

2.2. Preparing the Data Set

Data Set contains information about 404 students who are getting their GEP1006 – History of Civilization lecture over an e-learning portal which is in use at engineering faculty in software engineering department.
Table 2.1.4: ls_content table

<table>
<thead>
<tr>
<th>Classification</th>
<th>Definition</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ls_code</td>
<td>Generated Learning Styles Code depending on Learning Style scale values for the user, changes when a change on user preference is found.</td>
<td>e.g.: 1, 2...n</td>
</tr>
<tr>
<td>view_order</td>
<td>Suggested view order for the student depending on the ls_code field.</td>
<td></td>
</tr>
<tr>
<td>content_type</td>
<td>File extension/Content type of the suggested content. e.g.: html, swf, ppt etc.</td>
<td></td>
</tr>
</tbody>
</table>

Related information about those students is kept in a database and each database table is related to another table for being able to build a relational database where possible.

Table 2.1.5: tbl_learning_styles table

<table>
<thead>
<tr>
<th>Classification</th>
<th>Definition</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>user_Id</td>
<td>Student ID</td>
<td></td>
</tr>
<tr>
<td>tck</td>
<td>tck which is the Turkish application of Social Security Number</td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>Student’s name</td>
<td></td>
</tr>
<tr>
<td>surname</td>
<td>Student’s surname</td>
<td></td>
</tr>
<tr>
<td>act_ref</td>
<td>Index of Learning Styles scale representative of Active – Reflective learning</td>
<td>In a range between -12 and +12. If an attribute is close to be negative, such as: -7, it’s recorded as 7a; if the value is close to be positive than the value is 7a.</td>
</tr>
<tr>
<td>sen_int</td>
<td>Index of Learning Styles scale representative of Sensing – Intuitive learning</td>
<td></td>
</tr>
<tr>
<td>vis_verb</td>
<td>Index of Learning Styles scale representative of Visual – Verbal learning</td>
<td></td>
</tr>
<tr>
<td>seq_glo</td>
<td>Index of Learning Styles scale representative of Sequential – Global learning</td>
<td>In the next log in session the “ls_code” field is filled and associated with “ls_content” table.</td>
</tr>
<tr>
<td>ls_code</td>
<td>Generated Learning Styles Code depending on Learning Style scale values for the user, changes when a change on user preference is found.</td>
<td></td>
</tr>
</tbody>
</table>

“ls_content” table holds the suggested view order for students, depending on their learning style that we get to know after letting them solve a special test. The table includes an ls_code, view order that we suggest, e.g.: 1, 2...n and also a type of the content as a parameter.

“tbl_learning_styles” keeps the results of “learning style” test we applied on students. After each student is done with answering these questions, results are recorded to database with the following attribute values: userId (FK), tck (PK) which is the Turkish application of Social Security Number, name, surname as basics. act_ref, sen_int, vis_verb and seq_glo attributes has a range between -12 and +12. If an attribute is close to be negative, such as: -7, it’s recorded as 7a; if the value is close to be positive than the value is 7a. After those fields are successfully filled, in the next log in session the “ls_code” field is filled and associated with “ls_content” table.

2.3 Defining Students’ Learning Strategy – LASSI

“The Learning and Study Strategies Inventory (LASSI) [12], which is proposed by Weinstein in 1998-2002 and
still, is a 10-scale, 80-item assessment of students’ awareness about and use of learning and study strategies related to skill, will and self-regulation components of strategic learning. The focus is on both covert and overt thoughts, behaviors, attitudes, motivations and beliefs that relate to successful learning in post-secondary educational and training settings and that can be altered through educational interventions. Research has repeatedly demonstrated that these factors contribute significantly to success in college and that they can be learned or enhanced through educational interventions such as learning strategies and study skills courses.”

In this paper, I used a modified version of the LASSI with 100 questions and 15 scales. But since I was using a modified version of LASSI, I also ran SPSS and applied reliability analysis on my dataset in alpha model. The LASSI reliability results showed that 11 of 15 scales are available and reliable in 87 of 100 questions at total. Some questions were discarded as they had no effect on the scale averages and some question results did not differ much depending on users’ preferences on answering these questions. The LASSI questionnaire is applied on “History of Civilization” course students via Moodle Course Management System. History of Civilization course is a project dependent course where students are graded by their success on their projects as a group so the course is based of learning/teaching style. Also I figured out that successful students at this course paid more attention to the questionnaire.

3. Findings

The LASSI scale attributes I used can be seen in the table below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Range</th>
<th>Number of Questions</th>
<th>Related Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitude</td>
<td>DVERME</td>
<td>[1, 8)</td>
<td>6</td>
<td>47, 76, 81, 90, 94, 96</td>
</tr>
<tr>
<td>Concentration</td>
<td>CGÖST</td>
<td>[1, 8)</td>
<td>6</td>
<td>13, 18, 29, 54, 56, 68</td>
</tr>
<tr>
<td>Information Processing</td>
<td>YINELE</td>
<td>[1, 8)</td>
<td>7</td>
<td>15, 20, 27, 58, 59, 62, 83</td>
</tr>
<tr>
<td>Meta</td>
<td>USTBIL</td>
<td>[1, 8)</td>
<td>9</td>
<td>12, 31, 38, 46, 55, 73, 77, 82, 87</td>
</tr>
<tr>
<td>Motivation</td>
<td>DHEDEF</td>
<td>[1, 8)</td>
<td>7</td>
<td>57, 65, 66, 84, 86, 92, 97</td>
</tr>
<tr>
<td>Selecting Main Topics</td>
<td>ACIMLA</td>
<td>[1, 8)</td>
<td>6</td>
<td>39, 40, 41, 42, 48, 71</td>
</tr>
<tr>
<td>Self Study</td>
<td>YARDIM</td>
<td>[1, 8)</td>
<td>9</td>
<td>23, 26, 30, 37, 43, 45, 52, 61, 67</td>
</tr>
<tr>
<td>Self-testing</td>
<td>OZYETE</td>
<td>[1, 8)</td>
<td>9</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
</tr>
<tr>
<td>Study-Aids</td>
<td>BCALIS</td>
<td>[1, 8)</td>
<td>10</td>
<td>36, 63, 70, 72, 75, 79, 88, 89, 95, 99</td>
</tr>
<tr>
<td>Testing Strategies</td>
<td>STRATE</td>
<td>[1, 8)</td>
<td>12</td>
<td>21, 22, 24, 44, 49, 50, 51, 78, 80, 85, 93, 100</td>
</tr>
<tr>
<td>Time Management</td>
<td>ZAMAN</td>
<td>[1, 8)</td>
<td>6</td>
<td>11, 16, 17, 19, 32, 35</td>
</tr>
</tbody>
</table>

After the reliability analysis and results, the data set and scales were ready to use for data mining. Before I started using data mining techniques, I created an excel version of the data file then converted into *.arff and then converted into *.csv to be able to use these files in WEKA [13].

As long as I was looking for real associations and relations between groups of students, all the tests were done in supervised mode where the dataset is trained with in itself. %66 of the data is separated and used as dataset, rest is the training set. This supervised training might also lead the results to an unhealthy place where the data is not reliable and weak. But I did overcome this problem by applying reliability analysis on SPSS. [14] I also examined my dataset using WEKA on some data mining algorithm that is Apriori and Naive-Bayes [15]. Grade distribution for students are pointed by Naive-Bayes as follows:
Table 3.2: Grade Distribution

<table>
<thead>
<tr>
<th>Letter</th>
<th>A-</th>
<th>C-</th>
<th>F</th>
<th>C+</th>
<th>B</th>
<th>C</th>
<th>A-</th>
<th>B+</th>
<th>D+</th>
<th>I</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>.26</td>
<td>.03</td>
<td>.04</td>
<td>.06</td>
<td>.07</td>
<td>.04</td>
<td>.32</td>
<td>.007</td>
<td>.09</td>
<td>.02</td>
<td>.01</td>
</tr>
</tbody>
</table>

When I ran prepared my dataset for association rule generation with WEKA, Apriori Algorithm, I had 10 rules discovered from the dataset. These rules are given below:

**Rule 1:** If the student’s Motivation is Good, Self-testing is Good, and Self Study is Good then Testing Strategies is found to be Good 78  
conf: (0.97)

**Rule 2:** If the student’s Study-Aids is Good, and Motivation is Good, Self-testing is Good then Testing Strategies is found to be Good 80  
conf: (0.96)

**Rule 3:** If the student’s Attitude is Good, Motivation is Good, and Self-testing is Good then Testing Strategies is found to be Good 78  
conf: (0.96)

**Rule 4:** If the student’s Attitude is Good, Motivation is Good, and Self Study is Good then Testing Strategies is found to be Good 75  
conf: (0.96)

**Rule 5:** If the student’s Motivation is Good, Self-testing is Good, and Information Processing is Good then Testing Strategies is found to be Good 75  
conf: (0.96)

**Rule 6:** If the student’s Selecting Main Topics is Good, Motivation is Good, and Self-testing is Good then Testing Strategies is found to be Good 84  
conf: (0.96)

**Rule 7:** If the student’s Selecting Main Topics is Good, Motivation is Good, and Self Study is Good then Testing Strategies is found to be Good 79  
conf: (0.95)

**Rule 8:** If the student’s Motivation is Good, and Self-testing is Good then Testing Strategies is found to be Good 95  
conf: (0.95)

**Rule 9:** If the student’s Selecting Main Topics is Good, Testing Strategies is Good, and Information Processing is Good then Motivation is found to be Good 72  
conf: (0.95)

**Rule 10:** If the student’s Study-Aids is Good, Testing Strategies is Good, and Self Study is Good then Motivation is found to be Good 75  
conf: (0.95)

It can be seen that students had 10 different combinations of good aspects where 8 of them result at strategy scale’s being good where strategy – also known as test taking - scale assesses a student’s use of test preparation and test taking strategies. They are expected to learn effective techniques for preparing for and taking tests.

4. Conclusion and Discussion

These results were expected since the course was project depended and students needed to take a test on their projects to improve their learning, and best way for students to improve their learning was modifying their preferences on learning strategies and test-taking.

After all these rules generated by Apriori algorithm, I decided to check LASSI scales’ correlation with each other on SPSS and I ran Pearson Correlation and used a quick regression on the same dataset to see the data distribution, just to be sure that these rules are statistically true and whether these scales are significant.

Correlation showed that the scales named as Selecting Main Topics, Study-Aids, Attitude, Motivation, Self-Motivation, Self-Confidence, Self-testing, Testing Strategies have the highest correlation rates with others and Time Management has the lowest correlation rate. After this step correlation scatter plots are examined. Then I made a Linear Regression Analysis with SPSS and checked these scales’ significance.

The Linear Regression results were pointing at Selecting Main Topics, Study-Aids, Self-Motivation, Self-Confidence, Meta, Self Study and Information Processing scales being significant with a significance value less than 0.05; rest of the scales we used in LASSI were found to be insignificant and Time Management scale had the
biggest significance value that is 0.988 where it has no effect on the results and data, in a way time management does not effect learning Testing Strategies and does not help the students improve their learning.

As the linear regression analysis results show, the Apriori rules generated by WEKA are seem to be depended since these scales are statistically reliable, correlated and significant. When highly correlated but insignificant scales form an association rule with significant scales, we generally end up with a significant scale that is “Testing Strategies”. I also found out that if significant scales of LASSI form an association with in themselves, the result is also a significant scale, and relations between significant and insignificant scales rarely lead us to insignificant scales.

Acknowledgements

Due to limitation on paper size, some statistics and data is not included. These information will be provided by corresponding author on request.

References

[2] Hitendra Pillay - Cognitive processes and strategies employed by children to learn spatial representations @ Learning and Instruction, Volume 8, Issue 1, February 1998, Pages 1-18