An overview of Web Mining in Education

Alexandros Kleftodimos
Dept of Applied Informatics, University of Macedonia,
Thessaloniki, Greece
&
TEI of Western Macedonia, Greece
kleftodimos@kastoria.teikoz.gr

Georgios Evangelidis
Dept of Applied Informatics, University of Macedonia
Thessaloniki, Greece
gevan@uom.gr

ABSTRACT

This paper provides an overview of the advances on web mining in the domain of education by categorizing research on the field using the web-mining taxonomy: web content, web usage and web structure mining. Previous reviews focused on data mining methods applied to data derived from educational software applications (web based or not) and institutional administrative systems as well as from data gathered in typical classroom environments. The overview focuses on knowledge acquired from web based educational environments and the open web. More specifically it addresses: a) applications that dynamically update their content by extracting relevant educational information from the open web in order to meet user specific needs, b) applications that extract knowledge from usage data coming from educational web-based environments, and, c) applications that obtain knowledge coming from structure data such as links or social network connections that exist in e-learning applications and the open web.

Categories and Subject Descriptors

K.3.1 **[Computers and Education]:** Computer Uses in Education- *Computer-assisted instruction, Distance learning, Collaborative learning,* H.2.8 **[Database management]**: Database Applications-*Data mining,* I.5.3 **[Pattern recognition]**: Clustering-*Algorithms,* H.3.3 **[Information storage and retrieval]:** Information Search and Retrieval-*Clustering, Information filtering, Retrieval models, Search process*

Keywords

Web mining, Data mining, Education

1. INTRODUCTION

The term web mining was introduced by Etzioni in 1996 [8] to denote the use of data mining techniques to automatically discover web documents and services, extract information from web resources, and uncover general patterns on the Web.

Over the years, web mining research has been extended to cover the use of data mining and similar techniques to discover resources, patterns, and knowledge from the Web and web-related data [5].

Web mining is broadly interdisciplinary, attracting researchers from various science fields, such as databases, artificial intelligence, statistics, cognitive social theory, pedagogy, psychology, linguistics and so on.

The World Wide Web is today a huge and widely distributed source of information and applications, and has advanced to a point where it is affecting almost every science field and every aspect of the human activity and communication. As a consequence, web mining has applications in many areas, such as e-commerce, e-services, administration, medicine, and areas of societal benefit, such as e-government, politics, public security and crime investigation [13]. Another area affected by the advances in web mining is education and this is the focus of the present work.

Today there is a vast amount of educational resources that can be found all over the Internet, from University web sites to tutor personal web pages and recently the social web with sites like Facebook, Blogger and YouTube, where educational material is hosted in text and multimedia form. However, educational content that exists in the web is in unstructured or semi-structured form and in many different formats. This makes the task of discovering and organizing this content in an automated way a complex task.

The developments in educational technology have introduced a broad range of educational applications and platforms that are mainly web-based. These applications can be interactive and they can support collaborative learning. The educational content in these applications is also in the form of text or multimedia. These applications can also combine entertainment with educational tasks and in this case they are called edutainment applications. They may also contain other interesting features that aid the learning process. For example, they can contain "intelligence" in order to provide the learner with individualized instructions (i.e., intelligent tutoring systems) and adapt their content to the learners' needs (i.e., adaptive hypermedia systems). A variety of other platforms are also used in education today, for storing, organizing and distributing educational content and for providing the means of communication between all stakeholders (students, tutors, etc.). Examples of such applications are the Learning Management Systems (e.g., Moodle) and Social Networking engines that have been designed and developed for institutional use (e.g., ELGG Social Software).

Data mining (DM) research in education has been recorded since the mid 90s [25] and the interest in this field of research has been increasing since then.

PCI 2013, September 19 - 21 2013, Thessaloniki, Greece
Copyright 2013 ACM 978-1-4503-1969-0/13/09…$15.00.
http://dx.doi.org/10.1145/2491845.2491863

A considerable number of publications have been presented in conferences and journals related to Educational Technologies. The interest increased over the years and Special Workshops were dedicated to this particular research area in a number of conferences that took place from 2000 to 2007 (e.g., 5th International Conference in Intelligent Tutoring Systems - ITS'00, 20th National Conference on Artificial Intelligence AAAI'05, 13th Artificial Intelligence in Education Conference AIED'07, 7th IEEE International Conference in Advanced Learning Technologies - ICALT'07)[26]. Besides the ongoing publications that are presented in conferences and journals related to e-learning, this area has advanced today to the point where an International Conference on Educational Data Mining (EDM) is being organized annually since 2008. Annual conferences on EDM were joined by the Journal of Educational Data Mining, which published its first issue in 2009. The first Handbook of Educational Data Mining was published in 2010 and in 2011 the first International Educational Data Mining Society (IEDMS)[1] was formed with a mission to promote research in the interdisciplinary field of educational data mining, and to organize the EDM conferences and journal.

Educational data mining, according to the EDM community[1] definition, is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the settings in which they learn in. The educational settings include the typical classroom environments (offline education), e-learning applications and information systems that hold data related to the educational process such as student attendance, performance and profile information. Although there are research papers that apply DM on data gathered in classroom environments (e.g., student's be-havior/performance, curriculum, etc.), and data coming from institutional administrative systems, most of the research papers presented in EDM conferences deal with DM techniques applied on data coming from learning management systems (LMS), intelligent tutoring systems (ITS) and adaptive hypermedia systems (AEHS)[26]. These applications are mainly web based and data used from these educational applications for data mining purposes is mainly learner usage data and to a smaller extend content and structure data.

Three reviews have been produced by recognized members of the EDM community in order to record the developments in the field [25][26][1]. In these extensive and analytical reviews research was categorized according to different taxonomies. Research categorization was either based on the data mining techniques used (i.e clustering, classification and outlier detection; association rule mining and sequential pattern mining; and text mining) and on the areas of application of EDM methods (i.e., improvement of student models, pedagogical support, looking for empirical evidence to refine and extend educational theories etc.). However, besides the main focus of the EDM community which is knowledge extraction from e-learning applications, there is also research that focuses on extracting valuable educational knowledge from the open Web. More specifically, there is research that aims at discovering, organizing and aggregating valuable educational content from the open web, using web mining techniques, and incorporating this content in eLearning applications. The aim of this overview is to also identify and report this ongoing research. Therefore, this overview diverges from the focus of the EDM community, which deals mainly with knowledge extraction from data derived from eLearning applications because it provides a broader perspective of the applications of web mining in education by incorporating research on discovery and classification of educational content, as well as knowledge acquisition from open web resources in general.

The paper will organize research associated with web mining in education using the prevalent web mining taxonomy and its three main categories: web content, usage, and structure mining. Although we will be referring quite often throughout this text to the data mining techniques that are frequently used (i.e., classification, clustering, etc.) it is outside the scope of this work to describe in detail these techniques.

This paper is organized as follows: the web-mining taxonomy is described in Section 2, the research in educational web mining is categorized with respect to the web mining categories in Section 3, a discussion is carried out in Section 4 and the paper concludes in Section 5.

2. WEB MINING TAXONOMY

Etzioni in 1996 [8] proposed an organization of web mining into the following sub tasks: (a) resource discovery, for locating unfamiliar document and services on the web, (b) information extraction, for automatically extracting specific information for newly discovered documents, and, (c) generalization, for uncovering general patterns from web sites. A year later, in 1997, R Cooley, B Mobasher and J Srivastava [7] presented a taxonomy with two main categories: web content and web usage mining. Today, this taxonomy, extended by one more category (structure mining) is widely accepted and this is evident from a large number of publications. However, the definitions of the categories have been changing and will continue to change to map the web evolution and to embrace new research methods.

2.1 Web Content Mining

Web content mining describes the discovery of useful information and the extraction of knowledge from the contents of the Web. Application of text mining to web content has been the most widely researched topic. Issues addressed in text mining are: topic discovery, extraction of association patterns, clustering of web documents and classification of web pages [34].

Web content has been increasing steadily over the years and this process has been accelerated with the advent of Web 2.0 and the user generated content. Today the web is not dominated by commercial and institutional sites anymore and content can be uploaded by every individual with a connection to the Internet. Web content may encompass a very broad range of data. Besides the typical hypertext or the document data that exists in various formats (i.e., pdf), the Web also contains large volumes of multimedia data, such as images, video and animations. The advent of high speed broadband connections and the arrival of video and image sharing sites like YouTube and Flickr, contributed to the increase of multimedia data on the Internet. This gives particular importance to a research field related to web content mining called multimedia web mining. Multimedia web mining focuses on mining methods applied to multimedia data. Another field that is closely related to web content mining (or is part of content mining, according to some researchers) is opinion mining and sentiment analysis. This field uses natural language processing (NLP) and text mining techniques in order to automatically identify and recognize opinions and emotions in text derived from user posts and comments in web forums, blogs and other social media sites.

Web content mining is also closely related to information retrieval (IR). Researchers have claimed that resource or document discovery on the web is an instance of web content mining and others associate web mining with intelligent IR.

[1] www.educationaldatamining.org

Kosala and Blockheel [11] claim in their work that web mining is part of the (web) information retrieval (IR) and information extraction (IE) process since web mining methods such as document classification or categorization are frequently used in both in information retrieval (IR) and information extraction (IE). The authors propose that research in web content mining could be differentiated from two points of view: Information retrieval (IR) and Database (DB) views. The goal of web content mining from the IR view is mainly to assist or improve the information finding or the filtering of the information, while the goal of web content mining from the DB view is to model data on the web and to integrate them so that more sophisticated queries other than the keywords based search could be performed. They also state that IR view has global scope and spans the entire Web, while the DB view has local scope and spans a specific web page.

2.2 Web Usage Mining

Web usage mining focuses on the discovery of knowledge from user activity while browsing internet sites and web applications. This activity data can be found in various formats and in a number of places, such as web server access logs, proxy server logs, browser logs, cookies, and databases of online applications, depending on whether the collection is carried out at server level, client level or proxy level. Browsing activity can also be recorded with the aid of small software programs that enhance the functionality of the client browser (i.e., google toolbar, firefox addons, chrome extentions). The browsing activity captured by the browser can then be stored in remote databases. Web usage mining exploits user data in order to go one step further than log analysis tools, which provide statistics about site visitor activity (i.e., page hits, times of visits, hits per hour, etc.). Acquiring knowledge about the user behaviour and usage patterns has proven very useful for a number of application areas, such as e-commerce, video games and elearning that is the focus of this paper.

2.3 Web structure Mining

Knowledge extraction based on the structure of the Web is known as web structure mining. The Web contains a large variety of objects with no unifying structure, such as web pages, multimedia data, etc. These objects can be connected to one another either by hyperlinks or other types of social connections. Mining these types of connections is known as hyperlink network analysis (HNA) or social network analysis (SNA) or structural analysis. The social network notion is derived from sociology. A social network is a set of nodes (people, organizations or other social entities) connected by a set of relationships, such as friendship, affiliation or information exchange [40]. Social network analysis (or social network mining) is a set of research procedures for identifying structures in social systems based on the relations among system components. Hyperlink network mining on the other hand casts hyperlinks between web sites as social and communicational ties applying standard techniques from SNA [22]. Thus, web structure mining is the process of obtaining knowledge from a web graph (e.g., identifying communities and influential websites or blogs), which is formed either by extracting link connections between web pages or by extracting other types of connections that result from interaction and communication activity amongst users in the social web. Such activity can be the friendship relation between Facebook users, the “likes” and comments attached to a Facebook post, the follower connection in Twitter, the subscription in a You-Tube channel, the rating or the comments in a You-Tube video or a blog post. These social connections however can also be viewed as hyperlinks from a strictly technical point of view. HITS and PageRank are two page ranking algorithms that are based on web structure mining. These algorithms calculate the importance of web pages from the link structure of the web. Similar techniques based on the analysis of the linking behaviour amongst web pages, can also be applied to the social Web. For example these techniques can be used to calculate the importance of blogs in the blogosphere (e.g., Technorati authority measurement) or the importance of individuals in web communities (e.g., influential social media users). Similar techniques are used for ranking the importance of scientific journals (e.g., impact factor).

3. APPLICATIONS OF WEB MINING IN EDUCATION

Several procedures were followed to produce this overview. First the Scholar Google and DBLP databases were searched extensively based on a range of key terms. The key terms used were always a combination of terms associated to web mining and terms referring to education (i.e., education, educational content, e-learning etc.). The terms associated to web mining included content, usage and structure mining, classification, clustering, association rule, information retrieval and focused crawling, mashups, social network analysis, and link analysis. Furthermore the reference section for each relevant article found was searched in order to find additional articles. More than 100 articles were examined. Finally all EDM conference papers from 2008 to 2012 were scanned in order to be categorized according to the Web mining taxonomy. The overview concentrates mainly on research carried out from 2005 onwards

3.1 Web Content Mining in Education

Content mining methods can be viewed from an information retrieval (IR) or database (DB) view. IR has global scope while DB view has local scope. Global scope spans the entire web while local scope spans a specific web page. These web pages can either be ordinary sites with educational content or content that exists in e-learning applications. In this section we will examine applications of web content mining in education giving focus to the IR view.

Web content mining in education encompasses methods used to retrieve meaningful educational content from the Web to meet user specific needs, and valuable knowledge that could help educators in the learning process and in decision making. The web resources that contain educational content are following the general trend of the web where information is constantly changing. New sites are coming into existence and old sites are either updated or withdrawn on a daily basis. The popularity of exchange and dissemination of content through the web has created a huge amount of educational resources, and the challenge of locating suitable learning references specific to a learning topic has become a big challenge [23]. Various researchers tried to address this specific problem by using information retrieval methods and data mining techniques such as classification and clustering. These methods were used either to provide the users with efficient tools to track topic specific educational resources or to build digital libraries and e-learning applications that update their content dynamically.

For example, Prashant *et al*. [23] attempted to give a solution to the problem of locating learning materials in the web tailored to user needs, by developing an intelligent repository of educational resources that updates its content dynamically. Their work used crawling, classification, and information extraction techniques for the task of identifying useful softwares/tools for education from the web. Tang and McCalla [38] developed an e-learning environment that updates its content in a dynamic manner by retrieving topic specific scientific articles from CiteSeer. Besides acquiring knowledge from the web with the use of information retrieval methods, this system also extracts knowledge from the users' interaction with the system (usage data) to recommend new articles to the user according to his/her interests. Therefore, the authors in this system combine content mining with usage mining methods, which are considered in the next section.

Focused crawling is an information retrieval method that can be used in order to discover and categorize educational content by topic area. A focused crawler can be defined as a web crawler that seeks, retrieves, indexes and maintains pages on a specific topic, which represents a relatively narrow segment of the web**.** There are examples of research where focused crawling is used in order to search and deliver appropriate learning material to users. Focused crawling was used by Biletskiy, Wojcenovic and Baghi [2] who developed a technical solution for tracing learning objects from digital libraries on the Web, for further search and delivery to learners. Premlatha and Geetha [24] used focused crawling to traverse the Web and collect educational resources, categorized by topic area, for an e-learning content management system. Schmitz *et al.* [30] used focused crawling in their system called Courseware Watchdog, which aimed at finding and visualizing educational material on the web and in peer-to-peer networks according to user needs. Lawless, Hederman, and Wade developed a system [56] named Open Corpus Content Service (OCCS) that enables the discovery and classification of educational content from open corpus sources in the web and facilitated the incorporation of such content into elearning systems. OCCS discovers, harvests and indexes content with the use of a focused crawler, content classifier and indexer. These are some examples but we have also spotted some others in the literature [15][14].

The volume of the Web is increasing rapidly and this process has been accelerated by the introduction of the social web and the user generated content that is added to it (e.g., YouTube). The content of Web 2.0 can be exploited for educational purposes. APIs that provide interfaces to Web 2.0 services, crawling techniques and other content mining methods can be used in order to retrieve and aggregate useful educational content. An example of such an effort is the research conducted by Hong *et al*. [9] who used information retrieval, clustering and natural language processing (NLP) techniques to construct a multimedia encyclopedia called Mediapedia, which is automatically produced and dynamically updated by leveraging on the online Web 2.0 resources (i.e., Wikipedia and FlickR). The system crawls diverse images from Flickr and uses clustering to generate exemplar images for specific concepts. It then associates the exemplar images with relevant contents derived from Wikipedia, by using NLP techniques and noisy tag filtering. As a last step the system presents the text contents for each concept with a synchronized video presentation that is constructed from the exemplar images.

Content mining can also be used in combination with other techniques in order to produce new content from existing content on the web. Microsoft researchers Scott, Liu and Zhou [31] used web content mining and NLP techniques in order to develop Engkoo, a system for exploring and learning languages by mining translation knowledge (a massive set of bilingual terms and sentences) from across billions of web pages. Engoo was primarily used for Chinese users who are learning English, but the authors claim its underlying technology is language independent and can be extended in other language domains.

Although so far we focused on content mining from the Information Retrieval (IR) point of view, and specifically, in applications that extract knowledge from the open web, content mining methods can also be applied to content that exists inside web pages and e-learning platforms. Research in this field includes text mining and sentiment mining techniques applied to learner generated data (e.g., student discussions in forum threads) [16][10], text mining techniques applied on e-learning material in order to construct a concept map [33], natural language processing and other content based methods that automatically evaluate student knowledge about a topic by comparing student input with various sources of knowledge that describe the topic accurately [29], etc.

Finally, web content mining in the context of education can also be used to deliver knowledge to educators in order to support decision-making. Ciolac, Luban and Dobrea [6] proposed an automated web content mining framework to evaluate the compatibility between university curricula and market qualification needs. To achieve this, they used software information agents to derive information from university and e-recruitment web sites.

3.2 Web Usage Mining in Education

Advances in educational technology are constant and have a considerable impact on the way teaching is conducted. Technology developments affect almost every area of education. Interactive learning applications have been developed for a broad range of subjects such as mathematics, science, language learning, and for different areas of education, such as special education and teacher education. These applications can contain "intelligence" and adaptability to the user needs (i.e., Intelligent Tutoring systems, Adaptive Hypermedia Systems). In other words, these applications can match their instructional content or adapt their educational content to the learner's changing state of knowledge, performance or interests, motivations and identity. These applications can be operated offline but many are web-based and can be accessed through the Web. Learning management systems are also web-based. Almost all higher institutions use in one way or the other some form of web based environment for educational purposes and web usage data coming from these environments make up a gold mine for educational web usage mining.

Web usage mining has been used in the past by various disciplines, such as e-commerce, to obtain knowledge about the consumer interests and navigational behavior. This knowledge was then used to increase consumer sales. Equivalently, web usage mining in educational environments can be used to understand student learning patterns and the environment in which the learning process takes place. Most papers in the field of web mining in education are dedicated to data mining methods performed on usage data in order to acquire a better understanding on the student learning patterns. The EDM community research focuses on knowledge extracted from educational settings and these settings are mainly educational environments and to a smaller extend typical classroom environments and administrative information systems that hold student profile and performance information.

Web-based learning environments are able to record student actions and interactions (in log files and databases), and hence, are able to provide a huge amount of learning profiles. Web usage mining in educational environments can be used to answer questions like: "which are the most common navigational routes followed by the learners and which of these routes lead to effective learning?" [27], "how effective is an educational environment?", "Which are the pages/topics that students skip and what is the amount of time the students spend with a single page, a chapter or the full course?" "can we predict the performance of the learners by examining their activity in an LMS?" [42], "how likely a student is to give a correct answer to a problem in an intelligent tutoring system?" [4], "which are the main student categories according to their learning patterns in an LMS?" [28], "can we predict the student failure and dropout rate?" [18] etc.

In this category of Web Usage mining from data coming from educational environments we find a great number of publications mainly presented in the EDM conferences and workshops but also in other conferences and journals dedicated to education. Web usage data in EDM research is mainly derived from learning management, intelligent tutoring, and adaptive hypermedia systems, and to a smaller extend, from other web-based applications that assist and evaluate learning, such as, educational games, online tests and quizzes, etc. Until 2005, data universally came from the research group conducting the analysis. In other words, in order to do educational data mining research, a researcher first needed to collect one's own educational data [1]. This is no longer a necessity due to the fact that Pittsburgh Science of Learning Center opened a public data repository, the PSLC DataShop[2], which makes substantial quantities of data from a variety of online learning environments available for free to any researcher worldwide. In this way, a data mining researcher who does not have access to usage data stored in log or database files of educational environments, can use DataShop data in order to conduct research but also to compare research findings with other researchers that have used the same data sets, or to build on other researchers' past efforts [1].

3.3 Web Structure Mining in Education

Web structure mining also has applications in education. In the last few years, there has been an increasing focus on social software applications and services as a result of the rapid evolution of Web 2.0. Web 2.0 tools such as blogs, wikis and social networking platforms (i.e., Facebook, Twitter) and other types of collaborative learning tools are adopted today by many educators [20] [39][49]. Web 2.0 features are also incorporated in known e-learning platforms (e.g., Moodle). The exploration of social relationships developed in these environments but also in other collaborative learning environments through user interactions and collaboration can be used for drawing important conclusions about the learner behaviour and to discover communities that are formed in these environments. The SNA approach offers a method for mapping interactions amongst actors within a network, visualizing connectedness, and quantifying some characteristics of these processes within a community [12]. The actors (or nodes) of a network graph in e-learning settings typically represent students or teachers. Nodes may also represent other entities if the domain is the open web (e.g., web pages, digital libraries) such as institutions and articles. E-learning in work settings can also involve co-workers and collaborators.

In the literature we encounter a number of publications concerning SNA in online learning environments. De Laat *et al.* [12] provided a general overview of how SNA is applied in computer-supported collaborative learning research. Stepanyan, Borau and Ullrich [35][36] used the microblogging platform Twitter in the educational process and carried out SNA to identify the patterns and trends of network dynamics. Xu and Recker [41] collected user activity data in a peer production educational system, and followed a social network perspective in analysing this data. They focused their research on the relationships between users (teachers in this case) in order to identify important users and like-minded user groups. Crespo and Antunes [50] use effective techniques for social networks analysis to quantify the performance of students in teamwork. There are of course more instances of SNA applications in e-learning [32] [50][51].

In the category of structure mining, we also need to include the large number of applications that use books and articles stored in scientific databases (i.e., Citeseer, Scholar Google, DBLP) in order to obtain knowledge from article content and the citation graph. These applications use web content mining and web structure mining methods for tasks such as identification of web communities, areas impacted by specific research, co-authorship networks [17], the network of conference participants [19] etc. For example, Tang *et al.* [37] developed the ArtMiner System that aims at extracting and mining academic social networks. Their goal was to build a system that extracts researcher profiles automatically from the web and integrates these profiles with publication data derived from online digital libraries. Furthermore the system models the entire academic network and provides search services for this network. Cai *et* al. [3] used data mining techniques to discover hidden communities in heterogeneous social networks, and used the DBLP dataset to demonstrate the effectiveness of their method. The applications in this field are numerous and the whole field of citation analysis could fall under this category.

There are also publications that use link analysis to identify inter-institutional relations [21]. Although academic and research community identification is not directly related to the learning process, it may be of interest to its stakeholders and can be valuable to the educational progress.

Finally, according to some researchers [26], collaborative filtering or social filtering is also related to social network or structural mining. Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting taste preferences from users that share a number of similar interests. In this case, the graph examined is the graph created by the user preferences. Collaborative filtering is used for producing personal recommendations.

4. DISCUSSION

Having given examples of applications of web mining in education, it must be stated that the mining tasks of web content, usage, and structure mining can be used in isolation or in combination [11]. There are many applications that use more than one of these tasks in combination in order to accomplish their objectives. For example there is research that combines social network analysis with content analysis to gain a richer picture of networked learning, investigating not only who is talking to whom, but what they are talking about and why they are talking in this way. For example De Laat *et al.* [57] applied SNA to visualise the social structure of Networked Learning Communities, Content Analysis (CA) to identify learning and teaching processes and Context Analysis (CxA) to study students' personal experiences and intentions. Recommendation or personalization techniques (content based and collaborative filtering) rely mainly on usage data but also on educational content and structural information that reside in e-learning applications in order to provide users with the information they want or need, without expecting from them to ask for it explicitly. There are of course other examples where we have a combination of mining methods for achieving a task and some examples have been given in earlier sections [38][33].

[2] https://pslcdatashop.web.cmu.edu/

Another point of consideration is that the distinction amongst the web mining categories has never been clear-cut [11]. These distinctions are probably even more blurred today with the evolution of the participatory Web 2.0. For example, text posted by users on a social web platform can be subject to all three categories. Web content mining techniques such as text mining and opinion mining can be used to obtain knowledge from the textual data (e.g., the sentiment polarity of the document), web usage mining methods can be used to acquire knowledge about the user behavior by exploring more quantitative aspects, such as the number of posts a user makes and the amount of text in a post, and, finally, text postings (i.e., comments) in a blog, forum, or another social networking platform can be viewed as interaction data between the users, and therefore, be subject to web structure mining. The web mining taxonomy is heavily related to the Web 1.0 environment and the arrival of Web 2.0 may pose the need for new taxonomy theories or redefinition of the existing ones.

Web usage mining methods in education appear considerably more frequently in literature when compared to content mining and structure mining methods. As explained in the beginning of this paper, the EDM community research focuses on extracting knowledge from educational settings. Although these settings are not restricted to e-learning applications most of the research concentrates on data mining methods applied on usage data coming from intelligent tutoring systems (ITS), adaptive hypermedia systems (AEHS) and learning management systems (LMS), which are used in higher but also in other levels of education.

Content mining research focusing on discovering, organizing and aggregating educational content is limited. In the search carried out in scientific databases (DBLP and Scholar Google) and in EDM conference proceedings, we were able to spot only a handful of instances and most of them are included in the references (e.g., [23][30][38][2][24][56]). This research is of great interest since it proposes automated ways to utilize the vast amount of educational material that resides in the web. However the task of discovering and organizing relevant educational material is a complex task due to the fact that web content is in unstructured or semi structured form and educational material exists in many different formats (e.g., notes, articles, videos, powerpoint presentations etc.).

Social network analysis (or structural analysis) methods in web based education appear more frequently than IR content mining methods but considerably less frequently than usage mining methods. Research efforts that use social network analysis to extract knowledge from educational applications made their appearance after 2005 according to the EDM review carried out in 2010 [26]. In this review 15 out of the 306 references examined were related to social network mining. By looking at the publications from the top five Journals in the Scholar Google Educational Technology impact list (i.e., Computer Education, British Journal of Educational Technology, Journal of Computer Assisted Learning, Educational Technology & Society, Educational Technology Research and Development) for the year 2012 together with the EDM conference proceedings for the same year, we identify 6 publications that use social network mining techniques [50][51][52][53][54][55]. Thus, we could conclude that research on using social network analysis in the educational domain is increasing and will most likely continue to increase since the use of educational Web 2.0 applications and collaborative tools is a relatively new and emerging trend. The fact that Web 2.0 tools usage is an emerging trend in education is evident from the number of publications that appear in the five journals mentioned above for the year 2012. An examination of the aforementioned publications will reveal that there are two special issues dedicated to the use of Web 2.0 tools in education with 6 and 5 publications accordingly that are directly related to the topic [45][46]. Furthermore we were able to track at least 5 instances related to the use of Web 2.0 tools in education (i.e.,[43],[44],[47][48][49]) in the other 3 journals.

5. CONCLUSIONS

In this paper we provided an overview of the applications of web mining in education by categorizing applications with respect to the three web mining categories: web content, web usage, and web structure mining. Content mining methods are used to retrieve, organize and aggregate educational resources from the web and to extract knowledge from the content that lies inside e-learning platforms. This paper also reports research that uses content mining methods from the IR view, which was absent from previous reviews that focused mainly on knowledge extraction from usage data coming from e-learning environments. Web usage mining methods are used to obtain knowledge from learner activity data in web based educational settings. Web structure mining methods, such as social network analysis, extract knowledge from graphs that are formed from interactions and collaboration amongst users in e-learning applications, but also from links that reside in the web between educational related entities (e.g., inter- institutional web link connections) as well as co-citations in scientific articles.

Web usage mining methods in web based education appear considerably more frequently in literature when compared to content mining and structure mining methods. Research using content mining methods from the IR view in order to discover and organize or aggregate educational content from different sources is to the best of out knowledge limited. On the other hand research using social network analysis in educational environments started to appear from 2005 onwards and there are signs that it is increasing. This research will most likely continue to increase since the use of educational Web 2.0 applications and collaborative tools is a relatively new and emerging trend.

6. REFERENCES

[1] Baker, R., Yacef, K. 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1, 1, 3-17.

[2] Biletskiy, Y., Wojcenovic, M., and Baghi, H. 2009. Focused Crawling for Downloading Learning Objects. An Architectural Perspective. *Interdisciplinary Journal of E-Learning and Learning Objects,* 5, 169-180.

[3] Cai, D., Shao, Z., He, X., Yan, X., and Han, J. 2005. Mining hidden community in heterogeneous social networks. In *Proc. 3rd International Workshop on Link Discovery*, 58-65.

[4] Cetintas, S., Si, L., Xin, Y., and Hord C. 2009. Predicting correctness of problem solving from low-level log data in intelligent tutoring systems. In Proc. *2nd Int. Conf. on Educational Data Mining*, 230-239.

[5] Chen, H., and Chau, M. 2004. Web mining: machine learning for web applications. *Annual Review of Information Science and Technology*, 38, 289-329.

[6] Ciolac, C., Luban, F., Dobrea, R. 2010. Web Content Mining Framework for Discovering University Formations' Compatibility with the Market Needs. *Review of International Comparative Management*, 11, 5, 1001-1016.

[7] Cooley, R., Mobasher, B., and Srivastava. J. 1997. Web mining: information and pattern discovery on the World Wide Web. In *Proc. 9th International Conference on Tools with Artificial Intelligence* (ICTAI '97), 558-567. DOI= http://dx.doi.org/doi:10.1109/TAI.1997.632303

[8] Etzioni, O. 1996. The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39,11, 65–68. DOI= http://dx.doi.org/doi:10.1145/240455.240473

[9] Hong, R., Tang, J.Z., Zha, J., Luo, Z., and Chua, T.S. 2010. Mediapedia: Mining Web Knowledge to Construct Multimedia Encyclopedia. *Advances in Multimedia Modeling*, Lecture Notes in Computer Science, 5916, 556-566. DOI= http://dx.doi.org/doi:10.1007/978-3-642-11301-7_55

[10] Kim, S., and Calvo, R. 2010. Sentiment Analysis in Student Experiences of Learning. In *Proc of 3rd Int. Conf. on Educational Data Mining*, 111-120.

[11] Kosala, R., and Blockeel, H. 2000. Web mining research: a survey. *ACM SIGKDD Explorations Newsletter,* 2, 1, 1-15. DOI= http://dx.doi.org/doi:10.1145/360402.360406

[12] De Laat, M., Lally, V., Lipponen, L., Simons, R. 2007. Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *Computer-Supported Collaborative Learning*, 2,1, 87-103. DOI= http://dx.doi.org/doi:10.1007/s11412-007-9006-4

[13] Lappas, G. 2008. An Overview of Web Mining in Societal Benefit Areas. *Journal of Online Information Review*, 32,2 179-195. DOI= http://dx.doi.org/doi:10.1108/14684520810879818

[14] Lawless, S. 2007. Open Corpus Learning Content; Harvesting Knowledge to provide Equitable Access to Education for All. In *Proc. Education Without Borders Conference*, EWB2007, Abu Dhabi, United Arab Emirates.

[15] Lee, D., Kim, H., Kim, E., Yan, S., Chen, J., Lee, J. 2008. LeeDeo: Web-Crawled Academic Video Search Engine. In *Proc. 10th IEEE International Symposium on Multimedia* (ISM'08), 497-502. DOI= http://doi.ieeecomputersociety.org/10.1109/ISM.2008.105

[16] Lin, F., Hsieh, L., and Chuang, F. 2009. Discovering genres of online discussion threads via text mining. *Journal of Computers & Education*, 52, 2, 481–495. DOI= http://dx.doi.org/doi:10.1016/j.compedu.2008.10.005

[17] Liu, X., Bollen, J., Nelson, M., Sompel, H. 2005. Co-Authorship Networks in the Digital Library Research Community. *Information Processing and Management*, 41, 1462-1480. DOI= http://dx.doi.org/doi:10.1016/j.ipm.2005.03.012

[18] Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., and Loumos, V. 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53, 3, 950-965. DOI= http://dx.doi.org/doi:10.1016/j.compedu.2009.05.010

[19] Matsuo, Y., Tomobe, H., Hasida, K., and Ishizuka, M. 2003. Mining Social Network of Conference Participants from the Web. In *Proc. IEEE/WIC International Conference on Web Intelligence* (WI 2003), 190-193. DOI= http://dx.doi.org/doi:10.1109/WI.2003.1241192

[20] Minocha, S. 2009. Role of social software tools in education: a literature review. *Education and Training*, 51, 5, 353-369. DOI= http://dx.doi.org/doi:10.1108/00400910910987174

[21] Ortega, J., Aguillo, I., Cothey, V., Scharnhorst, A. 2007. Maps of the academic web in the European Higher Education Area – an exploration of visual web indicators. *Scientometrics*, 74, 2, 295-308. DOI= http://dx.doi.org/doi:10.1007/s11192-008-0218-9

[22] Park, H., and Thelwall, M. 2003. Hyperlink Analyses of the World Wide Web: A Review. *Journal of Computer-Mediated Communication*, 8, 4.

[23] Prashant, M., Ankit, D., Kumar, S.M., Kumar, A., and Sasikumar, M. 2010. Building A Knowledge Repository of Educational Resources using Dynamic Harvesting. In *Proc. IEEE International Conference on Technology for Education* (T4E), 157-163. DOI= http://dx.doi.org/doi:10.1109/T4E.2010.5550041

[24] Premlatha, K., and Geetha, T. 2011. Focused Crawling for Educational Materials from the Web. *International Journal of Computer Science & Informatics*, 1, 2, 26-29.

[25] Romero, C., and Ventura, S. 2007. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications,* 33, 1, 135-146. DOI= http://dx.doi.org/doi:10.1016/j.eswa.2006.04.005

[26] Romero, C., and Ventura, S. 2010. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions On Systems, Man, and Cybernetics,* Part C: Applications And Reviews, 40,6, 601-618. DOI= http://dx.doi.org/doi:10.1109/TSMCC.2010.2053532

[27] Romero, C., Gutiérrez, S., Freire, M., and Ventura, S. 2008. Mining and Visualizing Visited Trails in Web-Based Educational Systems. In *Proc of the 1st Int. Conf. on Educational Datamining*, 182-186.

[28] Romero, C., Ventura, S., Espejo, P., and Hervás, C. 2008. Data Mining Algorithms to Classify Students. In *Proc. 1st Int. Conf. on Educational Datamining*, 182-186.

[29] Rus, V., Lintean, and M., Azevedo, R. 2009. Automatic Detection of Student Mental Models During Prior Knowledge Activation in MetaTutor. In *Proc. 2nd Int. Conf. on Educational Data Mining*, 161-170.

[30] Schmitz, C., Staab, S., Studer, R., Stumme, G., and Tane, J. 2002. Accessing Distributed Learning Repositories through a Course-ware Watchdog. In *Proc. E-Learn 2002 World Conf. E-Learning in Corporate, Government, Healthcare, and Higher Education*, 15-19.

[31] Scott, M., Liu, X., and Zhou, M. 2011. Engkoo: Mining the Web for Language Learning. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, 44–49.

[32] Shen, D., Nuankhieo, P., Huang, X., Amelung, C., and Laffey, J. 2008. Using Social Network Analysis to Understand Sense of Community in an Online Learning Environment. *Journal of Educational Computing Research Issue,* 39, 1, 17-36.

[33] Šimko, M., and Bieliková, M. 2009. Automatic Concept Relationships Discovery for an Adaptive E-course. In *Proc. of 2nd Int. Conf. on Educational Data Mining*, 171-179.

[34] Srivastava, T., Desikan, P., and Kumar, V. 2005. Web Mining: Concepts, Applications and Research Directions. *Foundations and Advances in Data Mining, Studies in*

Fuzziness and Soft Computing, 180, 275-307. DOI= http://dx.doi.org/doi:10.1007/11362197_10

[35] Stepanyan, K., Borau, K., and, Ullrich, C. 2010. A Social Network Analysis Perspective on Student Interaction within the Twitter Microblogging Environment. In *Proc. of 10th IEEE Int. Conf. on Advanced Learning Technologies*, 70-72. DOI=http://dx.doi.org/doi:10.1109/ICALT.2010.27

[36] Ullrich, C., Borau, K., and Stepanyan, K. 2010. Who Students Interact With? A Social Network Analysis Perspective on the Use of Twitter in Language Learning, *Lecture Notes in Computer Science*, 6383, 432–437. DOI= http://dx.doi.org/doi:10.1007/978-3-642-16020-2_33

[37] Tang, J., Zhang, J., Yao, Li, .J., Zhang, L., and Su, Z. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks, In *Proc. 14th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining,* 990-998. DOI= http://dx.doi.org/doi:10.1145/1401890.1402008

[38] Tang, T., and McCalla, G. 2005. Smart recommendation for an evolving e-learning system. *International Journal on E-Learning*, 4, 1, 105–129.

[39] Tekinarslan, E. 2010. Web 2.0 Technologies in Higher Education: A Review from a Faculty Member's Perspective. *Int. Conf. on New Trends in Education and Their Implications.*

[40] Wasserman, S., and Faust, K. 1994. *Social network analysis: Methods and applications*. Cambridge, NY: Cambridge University Press.

[41] Xu, B., and Recker, M. 2010. Peer Production of Online Learning Resources: A Social Network Analysis. In *Proc. 3rd Int. Conf. on Educational Data Mining*, 315-316.

[42] Zafra, A., and Ventura, S. 2009. Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming. In *Proc. of the 2nd Int. Conf. on Educational Data Mining.* 307-314

[43] Junco, R. 2012. The relationship between frequency of Facebook use, participation in Facebook, activities, and student engagement. *Computers & Education* 58, 1, 162–171. DOI= http://dx.doi.org/doi:10.1016/j.compedu.2011.08.004

[44] Bennett, S., Bishop, A., Dalgarno, B., Waycott, J., and Kennedy, G. 2012. Implementing Web 2.0 technologies in higher education: A collective case study. *Computers & Education,* 59, 2, 524–534. DOI= http://dx.doi.org/10.1016/j.compedu.2011.12.022

[45] Pachler, N., Ranieiri, M., Manca, S., and Cook, J. 2012. Editorial: Social networking and mobile learning. *British Journal of Educational Technology*, 43, 5, 707-710.

[46] Ravenscroft, A., Warburton, S., Hatzipanagos, S., and Conole, G. 2012. Editorial: Designing and evaluating social media for learning: shaping social networking into social learning. *Journal of Computer Assisted Learning*, 28, 3, 177–182.

[47] Chuang, P., Chiang, M.-C., Yang, C.-S., and Tsai, C.-W. 2012. Social Networks-based Adaptive Pairing Strategy for Cooperative Learning. *Educational Technology & Society*, 15, 3, 226–239.

[48] Tambouris, E., Panopoulou, E., Tarabanis, K., Ryberg, T., Buus, L., Peristeras, V., Lee, D., and Porwol, L. 2012. Enabling Problem Based Learning through Web 2.0 Technologies: PBL 2.0. *Educational Technology & Society*, 15, 4, 238–251.

[49] Aydin, S. 2012. A review of research on Facebook as an educational environment. *Education Technology Research and Development,* 60, 6, 1093-1106. DOI= http://dx.doi.org/doi:10.1007/s11423-012-9260-7

[50] Crespo, P., and Antunes, C. 2012. Social Networks Analysis for Quantifying Students' Performance in Teamwork. In *Proc. of 5th Int. Conf. on Educational Datamining,* 234-235

[51] Obsivac, T., Popelinsky, L., Bayer, J., Geryk, J., and Bydzovska, H. 2012. Predicting drop-out from social behaviour of students. In *Proc. of 5th Int. Conf. on Educational Datamining,* 103-109.

[52] López, M.I., Luna, J.M., Romero, C., and Ventura, S. 2012. Classification via clustering for predicting final marks based on student participation in forums. *In Proc. of 5th Int. Conf. on Educational Datamining*. 148-151

[53] Cadima, R., Ojeda, J., and Monguet, J. 2012. Social Networks and Performance in Distributed Learning Communities. *Educational Technology & Society*, 15, 4, 296–304.

[54] Oshima, J., Oshima, R., and Matsuzawa, Y. 2012. Knowledge Building Discourse Explorer: a social,network analysis application for knowledge building discourse. *Educational Technology Research and Development* ,60, 5, 903-92, DOI= http://dx.doi.org/doi:10.1007/s11423-012-9265-2

[55] Jimoyiannis, A., and Angelaina, S. 2012. Towards an analysis framework for investigating students' engagement and learning in educational blogs. *Journal of Computer Assisted Learning,* 28, 3, 222-234. DOI= http://dx.doi.org/doi:10.1111/j.1365-2729.2011.00467.x

[56] Lawless, S., Hederman, L., and Wade, V. 2008. OCCS: Enabling the dynamic discovery, harvesting and delivery of educational content from open corpus sources. In *Proc. 8th^{h} IEEE Int. Conf. On Advanced Learning Technologies*, 676-678.

[57] De Laat, M., Lally, V., Lipponen, L., and Simons, R.-J. 2006. Analysing student engagement with learning and tutoring activities in networked learning communities: A multi-method approach. *Int. Journal of Web Based Communities*, 2, 4, 394-412.