Applying Learning Analytics in an Open Personal Learning Environment

A quantitative approach

Eleni Koulocheri  Alexander Soumplis  Michalis Xenos
Software Quality Research Group
School of Science and Technology
Hellenic Open University
Patra, Greece
e-mail: [ekoulocheri, asoumplis, xenos]@eap.gr

Abstract—Every activity in the web leaves a digital trace. Analytics are used to measure all gathered traces along with the activities and the traffic that a user provokes, aiming to exploit the application-under-analysis with an optimal way in terms of time and cost. As a result, the question about how they applied into learning environment, raises. This paper discusses analytics dedicated to the learning process named learning analytics. After presenting their context, and a literature review regarding this relatively new term, this paper presents a methodology that is applied the academic year 2011-12 on HOU2LEARN Environment, an open educational environment set by Hellenic Open University. This methodology consists of social network analysis along with a set of metrics that is also presented. The procedure and the outcomes of this ongoing research so far are discussed.

Keywords: Learning analytics; social network; data; visualization; metrics; assessment

I. INTRODUCTION

Analytics are part of our daily digital life since they measure with a discreet way the activities and the traffic that a user provokes while interacting with an online application. Until now they were used aiming to exploit the application-under-analysis with an optimal way in terms of time and cost potentially. However, part of our daily digital life is also learning activities and since learning applications cannot be faced as common web applications, they cannot be analyzed with common means such as web analytics. Learning analytics come to cover this special and upcoming research area; This paper presents the definition of learning analytics (hereinafter, LA) and presents their application to HOU2LEARN, an open personal learning environment. LA are applied through two basic methods; through social network analysis (hereinafter, SNA) and through a set of metrics that is proposed, special designed for HOU2LEARN (hereinafter, H2L), as a part of an ongoing research in this upcoming field.

Section II discusses the large amount of data, called ‘Big Data’, that occurred from the digital learning traces that a learner or instructor leaves in a learning environment. It also presents the definition of LA that exploits Big Data in a way that helps the participants to understand and optimize learning and the environment that it occurs. Section III includes most common types of LA and focuses on types that are based on SNA and the measurement of discourse and learning trails. Section IV presents H2L, the open educational environment that runs under the auspices of the Hellenic Open University (hereinafter, HOU) and it is the test bed for this LA research. It also describes the tools that have been used for the research so far and it includes the research has been designed regarding SNA and the measurement of discourse and learning trails using metrics. Early results are also included where available. For reference purposes, web analytics (hereinafter, WA) data are also presented aiming to make clear the differences of these two terms. Finally, section V summarizes the current research results and discusses the future research in this scientific area.

II. FROM (BIG) DATA TO LEARNING ANALYTICS

Learning systems that facilitate education through e-learning have numerous advantages in comparison to the conventional learning means [1], [2]. In addition to all obvious advantages, such as provision of learning material and support with no geographical or time constraints, there are other less obvious but equally important advantages such as the activity tracking and monitoring; since all activities are monitored and stored in the database that is behind every learning system, there is a large amount of data, that can be exploited in order to lead us to useful outcomes regarding the learning process.

All digital traces that a learner leaves while using a learning system constitute data that can be further analyzed. Deibold [3] coined the term “Big Data” in order to illustrate the power of such data; Big Data refer to “the explosion in the quantity (and sometimes in the quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology”. Boyd supports the “Big Data” development on the Internet, which has “created unprecedented opportunities.
for people to produce and share data, interact with and remix data, aggregate and organize data...” [10], [14]. Beyond that, Norris et al. emphasize on the use of analytics, considering that analytics can be used to measure, compare and improve the performance of individuals and not just to enhance the experience but also to facilitate better outcomes to the activity [10], [15]. Dawson et. al relate the analysis with potential quality improvement and believe that learning-related data analysis might be used to improve the student learning experience, which would not only require a quantitative analysis but also a qualitative interpretation of findings, if not a qualitative analysis as well [16].

A. Web Analytics

Most of existing learning environments are monitored by Web Analytics (WA) methods. In order WA not to be confused with LA, the official definition of WA, it is vital to be included. According to the Digital Analytics Association, the official WA definition is “the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage” [4].

WA methods focus on internet traffic that a learning system provokes and track number of visits, geographical origin of visitors, page views, visit duration etc. WA are applied to every web page since many years.

B. Beyond Web Analytics: Learning Analytics

However, applying WA monitoring, appears as facing a learning system just like another website. How convenient does it sound? LA come to investigate further exploitation of all traffic and activity data that are produced behind a learning system from learning perspective.

Initially, analytics were mostly related to Learning Management Systems (LMSs) and appeared to be more relative with academic analytics that refer to the application of business intelligence in education and focus on institutional or regional level [7], [8], [9], [10]. But moving web and academic analytics one step beyond, Dawson et al. [8] considered that the analysis of data gathered through the learning activity could be used in order to improve the activity and the student learning experience [10]. As a still young concept, the term Learning Analytics (LA) was initiated and officially defined during the 1st International Conference on Learning Analytics and Knowledge of 2011 as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” [5], [6]. In contrast to WA, LA do not focus on web traffic produced by a learning environment but aim to exploit all digital traces that a learner leaves while he navigates in such an environment, contributing thus, to the improvement of the teaching procedure, curriculum shaping as well as learning efficiency and experience.

III. TYPES OF LEARNING ANALYSIS

According to Fournier et al. [10], learning-related data collection and analysis would not only be related to the increase of the effectiveness of learning, but also with possible changes in the learning process. Downes [16] argues that there are different tools for measuring learning engagement, and most of them are quantificational, obvious and measure trivial variables such as page access. He also believes that these tools are suitable for a basic level of assessment and highlights the necessity to future analytics systems that analyze even the learners’ contributions quality. Other researchers such as Kop and Parry, advocate that analytics could be used for providing learners and instructors with information that could be used for both learning improvement and provision of recommendations based on earlier learning activity [17], [18]. Fournier et al. advocates that this kind of data “would entail that it is not just analyzed with steps being taken by people to improve the performance in formal education but rather that technological means are being used to link data and use it to improve learning” [10].

According to Siemens and Long [5] proposed cycle, LA are reflected by five axes as following:

- **Course level:** social network analysis, discourse analysis, learning trails.
- **Educational data mining:** pattern recognition and predictive modeling.
- **Intelligent curriculum:** development of semantically defined curriculum resources
- **Adaptive content:** provision of adaptive content using recommendation procedures, based on learner behaviour
- **Adaptive learning:** social interaction and learner support as an adaptive learner process.

This paper, as already aforementioned, focuses on analytics on learning networks on the context of Personal Learning Environments (PLEs) through the H2L Open Educational Platform. More specifically, focuses on the first axis which refers to course level and divides the research into SNA and learning trails and discourse measurement using metrics.

SNA acts as a quantitative method of LA. Using network graphs, it aims to reveal the central nodes of the connections’ grid and thus, the people on the network who executed a significant role of connecting to the otherwise un-connected ones [10]. It consists of two basic building blocks: Nodes (or vertices or agents) and edges (or connections). Nodes usually represent physical members of a network, such as learners and instructors in case of an educational social network, but they can also represent other social structures such as teams, organizations, institutions, states or even countries [19]. Edges represent the connection between two nodes and consist of lines that point from one node to another. Edges may be directed or undirected; directed edges, represented as arrows, have an origin and a destination and may represent...
“who is following who”, ”who sends an email to who” etc. [19]. Connections like “who is friend with who” have a bidirectional nature and thus they don’t need to have any direction. In addition, quantitative network metrics have been set by social researchers in order to slice up the social community and create a basis on which to determine the relative position of individuals and clusters within a network, to track changes over time or to compare networks. Some these metrics are: Degree Centrality, Betweenness Centrality, Closeness Centrality and Eigenvector Centrality. At this very early stage of our research, only Degree Centrality that counts the total number of connections linked to a vertex [11], [19] is discussed.

On the other hand, metrics can quantize in more detail the activity of the learners (or members in case of an educational social network). Who does produce more content? What content is produced per learner? Who comments on every type of content and how much? Who feeds the most the public discussion of the course? Who launches his/her own discussion topic? This a sample of questions that can be interpreted into metrics and contribute to the quantification of the learning activity using data behind the PLE.

IV. ANALYZING AN OPEN PERSONAL LEARNING ENVIRONMENT

A. The Environment: HOU2LEARN

The increasing correlation between broadband penetration and e-learning across the European Union [20] indicates the importance of e-Learning and consequently the tools associated with it. While the most widely used tools are Learning Management Systems (LMS), those focus on course delivery without rich opportunities for interaction, content-creation and collaboration. The emergence of Web 2.0 influenced the e-Learning tools, leading to the development and continuous growth of PLEs. Such environments tend to adapt to the Web 2.0 attitude [21] and allow users to create, maintain and redistribute their own content. Also they incorporate strong social networking characteristics as well as a loosely structured collection of various widely used Web 2.0 tools such as wikis, blogs, fora, glossaries, RSS, multimedia sharing and real time conference [22], [23].

The Hellenic Open University (HOU) has set up an informal PLE in order to support the e-Learning process that takes place through the official established channels of traditional LMS environments and conferencing tools. This PLE, named HOU2LEARN (H2L), initially launched in September 2010 as part of an ongoing research within the Software Quality Research Group (http://quality.eap.gr) related to the correlation between Learning 2.0 challenges, social networking activities and informal learning. The environment has been built on top of the Elgg framework (http://www.elgg.org) and provides useful tools, the most important of which are blogs, social bookmarks, personal pages, file sharing, polls and status updates. Users have the ability to customize their views, disable features and interact with other users. Also, H2L supports the creation of Groups with private or public access in order to allow uses with common interests to meet, collaborate and share content.

One of the most important features of H2L is the ability of users to create relations between each other, by introducing the concept of “Follower”. Any user can “Follow” any other, so he receives immediate and through various channels, notifications about content added or manipulated by the user being followed. This instant notification and the collective delivery of information out of a noisy and crowded environment such as that of an open social networking educational platform, allows a user to create a network of interest with focus on his own personal needs, abilities and requirements.

An important checkpoint in the evolution of H2L occurred in September 2011. At this point the professors responsible for the postgraduate course “Special Issues on Software Engineering” (PLH42) of HOU, promoted H2L as the main environment of use for the objectives of this course. Students were encouraged to exploit the full potential of the environment and actively collaborate with each other, professors and other users. Also, a special group named PLH42 has been established in order to allow students to have a common reference point for content strongly related to this specific course. Hence, students and instructors of the course PLH42 were invited to make a registration to H2L if not already, and then to join the Group “PLH42”. Being member of this group, users can follow other group members so as to be kept posted about their activity. They can also launch a new Group Discussion topic, or upload a blogpost on the Group Blog, a file on the Group Files page, a bookmark on the Group Bookmark page and all these items can be commented by other group members. All students were also encouraged to express their questions through this environment and to transfer all their communication through the communication and collaboration channels than H2L provides.

B. Research and tools

As aforementioned, during the academic year 2011-2012, the course PLH42 ran completely through H2L instead of other well-established learning environments such as Moodle. Students and instructors were encouraged to register as members to H2L and then to follow the Members Group dedicated to the course PLH42 named “PLH42”. Behind that, the HOU learning analytics research team had set up direct access to the H2L Database using Navicat Premium and then Toad for MySQL 6.0.1. Database queries that extract all PLH42 members (nodes) along with the connections (edges) lead to tables that had to be in format compatible with the SNA software.

For SNA, a notable number of tools (such as Gephi, Sentinel, NodeXL, GraphViz, Ucinet, Touchgraph, Sentinel, NodeXL, GraphViz, Ucinet, Touchgraph,
GraphInsight) were tested and based on their capabilities and user friendliness, Gephi (v0.8 beta) and NodeXL (1.0.1.209) were finally used.

1) Social Network Analysis

Since H2L in an educational social network, social connections have been developed among the members (learners and instructors). This was initially depicted with a SNA diagram that illustrates the connections of members that related to the course “PLH42” that is the case study course. Every PLH42 member is represented with a node and the member username is also displayed. Directed edges (arrows) have been used and the arrow from member A to member B means that A follows B. The node size is proportional to the Degree Centrality of the member that corresponds to the node. Sole nodes (unconnected) are also represented using the smallest but visible node size.

Fig. 1 presents the connections among all 80 H2L PLH42 members in the middle of December 2011. Low connectivity is clear; only 18 members have at least one connection and the most central member is the instructor with a total number of connections, 22. The rest 62 members have no connections. Maximum geodesic distance, or in other words, the diameter of the graph (i.e. the length of the shortest path between two members) is 4. The ‘cloud’ of nodes at the low left side of the figure includes all unconnected, isolated nodes. As a part of a wider research, some assignments were given to all members aiming to encourage them to connect with other ones. In three months, the situation has been changed. Fig. 2 presents the increased connectivity; the number of edges has become 73, which means 232% increase. Maximum geodesic distance remains 4. It is apparent that during the first weeks of the academic year, Group members didn’t connect each other and it was necessary to provide encouragement so as to increase the connections and the activity in general. Both figures have been produced using Gephi (v0.8 beta). On every node, the username of the respective PLH42 member is displayed as a label. The node and label size and colour are proportionate to the member degree centrality which is defined as the number of links incident upon a node (i.e., the number of edges that a node has) [11]. The ‘cloud’ of nodes at the upper right side of the figure includes all unconnected, isolated nodes.

One other screenshot was taken in the middle of May 2012. The representation was achieved using NodeXL (1.0.1.209) due to stability issues that occurred using Gephi. In spite the fact that these two tools don’t provide the same possibilities in order to make the graphs to have the same format, it was tried to apply all common settings. Nodes and labels’ size and color are also proportionate to the degree centrality. Fig. 3 presents the connections among the PLH42 members in the middle of May 2012. Unconnected nodes are located in the low left side of the graph.

Two months after the situation illustrated in Fig. 2, the connectivity has increased even more. The connected nodes
are now 54 with a total number of edges, reaching 404. Maximum geodesic distance is now 5. It is also vital to mention that, there is a number of algorithms that arrange the nodes in the graph pane, and aim to make the graph more readable. The most common is called Fruchterman-Reingold algorithm that is a force-directed algorithm, which threads edges like springs that move nodes closer or further from each other in an attempt to find equilibrium that minimizes the “energy” of [19], [24]. The algorithm results are dependent on two parameters: the strength of repulse force between nodes (higher numbers make nodes push away from each other more), and the number of iterations (that determines the number of times the Fruchterman-Reingold algorithm layout will be performed). The values for Fig. 3 are: repulsion = 8 , iterations = 100, since the default values repulsion = 3 and iterations = 10 gave a not so readable result. Same conditions were applied in both Fig. 1 and Fig. 2.

Furthermore, precise SNA in a graph can reveal groups that act in competition or competitively or that have higher or powerful action. SNA software incorporates tools that identify such groups or clusters, i.e. pockets of densely connected nodes that are only sparsely connected to other pockets [19], [25]. Using the Wakita-Tsurumi algorithm that looks for groups of densely clustered nodes that are only loosely connected to nodes in another cluster [19], [26], it appears that the graph includes 7 groups, as illustrated in Fig.4.

2) Metrics
In order to enable learning trails and discoursing and content trafficking to be measurable, all members’ activity needed to be operationalized as metrics.
These metrics aim to quantitatively analyze data from activities that take place during the course lifecycle [11], [12]. Hence, data available can be potentially used to improve the helpfulness and effectiveness of the learning procedure within H2L.

This paper proposes a set of metrics that aim to measure the traffic and activity within the pages of the course “PLH42”. These metrics are adopted to be calculable from data available while the course PLH42 runs through H2L and they are a first step towards the setting of a stable baseline implementation with which learning activity related metrics can be used for effective and efficient tracking of the quality of learning outcomes and the improvement of the learning procedure.

According the aforementioned possibilities that H2L provides to the members of course PLH42, the metrics intend to quantize the content that is uploaded, commented and created. In this case the term “content” includes group discussion topics, comments, blogposts, wire posts, files and bookmarks. The proposed metrics follow:

- Number of:
  - topics that each user has uploaded on Group Discussion.
  - comments on topics of Group Discussion
  - new blogposts in Group Blog
  - comments on blogposts in Group Blog
  - comments on wireposts of other Group Members
  - uploads of new files on Group Files page
  - comments on files uploaded by other Group members
  - new bookmarks in Group
  - comments on bookmarks uploaded by other Group members

These metrics are based on the activity that HOU2LEARN permits user to achieve and aim to quantitize the activity that takes place within the Group of course PLH42. The next steps are to translate these metrics into SQL queries and to calculate the values per member.

3) Web Analytics (as a reference)
Aiming to make the difference between LA and WA clearer, and for reference purposes, it is included below some WA data that HOU2LEARN collects in parallel with the LA research.

In HOU2LEARN, monitoring conducted through Piwik. Piwik is an open source WA system written by a team of international developers, and runs on a PHP/MySQL webserver. Piwik is a widely accepted system, since it is used by over 250,000 websites. Using a modern user interface, and capabilities such as real time reports, extended privacy features, it is released under GNU/GPL.

Fig. 5 presents a number of examples of the WA results that have been tracked in H2L during the period October 2011 to April 2012.

Figure 4. SNA Diagram of H2L members (Course PLH42) with groups using NodeXL. Screenshot taken 16/05/2012.
It is evident that WA provide an overall approach on web traffic that are generic and don’t focus on specific evolving activities. They don’t illustrate neither the interaction among the members of H2L, nor specific actions of each member regarding the content he/she manipulates (content creation, content exchange, content commenting etc.).

V. DISCUSSION AND FUTURE RESEARCH

This paper presents a part of an ongoing research regarding the application of LA in H2L, an open personal learning environment. After the presentation of the concept and definition of LA, the applied methods along with the progress of each of them so far were discussed. At the end of the academic year, useful outcomes are expected to be produced; how the connectivity is changed though the year, what is the relation of the most connected members with the values of the metrics etc.

At this phase, the research about LA that is conducted in H2L sticks on the quantitative paths and it is in our next steps to expand it in qualitative paths as well covering in a more complete way the nature of LA. One step towards this, is to consider learners’ grades; it is possible that learners with high grades may find the course too easy to be interesting for them and thus they have low activity in H2L. PLH42. Hence, institutional assessment including grades etc. will be co-calculated along with the metrics’ values.

Furthermore, as soon as the metrics are finalized and give values, it is planned to be combined with the SNA results so as i.e. in a SNA diagram, nodes’ size will be proportional to the value of a metric. The degree centrality of each member along with his/her metrics’ values will be compared. Do most centered members have high metrics’ values as well? Useful outcomes that compare the degree centrality and the metrics are expected to be produced.

REFERENCES


© IEEE 2012. This is the authors’-created version of the paper. It is posted here for your personal use, not for redistribution. The IEEE-published version is available at doi: 10.1109/PCI.2012.38