LeMo: a Learning Analytics Application Focussing on User Path Analysis and Interactive Visualization

Albrecht Fortenbacher[1] , Liane Beuster[1] , Margarita Elkina[2]
Leonard Kappe[1] , Agathe Merceron[3] , Andreas Pursian[2]
Sebastian Schwarzrock[3] , Boris Wenzlaff[1]

[1] HTW Berlin, University of Applied Sciences
Wilhelminenhofstraße 75A, 12459 Berlin, Germany
{liane.beuster,forte,kappe,boris.wenzlaff}@htw-berlin.de
[2] HWR Berlin, Berlin School of Economics and Law
Alt Friedrichsfelde 60, 10314 Berlin, Germany
{margarita.elkina,andreas.pursian}@hwr-berlin.de
[3] BHT Berlin, University of Applied Sciences
Amrumer Straße 10, 13353 Berlin, Germany
{merceron,sschwarzrock}@beuth-hochschule.de

Abstract **– LeMo is the prototype of an application for learning analytics, which collects data about learners' activities from different learning platforms. The article describes design principles of LeMo and their implications for efficient learning analytics. Focus is on the LeMo system architecture, user path analysis by algorithms of sequential pattern mining, and visualization of learners' activities.**

Keywords **– learning analytics; visual analytics; sequential pattern mining; user path analysis**

I. Introduction

The Horizon Report 2011 [1] identifies learning analytics as an emerging technology, which most likely will have a significant impact on higher education within the next years [2]. On the first LAK conference 2011 [3], learning analytics was defined as

> ... the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs ...

In 2011, the research project "LeMo: monitoring of the learning process on personalizing and non-personalizing learning platforms" started at three universities of applied sciences in Berlin [4]. It aims at developing an analytics and visualization tool, to facilitate learning analytics for teachers, E-learning providers and researchers. Basic requirements were connectivity to various platforms for online learning, powerful analysis and mining of data ob-

tained from the platforms, and an intuitive visualization of analytic results, providing insight into learning processes.

The following section describes the methodical approach of LeMo, which guided requirements analysis and refinement throughout the project. Then, basic design principles of the LeMo application are explained. The following sections focus on user path analysis and visualization. The last section concludes the findings of the LeMo project, and refers to future work.

II. Methodology

Requirements analysis in LeMo is based on a survey amongst stakeholders (project partners), representing the three target groups teachers, E-learning providers and researchers. This generic qualitative method was targeted at supporting teaching with a variety of didactic methods and different degrees of technology support, e.g. E-learning as a supplement to face-to-face teaching, or online learning as a basis for distance learning.

The result of the survey was a catalogue of about 80 questions and assumptions regarding students' learning behavior and the use of online media. Some of the questions were beyond the scope of an analytics tool which relies on learners' activities obtained from a learning (management) platform.

The catalogue of questions was segmented into six categories representing different areas of interest. The first two categories contain questions about "users and groups of users", as well as "learning objects and type of learning objects". Interactions with the learning environment are modeled through the categories "usage analysis", "user path analysis", "communication and collaboration", and "performance".

Based on the catalogue of questions, 27 indicators were derived, that provide data analysis to help answering the referred questions. The most important indicators were "activity / time" and "activity / learning object", both related to the category "usage analysis", as well as "frequent paths" and "activity graph", related to "user path analysis".

The survey showed a considerable interest in analyzing the sequence of user interactions. This interest can be contextualized and supported by the theoretical approach of connectivism. According to Siemens, learning is organized in a network – internally as a neural network and externally 'as a network of nodes, representing a specialization and aggregates as actions concerning the environment" [5]. In a first step, this network is visualized as an "activity graph" in LeMo.

Consequently, the four indicators mentioned above were the first indicators realized in LeMo. Whereas "usage analysis" is quite common in comparable tools, the indicators "frequent paths" and "activity graph", together with intuitive visualization, are unique characteristics of the LeMo tool.

III. LeMo Application

A. System Architecture

LeMo is designed as a 3-tier application, which facilitated distributed development among the three universities [6]. The first tier (data) contains a data model as well as functionality for data management, which includes connectivity to different learning platforms. Data analysis takes place in the second tier (data). Results of data analysis are passed to the third tier (presentation), where the data are visualized. Filtering data, e.g. selecting male or female students within a course, is presently done in the data tier, but eventually should be performed in the third tier, thus reducing the number of requests to the data tier.
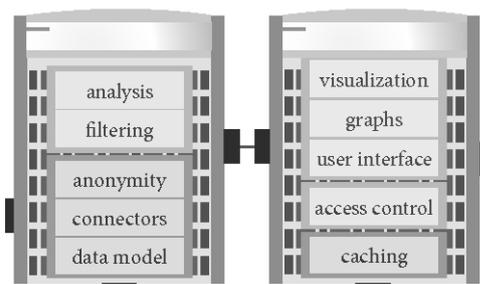

Figure 1. 2-server architecture

The LeMo application consists of a *data-management server*, which basically comprises the first and second tier, and a *application server*. From user interaction, the application server generates requests, which are passed via a web-service interface to the data-management server.

Both servers are realized via web apps, which could be run by a simple Tomcat server, but also, for performance and security reasons, be distributed within a network.
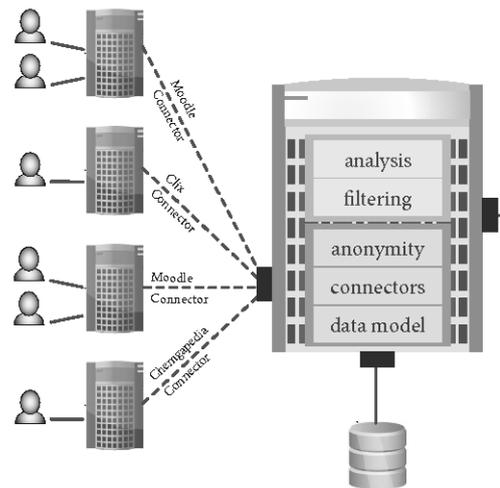

Figure 2. data-management server

A data-management server may connect to more than one LMS (learning management system). Presently, connectors for different versions of moodle, clix and Chemgapedia are implemented.
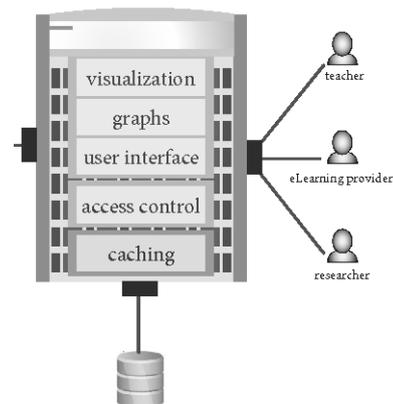

Figure 3. application server

The application server provides for visualization and user interaction. Using a JavaScript framework, this is done locally on the client, reducing server load and network traffic.

B. Data Management

In [7], three factors for the development of learning analytics are listed: online learning, big data, and political concerns. This list is reflected by the design principles of LeMo.

Data analyzed by the LeMo tool are basically user activities obtained from platforms for online learning. LeMo connects to various platforms, including the learning management systems (LMS) Moodle and Clix, as well as the online encyclopedia Chemgapedia. The LeMo connectors for Moodle and Clix are implemented to access the underlying databases directly. In case of Chemgapedia, which is a web application, data stem from the server log file.

During the ETL (Extraction – Translation – Load) phase, data are imported into a database with a unified data model. While LMS provide detailed information on content and on users, server log files just contain page visits. Here, session information is needed to identify anonymous users (students). The LeMo data model contains entities for learning objects, e.g. courses, resources, wikis, or tests, and associations between learning objects, e.g. resource X belongs to course A. There exist students and teachers related to courses. Degree programs can group courses, a department is responsible for different degree programs, and an institution consists of several departments. The data model also includes *user activities*, which basically are represented as tuples $[U, L, T]$ (user, learning object, time stamp). Additionally, an *action* is stored with each user activity, which may be one of *view*, *download*, *modification*, *creation*, *attempt*, or *submit*. And finally, platforms are represented in the data model. Given an element E, which was imported from platform P, a unique primary key for this element can be constructed from the pair $[E, P]$.

In the case of Chemgapedia, some data are not available, examples being department/faculty, or a student's enrollment in courses. This data has to be left empty, leading to sparse data, and eventually reducing the number of available analyses.

Big data denote huge sets of, generally unstructured, data, which typically originate from social networks, and which are processed using analytic methods ("social media analytics"). Recording activity data on platforms for online learning, over a long period of time, a huge amount of data is produced. For interactive analytics applications, it is critical to handle big data efficiently, which can be achieved via data compression, efficient data retrieval and efficient mining algorithms.

Reviewing usage statistics on the Chemgapedia web server, it became apparent that log data have to be pre-processed in order to exclude activities not corresponding to learners (e.g. traffic generated by web crawlers). To solve this problem, an optional functionality was added which excludes all log entries with one of the following characteristics: multiple accesses per second, frequent repetition of time intervals between accesses, and many accesses to the same page. The use of this filter reduced the number of log entries by about 47%, while excluding just 5% of all users.

With the actual LeMo prototype, data are stored in a relational data base. While this is sufficiently efficient for the (relatively small) amount of data considered so far, non-standard ways of storing data, e.g. in a NoSQL data base, will be investigated.

C. Data Privacy

One political concern of learning analytics is data privacy. Traditionally, personalized interactions and user modeling have significant implications on data privacy. Personal information about a user is collected and analyzed, which might not be in the interest of the user. Recently, collection of user data in social networks, and possible violation of data-privacy legislation, have been published frequently. As a result, even more strict data-privacy regulations at universities are established, which in turn limits the use of learning analytics.

The LeMo approach towards data privacy aims at achieving a high level of anonymity. The activity sequence, i.e. "learning path" of a particular student may be analyzed, but personal data, which could serve to identify the student, must not be exposed. This is done by omitting all personal data in the ETL process, the only exception being gender information which is valuable for learning analytics. Furthermore, small data samples must not be used for a specific analysis, if this leads to the identification of a particular student (k-anonymity [8]). An example is a course with just one female (male) participant; this student could be identified easily using a gender filter.

For most of the analyses, courses must be identified by their name. Also, the title of learning objects like learning material (e.g. powerpoint presentations), the name of threads in a course-related forum, or results of a test provided in the course, are essential for efficient analysis, but must not imply the identification learners (students).

To implement a role *teacher*, courses taught by a given teacher must be identified. This cannot be done using some unique attribute, like name, personal ID or login name (user ID). All these attributes are omitted during ETL process. Instead, a mapping $\phi : Teachers \rightarrow Courses$ can be realized by a pseudonym (hash value) derived from a unique attribute during ETL. If the LeMo application uses the same authentication scheme (e.g. authentication via LDAP) as the connected platform, the pseudonym is derived from the login name. When a teacher logs in, he sees the list of "his" courses.

IV. User-Path Analysis

Regarding sequences of learning activities, we can construct user paths reflecting navigation of students (within a course). A user path is a sequence of learning objects. For each user path, there exists a set of users who access the learning objects in the order given by that user path. Let

$$P = (L_1 ... L_n) \tag{1}$$

be a user path, and

$$\mathcal{A}(U) = (A_1...A_k) \qquad (2)$$

the sequence al all user activities of user U, ordered by time. Now we can define $\mathcal{U}_f(P)$, the set of all users U supporting user path P, by the following conditions

$$\forall\, 1 \leq i \leq n\, \exists\, 1 \leq j_i \leq k : A_{j_i} = [U, L_i, T_{j_i}] \qquad (3)$$

$$j_1 < ... < j_n \qquad (4)$$

In the context of a course, or of several courses, with a defined number k of students, the *support* of a user path is the number of students following this path, relative to the total number of students:

$$\mathcal{S}_f(P) = \mathcal{U}_f(P)/k \qquad (5)$$

Given some *minimal support S*, we can define the set of *frequent paths*:

$$\mathcal{FP}(S) = \{P | \mathcal{S}_f(P) \geq S\} \qquad (6)$$

Choosing an appropriate minimal support, a teacher can investigate if the majority of students access the learning objects in the intended order.

Direct paths, which are followed by students who don't have any extra user activity between accessing the first and last object, might provide further insight into the learning process. The set $\mathcal{U}_d(P)$ consists of all users U supporting directly a user path P. U supports P directly, if a position p within $\mathcal{A}(U)$ can be found with

$$\forall\, 1 \leq i \leq n : A_{i+p-1} = [U, L_i, T_{i+p-1}] \qquad (7)$$

i.e. if the user path P can be embedded as a subsequence into the sequence of activities of user U.

Similar to the definition of frequent paths, direct paths can be defined as a function of minimal support:

$$\mathcal{S}_d(P) = \mathcal{U}_d/k \qquad (8)$$

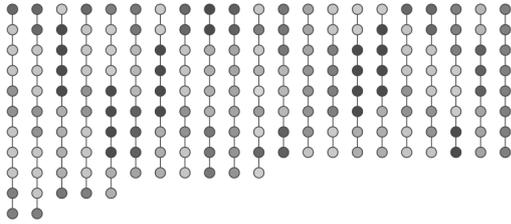$$\mathcal{DP}(S) = \{P | \mathcal{S}_d(P) \geq S\} \qquad (9)$$

Figure 4. frequent paths

In the LeMo application, the BIDE algorithm [9] is used to identify frequent paths. If the support parameter is lowered, the number of frequent paths increases, as well as the response time of the BIDE algorithm.

The drastically increased processing time of BIDE for low support values poses a serious problem to the usability of the LeMo application. As a solution, a "default value" for minimal support is determined, depending on parameters "number of user paths", "average length of user paths", and "number of learning objects" within a given course. Since BIDE uses an a-priori approach to determine reoccurring sub-paths, the processing time increases dramatically when the number of frequent sequences is large [10].

A typical users' behavior is to try out low support values (which might cause BIDE to run "for ever"), and then increase the values until a result is achieved within a reasonable response time. This might start a couple of BIDE algorithms running as parallel threads, and eventually bring down the system. To avoid this, the algorithm must be terminated when a user requires the next analysis. Thus, in a multi-user environment, the number of running BIDE algorithms is limited to the number of active users. Furthermore, after a pre-defined lapse of time, the algorithm is terminated (default value: 5 minutes), which does not leave hung threads in the system.

Not surprisingly, computing direct paths is a much easier task, allowing for interactive learning analytics. The Fournier-Viger algorithm [11], which is used for direct paths, returns results within a very short time, even for low support values like 10%, which significantly increases the number of direct paths.

V. VISUALIZATION

The design of the user interface in LeMo follows the explorative metaphor "overview first, filter and details on demand", which assists analysis of learning data in an explorative way [12]. Especially when displaying large amounts of data, this approach supports a highly intuitive reception of the results of analyses.

Visualization was designed according to the following guidelines:

- **support pre-attentive perception**: through the use of visual attributes, e.g. shape, size, color and position
- **provide details on demand**: detailed information on a specific learning object are available (context sensitive), without overloading the initial visualization
- **facilitate exploration**: the interface provides the possibility to customize the visualization, to best fit the personal expectations, through features like translation, rotation, filtering, and zooming

The results of an analysis "activity graph" are visualized by a navigation network. This analysis can be used to obtain information on the degree of cross-linking between learning objects. The activity graph is visualized by a graph, with learning objects as nodes and navigation steps as edges. Nodes are color-coded to support the visual
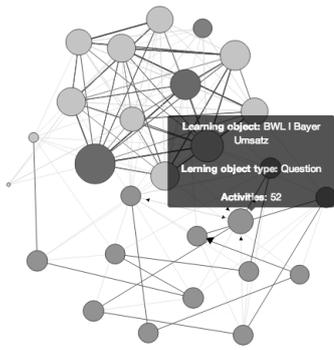
Figure 5. activity graph

Figure 7. alternative form of activity graph

perception of content type transitions. Learning objects are adjusted in size to encode the absolute number of user requests. Edges are weighted and color-coded to encode the amount of navigational steps.

Detailed information on specific elements of the graph can be made visible by interacting with the visualization, including tool tips with information on learning objects, and rearranging the graph, focussing on nodes of interest.
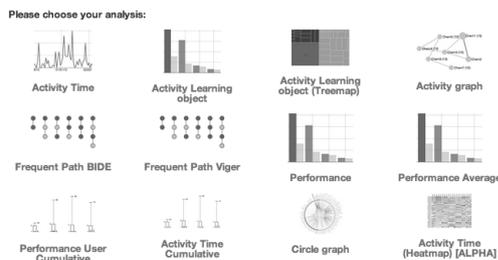

Figure 6. different visualizations

For the visualization of "frequent paths" (results of the BIDE algorithm), the following premises are important:

- A path is visualized by a sequence of nodes (learning objects) and edges (navigation steps). Learning objects are color-coded to allow a visually easy correlation of navigational patterns across different paths.
- Detailed information on specific elements of the path is visible through user interaction.

LeMo provides a variety of different visualizations, which are suitable for analysis of specific learning situations and processes. For example, For example, the performance of students based on self-tests can be visualized. Alternatively, students' performance could be used as a filter for other analyses, exploring correlations between performance and navigation [13].

In some cases, LeMo provides even different "visual styles" for the same analysis, an example being "activity graph". The first version features a volatile visualization
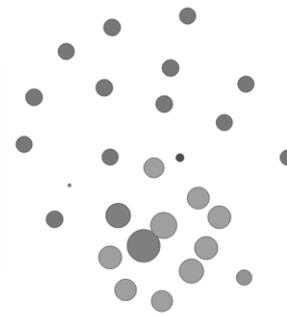
of the navigation graph, which facilitates user interaction. The second version is a more static visualization, which, as an overview, just displays learning objects, where the size of the circles corresponds to the number of related user activities. Edges are omitted - selected edges appear, when a specific learning object is selected. The two visualizations support different ways of perceiving information, allowing users to develop their personal style of doing learning analytics.

VI. CONCLUSION AND FUTURE WORK

The actual LeMo prototype proved to be a useful tool for teachers, giving hints on the learning behaviour of their students, and helping with the evaluation of their didactic concept. The combination of user path analysis combined with a powerful interactive visualization distinguishes LeMo from other tools for learning analytics.

Focus of future work will be on improving the LeMo application with respect to efficiency and usability, as well as on enhancing data analytics and visual analytics. System improvements will include a standardized ETL interface, connectivity with further platforms, a (non-standard) data model for efficient data retrieval, and additional pattern-mining and data-mining algorithms.

Functionality will be enhanced by realizing more indicators derived from the stakeholders' original list of requirements (questions). Of special interest are indicators referring to "group of users", "students' performance", and "communication and collaboration".

The actual prototype realizes the role "teacher", where a teacher can analyze his/her own courses. In the future, the roles "E-learning provider" and "researcher" should be defined and implemented.

For all improvements on the functionality and usability of the LeMo tool, an evaluation study is essential and will be conducted in the immediate future.

REFERENCES

[1] L. Johnson, R. Smith, H. Willis, A. Levine, and K. Haywood, "The 2011 horizon report," The New Media Consortium, Austin, http://www.nmc.org/pdf/2011-Horizon-Report.pdf, 2011.

[2] A. Fortenbacher, "Learning analytics for higher education - perspectives and challenges," in *Pratsi: Scientific, science and technology collected articles*, issue 1(40) ed. Odessa National Polytechnic University, 2013, pp. 184–187.

[3] *1st International Conference on Learning Analytics and Knowledge.* ACM Digital Library, 2011.

[4] A. Beuster, M. Elkina, A. Fortenbacher, L. Kappe, A. Merceron, A. Pursian, S. Schwarzrock, and B. Wenzlaff, "Learning analytics und visualisierung mit dem lemo-tool," in *Proceedings of Interaktive Vielfalt - DeLFI, Bremen*, 2013.

[5] G. Siemens, "Connectivism: Learning theory of pasttime of the selfamused," http://www.elearnspace.org/Articles/connectivism_self-amused.html, 2006.

[6] L. Beuster, M. Elkina, A. Fortenbacher, L. Kappe, A. Merceron, A. Pursian, S. Schwarzrock, and B. Wenzlaff, "Prototyp einer plattformunabhängigen learning analytics applikation – fokussiert auf nutzungsanalyse und pfadanalyse," in *E-Learning Symposium 2012 Aktuelle Anwendungen, innovative Prozesse und neueste Ergebnisse aus der E-Learning-Praxis*, U. Lucke, Ed. Universitätsverlag Potsdam, 2012, pp. 69–72.

[7] R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," http://kmi.open.ac.uk/publications/techreport/kmi-20-01, Knowledge Media Institute, The Open University, UK, Tech. Rep. KMI-12-01, 2012.

[8] V. Ciriani, S. D. C. di Vimercati, S. Foresti, and P. Samarati, "K-anonymity," in *Secure Data Management in Decentralized Systems*, ser. Advances in Information Security, T. Yu and S. Jajodia, Eds. Springer, Berlin Heidelberg, 2007, vol. 33, pp. 323–353.

[9] J. Wang and J. Han, "Efficient mining of frequent closed sequences," in *Proceedings of the 20th International Conference on Data Engineering, Boston, MA, USA*, 2004.

[10] S. Schwarzrock, "Analysis of learner navigation on web-based platforms using algorithms for sequential pattern mining," in *Pratsi: Scientific, science and technology collected articles*, issue 1(40) ed. Odessa National Polytechnic University, 2013, pp. 18–21.

[11] P. Fournier-Viger, R. Nkambou, and E. M. Nguifo, "A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems," in *Proceedings of MICAI*, ser. LNAI 5317, A. Gelbukh and E. Morales, Eds. Springer-Verlag Berlin Heidelberg, 2008, pp. 765–778.

[12] A. Pursian, "An interactive approach to information visualization in the context of learning analytics," in *Pratsi: Scientific, science and technology collected articles*. Odessa National Polytechnic University, 2013.

[13] A. Merceron, L. Beuster, M. Elkina, A. Fortenbacher, L. Kappe, A. Pursian, S. Schwarzrock, and B. Wenzlaff, "Visual exploration of interactions and performance with lemo," in *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, USA*, 2013.