Data Warehousing for Lifelong Learning Analytics

M. C. Pham and R. Klamma

Abstract— **Lifelong learning aims at improving our personal knowledge, skills, and competences during our professional life. Traces or historical data of our lifelong learning processes are stored in a number of distributed portfolios. Lifelong learning analytics (LLA) involves harnessing these portfolios. However, data management of heterogeneous learning portfolios can become cumbersome and therefore limiting LLA, if done manually. Data warehousing is a data management technology ideally suited for automatic extraction, transformation and loading (ETL) of LLA data from heterogeneous portfolios making it perfectly suitable to be employed as a backend for visual LLA. A case study on teacher networks in Europe is demonstrating the feasibility of our data warehouse approach. We employed the data warehouse to analyze the interaction between teachers using social network analysis (SNA) for competence management and assessment of teachers' performance via measuring social capital of teachers in the networks over time. Since all core ETL processes are running automatically, only little maintenance effort is needed to run the LLA processes.**

***Index Terms*—Lifelong learning, learning analytics, data warehouse, social learning.**

I. Introduction

Learning analytics is recently attracting significant attention from research communities in learning technology, education, information science and computer science. The *First International Conference on Learning Analytics & Knowledge* (LAK 2011) defined learning analytics as the measurement, collection, analysis and reporting of data about learners and their context, for purposes of understanding and optimizing learning and the environments in which it occurs. Previously, learning analytics mainly focuses on supporting learning adminostrators, funders, departments, education authorities in decision-making. Currently, there is a shift from an institutional perspective to the concerns of learners and teachers, which are more personalized.

Manuscript received April 08, 2013. This work was supported in part by the DFG-funded excellence cluster UMIC, EU FP7 ICT Tellnet project and EU FP7 ICT Integrated Project Learning Layers.

M. C. Pham and R. Klamma are with Information Systems and Databases, RWTH Aachen, Ahorn Str. 55, D52056, Aachen, Germany (email: {pham, klamma}@ dbis.rwth-aachen.de).

Lifelong learning (LLL) is all learning actvitity that is undertaken throughout learners' life, with the aim of improving personal knowledge, skills, and competences [5]. The main focus of LLA is on learners and learning communities, where analysis helps learners to adapt to the learning environment and improve outcomes. Unlike formal learning where learning activities take place in certain places (schools, classes) and in certain time periods, lifelong learning is informal, where learning activities take place from the daily interactions with others and with the environment in different contexts, e.g. home schooling, continuing education, professional development and personal learning environments. Long time period of lifelong learning activities and the diversity of learners' interactions raise a challenge for learning analytics in both data management (e.g. data quality management, data uncertainty) and analytical applications for competence management, learner progression analysis and social interactions.

The central of any LLA system is a data warehouse that integrates information from various learning sources for use in analytical applications. Although designing and implementing a data warehose depends on the applications, there are several key principles in design and implementation of a LLA data warehouse, which we want to discus in this paper. The paper is organized as follows. In section II, we discus about the data warehouse model for LLA. In section III, a case study on the teacher network in Europe is presented. The paper finishes with conclusions and outlook.

II. Data Warehouse for Lifelong Learning Analytics

The data warehouse model for LLL analytics is given in Fig. 1. First, data about learning activities from learning systems such as Content Management Systems (CMSs), Learning Management Systems (LMSs), Student Information Systems (SISs), Social Networking Sites (SNSs) and other sources, e.g. Personal Learning Environments (PLEs) needs to be integrated. Data from those sources can be available in different formats, e.g. relational data tables, flat files or XML. Those different sources might contain data of varying quality, or use inconsistent representations, codes and formats. Data integration process needs to check the quality of data, perform data cleaning, transform data sources and integrate them.

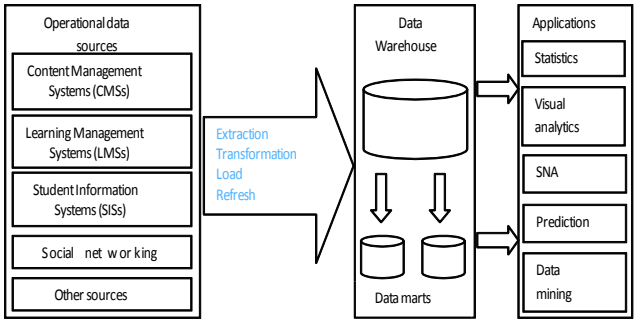

Fig. 1. Data warehouse for LLL analytics.

Since data warehouses are targeted for decision support, historical, summarized and consolidate data is more important than individual records in operational systems. For example, student performance in each subject over semesters, or social interaction of learners over time might be more important than the current data. Since a data warehouse contains consilidated data from several databases over long time periods, it should be seperated from operational learning systems to ease the data management and to allow complex queries and analysis to take place without disrupting or slowing learning systems.

After data integration, the resulting data is processed ETL standard steps. For automatic data proccessing, back end tools and utilities are implemented for data extraction, cleaning, transformation, loading and refreshing. Those tools connect to the data warehouse via standard interfaces, such as JDBC, Oracle Open Connect, etc. After extracting, cleaning and transforming, large amount of data is loaded into the warehouse. Then analytical applications such as visual analytics, SNA, recommendations or other data mining applications can be built on top of the data warehouse. Refreshing updates the warehouse (data and derived measures) based on the updates in operational databases.

To store information in a data warehouse, a data schema needs to be designed based on a *multidimensional* data model which views data in the form of a data cube and is commonly modeled within a relational databse management system (RDBMS). In this model, there is a set of measures that are derived from operational data. Each measure is associated with a set of dimensions, which give the context for the measure. For example, the dimensions associated with the average grade of students can be subject, semester and programm. Normally, the data schema is modeled using a *star schema* where a fact table (to store measures) is linked to dimension tables [6]. In the following, we present a case study on teacher network in Europe to give an example of the design of a data warehouse for a particular LLA application.

III. A CASE STUDY WITH TEACHER NETWORK IN EUROPE

eTwinning[1] is the community for schools in Europe, which creates a professional development network for European school teachers. It aims to promote European teachers' collaboration through the use of Information and Communication Technologies (ICT). The Teachers' Lifelong Learning Network (Tellnet) project aims to provide a learning analytics framework to analyze eTwinning community. The framework targets several stakeholder groups, including:

- **Decision-makers** pursue a better future for European school education and for European teachers' professional development and lifelong learning, which influence school education. They may be interested in different aspects of eTwinning community, especially how it develops over time, interactions between schools, regions and countries.
- **The Central Support Service** (CSS) i.e. learning network management staff. They are interested in questions like "What is the relationship between eTwinning and professional development? How do they influence each other and how do they support each other? "
- **The National Support Service** (NSS) promotes eTwinning by providing training and support, organizing meetings and national competitions, and managing public relationships based in each country. They may be more interested in the interaction of eTwinners in their country and, the interaction between their countries with other countries.
- **School management staff** is interested in teachers' activities in eTwinning, especially regarding the development of teachers' ICT skills and project cooperation. The interaction between teachers in the school with teachers in the same other schools is of special interest.
- **Teachers** (eTwinners) from pre-school, primary, secondary and upper schools can all participate in eTwinning to exchange and collaborate, as well as to learn new ICT skills, communication skills, teaching skills, and interdisciplinary working skills. They are interested in their own performance in eTwinning as well as their communities.

To answer the questions raised by different stakeholder groups, we employ data warehouse technology and social network analysis to analyze the interaction between teachers.

A. *eTwinning Data Warehouse Model*

The eTwinning data warehouse model is given in Fig. 2. eTwinning data is provided by European SchoolNet in CSV files. The data set consists of all teachers registered in eTwinning portal, their intitutions (schools), regions and countries. The interactions between teachers are reflected via several communitication mechanisms provided by the systems, including project collaboration, messages (posting messages on the profile wall of each other), contact lists and blogs. Other information includes prizes and awards of teachers and projects. Data also records time information of teachers' activities.

Before performing standard ETL steps, data is checked for consistency, e.g. missing values, wrong references, etc. When we have identified errors, appropriate data cleaning methods can be applied to correct them. For example, depending on the role of the missing values, they can be ignored, be filled by default values, or be deleted.

Based on the cleaned data, we extract several social networks at different levels as well as other objects such as teacher, institution, project, blog and message. Those networks include: *teacher networks* where nodes are teachers and connections between denote the interactions, e.g. project collaboration, contact list, commenting on blog posts, posting on the profile wall of each other; institution networks, region

[1] www.etwinning.net

network and contry networks where nodes are institutions, regions and countries and connections between them are aggregated from teacher networks based on teacher-institution relationship and the information about location of institutions.

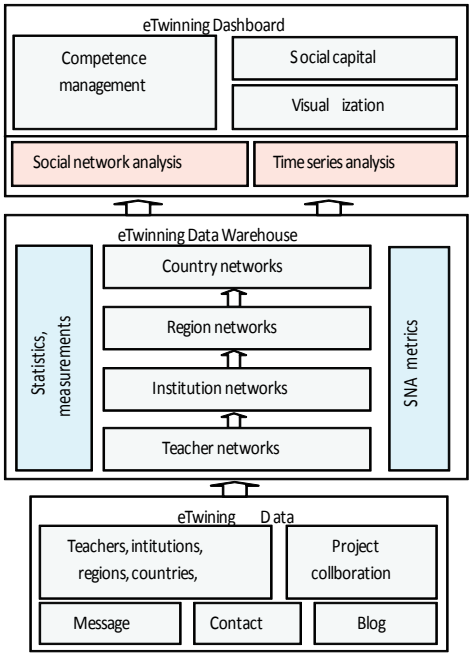

Fig. 2. eTwinning data warehouse model.

We develop a data model to store the above networks together with time information and network measures over time. For each time point, we store a snapshot of the each network and for each snapshot we computed a set of network parameters, including numbers of nodes and edges, network and node clustering coefficient and betweenness, node degree [1]. Together with other statistics, e.g. number of quality labels, prizes and projects, subjects, etc., this data allows us to analyze the development of the networks, assessing social capital of teachers and their professionnal development via the interactions with each other in eTwinning. eTwinning dashboard can be customized to serve different stakeholder groups according to their information needs. In the follows, we focus on three applications of this data warehouse: visualization of the networks provides an overview of teachers' interactions; comptetent management helps teachers to monitor their own performance in eTwinning network; and social capital assessment identifies the form of social capital in eTwinning, which can be used to make personalized recommendations to teachers.

B. Network Visualizations

Visualizations are useful for understanding teachers' interactions at different levels. Networks can be visualized directly from eTwining data warehouse. By combining the ranking of nodes in the networks by SNA measures, the visualizations can show the structure of the networks (e.g. communities) as well as the important of the nodes. The visualizations are especially useful for decision-makers, central support service, national support service and school management staffs, who are interested in an overview of teachers' interaction at Europe, country and school levels.

Fig. 3. Institution project collaboration network (geo layout).

For example, Fig. 3 shows the visualization of the institution collaboration network where nodes are institutions and are place on the map according to their coordination, and the color of nodes and edges denote the community that they belong to. One can see that eTwinning members come from schools which cover almost all countries in Europe. The collaboration is also very international, where members collaborate across geographical boundaries, not only within local members. eTwinning dashboard can visualize teacher networks at all level (i.e. teacher, institution, region and country levels) to provide different fine-grained views.

C. Competence Management

The term competence is defined as "The knowledge, skills, traits, attitudes, self-concepts, values, or motives directly related to job performance or important life outcomes and shown to differentiate between superior and average performers" [7]. eTwinning dashboard provides tools for learner self-monitoring and reflection. In competence management, competence assessment plays a central role. We consider three types of teacher competences and define indicators to assess them:

- *Professional competence*: involves all abilities or skills of teachers, which are necessary for performing professional tasks in eTwinning. For example, sufficient knowledge of a subject and a language are essential for organizing a successful consortium that focuses on the subject. We define *project performance* indicator to measure how a teacher performs in project by the number of projects she participated in and the number of quality labels/awards sh received.
- *Social competence*: considers the interactions of a teacher with others and with communities via different communication mechanisms, e.g. project collboration, messaging, blog posting and commenting. SNA measures on the networks formed by teachers' interactions can be used as indicators to assess social competences.
- *Meta-competence*: considers the self-monitoring ability of teachers, which describes teachers' ability to take advantage of monitoring their activities in the eTwinning network. It depends on various features, such as basic knowledge about SNA, understanding about graphs, charts, observation ability or abstract thinking ability etc.

eTwinning dashboard helps teachers to monitor their competences through plots of indicators. It also provides tools to monitor community (e.g. school, project) competences, i.e. teachers' competences in a particular communitys.

D. *Assessing Teacher Social Capital*

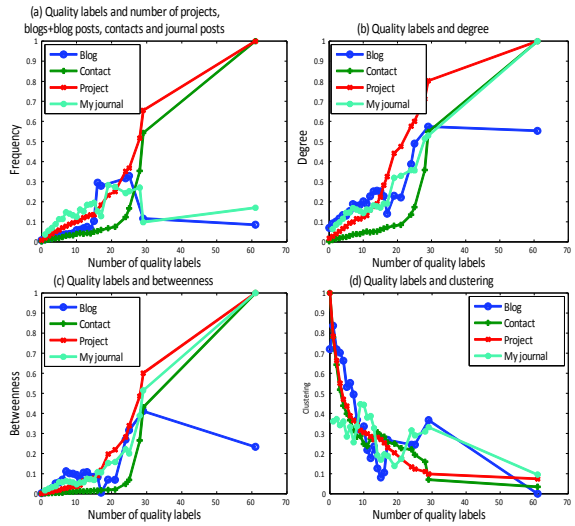

Fig. 4. Teachers' performance and network parameters.

Social capital stands for the ability of actors to derive benefits from the membership in social networks or other social structures [2]. In SNA, studies are concerned with the network structures that are the most effective factor for creating social capital. Two types are identified. Coleman [3] emphasizes the benefits of being embedded into densely connected groups, as regards to the confidence, trust and secured relationship in the community. This form of social capital is referred to as *closure*. On the other hand, Burt [4] discusses social capital as a tension between being embedded into communities and brokerage - the benefits arising from the ability to "broker" interactions at the interface between different groups. We call this form *structural hole*.

We analyze the correlation between teachers' performance (indicated by the quality labels and prizes) and their positions in eTwinning network to reveal the social capital form. *Closures* and *structural holes* can be differentiated by two network measures on nodes. *Local clustering coefficient* measures the extent to which a node is positioned in a densely-connected community. If a node is in a dense community, its local clustering coefficient is very high. *Node betweenness* measures the extent to which a particular node lies between the communities. Nodes connecting different communities together have very high betweenness. Therefore, closures are characterized by high local clustering coefficient, while nodes with high betweenness are structural holes.

The correlation between network parameters and teachers' performance (indicated by quality labels) is given in Fig. 4. It shows that teachers who are positioned at the interface between communities (i.e., *structural holes*) have a big advantage. Teachers with a high number of quality labels connect different communities together (high betweenness and low clustering coefficient), while teachers with low number of quality are clustered within communities (high clustering coefficient). The result is useful to make recommendations to teachers, e.g. suggesting projects and contacts, helping them to build their communities as well as selecting members for a particular project in the way that maximizes the social capital.

IV. Conclusion

In this paper, we discus about the application of data warehouse for LLA. The characteristics of lifelong learning such as the long time periods of learning activities and the heterogeneous interactions of learners raise a challenge for data management as well as analytical applications. Data warehouse is ideally suitable for LLA as a backend with automatic data extraction, transforming and load process. We present a case study on teacher network in Europe where we employed data ware house to deal with heterogeneous interactions between teacher and large-scale data derived from those interactions. The case study demonstrate the usefulness and efficiency of data warehouse for LLA.

References

[1] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, New York, NY, 1994.

[2] Borgatti, S. Network measures of social capital. Wire, 21(2), 1998, pp. 27-36.

[3] Coleman, James S. Social Capital in the Creation of Human Capital. American Journal of Sociology Vol. 94, 1988, pp. 95-125, the University of Chicago Press.

[4] Burt, R. S. "Structural Holes versus Network Closure as Social Capital". In N. Lin, K. Cook and R. S. Burt: *Social Capital: Theory and Research.* Sociology and Economics: Controversy and Integration series. New York: Aldine de Gruyter, 2001, pp. 31-56

[5] Fischer, G. "Lifelong Learning—More Than Training". *Journal of Interactive Learning Research, 11*(3), 2000, pp. 265-294. Charlottesville, VA: AACE.

[6] Chaudhuri, S. and Dayal, U. "An overview of data warehousing and OLAP technology". *SIGMOD Rec.* 26, 1 (March 1997), pp. 65-74.

[7] McClelland, D.C. "Testing for competence rather than for intelligence". *American Psychologist* 20 (1973), pp. 321-333.

Manh Cuong Pham, born on September 7th 1977 in Thai Nguyen, Vietnam, received his master degree in Software Systems Engineering in Computer Science at RWTH Aachen University in Aachen, Germany in 2008. Currently, he is a researcher at Information Systems and Databases Institution, Department of Computer Science, RWTH Aachen, Germany. His main research areas include social network analysis, recommender systems, data mining, information retrieval and web science. He received BIT Research School scholarship for PhD study by University of Bonn and RWTH Aachen in 2008. He worked for Tellnet project funded by EU which deals with dynamic social network analysis on the teacher networks to study the interaction between teachers in Europe.

Ralf Klamma has diploma, doctoral and habilitation degrees in computer science from RWTH Aachen University. He leads the "Advanced Community Information Systems (ACIS)" research group at the information systems chair, RWTH Aachen University. He has worked in EU projects for Technology Enhanced Learning (PROLEARN, ROLE, TELMAP, GALA and Tellnet). He is on the editorial board of IEEE Transactions on Technology Enhanced Learning, International Journal on Technology Enhanced Learning (IJTEL) and Social Networks Analysis and Mining (SNAM). His research interests include community information systems, technology enhanced learning, data management, and social network analysis.