Business Intelligent in Digital Library Data

Sunil Kumar Jain[1] , Sanjivani Mahor[2]
[1]Asst Prof in Department of CSE , Bhopal ,India
[2]Asst Prof in Department of CSE, Bhopal, ,India
Email : sunil.jain.ips@gmail.com[1] , sanjivanichhaya@gmail.com[2]

ABSTRACT: **Teachers and students increasingly enjoy unprecedented access to abundant web resources and digital libraries to enhance and enrich their classroom experiences. However, due to the distributed nature of such systems, conventional educational research methods, such as surveys and observations, provide only limited snapshots. In addition, educational data mining, as an emergent research approach, has seldom been used to explore teachers' online behaviours when using digital libraries. Building upon results from a preliminary study, this article presents results from a clustering study of teachers' usage patterns while using an educational digital library tool. The clustering approach employed a robust statistical model called latent class analysis. In addition, frequent item sets mining was used to clean and extract common patterns from the clusters initially generated. The final clusters identified three groups of teachers in this tool: key brokers, insular classroom practitioners, and inactive islanders. Identified clusters were triangulated with data collected in teacher's registration profiles. Results showed that increased teaching experience and comfort with technology were related to teachers' effectiveness in using this tool.**

Keywords: **Educational Data Mining, Educational Web Mining, Clustering, K-means, Digital Libraries, Teacher Users.**

I. INTRODUCTION

Increasingly, education and training are delivered beyond the constraints of the classroom environment, and the increasingly widespread availability of online repositories, educational digital libraries [1] and their associated tools are major catalysts for these changes [3][5](Borgman et al., 2008; Choudhury, Hobbs, & Lorie, 2002). Teachers, of course, are a primary intended audience of educational digital libraries. Studies have shown that teachers use digital libraries and web resources in many ways, including lesson planning, curriculum planning [4][15][19](Carlson & Reidy, 2004; Perrault, 2007; Sumner & CCS Team, 2010), and looking for examples, activities as well as illustrations to complement textbook materials [2][19][20](Barker, 2009; Sumner & CCS Team, 2010; Tanni, 2008). Less frequently mentioned ways are learning about teaching areas networking to find out what other teachers do [16](Recker, 2006), and conducting research [17](Recker et al., 2007). These studies, however, were generally conducted in laboratory-like settings, using traditional research methods, such as interview, survey, and observation.

Due to the distributed nature of the Web, traditional research methods and data sources do not support a thorough understanding of teachers' online behaviours in large online repositories. In response, web-based educational applications are increasingly engineered to capture users' fine-grained behaviours in real-time, and thus provide an exciting opportunity for researchers to analyze these massive datasets, and hence better understand online users [18] (Romero & Ventura, 2007).

These records of access patterns can provide an overall picture of digital library users and their usage behaviours. With the help of modern data mining techniques—the discovery and extraction of implicit knowledge from one or more large databases [8][14][18](Han & Kamber, 2006; Pahl & Donnellan, 2002; Romero & Ventura, 2007)—the data can further be analyzed to gain an even deeper understanding of users. Yet, despite the wealth of fine-grained usage data, data mining has seldom been applied to digital library user datasets, especially when studying teacher users.

The study reported in this article used a particular digital library tool, called the Instructional Architect (IA.usu.edu), which supports teachers in authoring and sharing instructional activities using online resources [16] (Recker, 2006). The IA was used as a test bed for investigating how the data mining process in general, and clustering methods in particular, can help identify the different and diverse teacher groups based on their online usage patterns. This study built substantially on results from a preliminary study that also used a clustering approach [09] (Xu & Recker, in press). In particular, both studies relied on a clustering approach that used a robust statistical model, *latent class analysis* (LCA)[12]. In addition, this study used more refined user feature space, and frequent item a set mining was used to clean and extract common patterns from the clusters initially generated. Lastly, as a means of validation the clustering results, we explored the relationship between teachers' characteristics (comfort level with technology and teaching experience) and the teacher clusters that emerged from the study.

II EDUCATIONAL DATA MINING

There is increasing interest in applying data mining (DM) to the evaluation of web-based educational systems, making educational data mining (EDM)[18] a rising and promising research field. Data mining is the discovery and extraction of implicit knowledge from one or more large databases, data warehouses, and other massive information repositories [8][14][18] (Han & Kamber, 2006; Pahl & Donnellan, 2002; Romero & Ventura, 2007). When the context is the Web, it is sometimes explicitly termed web mining [6] (Cooley, Mobasher, & Srivastava, 1997). Educational data mining, as an emerging discipline, is concerned with applying data mining methods for exploring unique types of data that come from educational settings [2](Baker & Yacef, 2009). As web-based educational applications are able to record users' fine-grained behaviours in real-time, a massive amount of data becomes available for researchers to analyze in order to better understand an application's impact, usage, and users.

The knowledge discovery and data mining (KDD) process typically consists of three phases:

1) Pre-processing datasets.
2) Applying data mining algorithms to analyze the data.
3) Post-processing results

Data pre-processing refers to all the steps necessary to convert a raw dataset to a form that can be ingested into a data mining algorithm. It may include any of the following tasks: data cleaning, missing value imputation, data transformation, and data integration. The application of data mining algorithms usually has one of two purposes: description and prediction. Description aims at finding human-interpretable patterns to describe the data; prediction attempts to discover relationships between variables, in order to predict the unknown or future values of similar variables. Currently, there is no universal standard for post-processing and evaluating data mining results. Typical interpretation techniques draw from a number of fields such as statistics, data visualization, and usability studies.

III. CLUSTERING STUDIES IN EDUCATIONAL SETTINGS

The increasing availability of educational datasets and the evolution of data mining algorithms have made educational data mining a major interdisciplinary area, lying between the fields of education and information/computer sciences. Based on [18] (Romero and Ventura's (2007)) educational data mining survey, most commonly used data mining techniques include statistical data mining, classification, clustering, association rule mining, and sequential pattern mining. This study focused on using clustering approach to analyze teachers' online behaviours when using a digital library tool. As such, several clustering studies using in educational datasets are reviewed. [10] (Hübscher, Puntambekar, & Nye (2007)) used K-means and hierarchical clustering techniques to group students who used CoMPASS, an educational hypermedia system that helps students understand relationships between science concepts and principles. K-means is a clustering analysis method that aims to partition *n* data points into *k* clusters in which each data point belongs to the cluster with the nearest cluster centre. Hierarchical clustering is a clustering analysis method that seeks to build a hierarchy of clusters. In CoMPASS, navigation data was collected in the form of navigation events, where each event consisted of a timestamp, a student name, and a science concept. After pre-processing, K-means and hierarchical clustering algorithms were used to find student clusters based on the structural similarity between navigation matrices. [7] (Durfee, Schneberger, & Amoroso (2007)) analyzed the relationship between student characteristics and their adoption and use of particular computer-based training software, using factor analysis and self-organizing map (SOM) techniques. Survey responses to questions regarding user demographics, computer skills, and experience with the software were collected from over 40 undergraduate students. They used SOM to cluster and visualize the dataset. By visually analyzing the similarity and difference of the shades and borders, four resulting student clusters were identified. Finally, a *t* test on performance scores supported the clustering decisions. [21] (Wang, Weng, Su, & Tseng (2004)) combined sequential pattern mining with a clustering algorithm to study students' learning portfolios. The authors first defined each student's sequence of learning activities as a learning sequences, LS = $<s_1 s_2 \ldots s_n>$, where s_i was a content block. They then applied a sequential pattern mining algorithm to find the set of maximal frequent learning patterns from learning sequences. The discovered patterns were considered as variables in a feature vector. For each learner, the value of bit *i* was set as 1 if the pattern *i* was a subsequent of the original learning sequence, 0 otherwise. After the feature vectors were extracted, a clustering algorithm called ISODATA was used to group users into four clusters.

The literature review only identified one clustering study investigating teachers' use of an educational digital library tool. In this study, a clustering approach was applied to model and discover patterns in teachers' using an online curriculum planner [13] (Maull, Saldivar, & Sumner, 2010). In this study, user sessions were first abstracted, and 27 features were selected for clustering experiments. The study then used K-means and expectation-maximum (EM) likelihood to cluster the user sessions. The two algorithms identified very similar patterns in the largest clusters, such as clicking on *instructional support materials*, *embedded assessments*, and *answers and teaching tips*. However, the authors acknowledged that their study was preliminary, in

that there was not complete agreement between the different algorithms on top cluster features or cluster sizes.
There are other clustering studies documented in the literature on educational web mining, however, the above examples are sufficient in revealing some major considerations in discovering user groups in the context of online environments, as follows:

- A user-model must be carefully defined that accounts for the task and domain. Navigational paths, online performance, user characteristics, and a user's prior knowledge are all good candidates for user features.
- Clustering is a generic definition for a certain type of data mining method. Researchers must select the clustering algorithm appropriate for their studies; however, different approaches may produce different results.
- Other data mining methods such as rule discovery, dimensionality reduction, and filling in missing values can be used with clustering algorithms to achieve a better grouping effect.
- To better understand online user behaviors and produce more useful information, the data mining results should be used in conjunction with other data.
- As an indispensible component of the KDD process, evaluation of the clustering results should be conducted if at all possible.

A. Data Mining Techniques:

Clustering is often used for discovering classes, patterns or structure in data. Cluster analysis for grouping together genes or samples with similar expression patterns. Objects in the same cluster are similar to each other than objects in different clusters. Applications of clustering algorithms to Library data are for the find the informative data like no of student in library, no. of issued book etc. Pattern recognition methods can be divided into two categories:

- supervised
- Unsupervised.

In unsupervised clustering we used k-mean clustering algorithm.

Unsupervised methods (K- Means) because the k-means algorithm is an evolutionary algorithm that gains its name from its method of operation.

Input: The number of clusters k and a database with n objects.

Output: A set of cluster that minimizes the squared-error criterion.

Complexity O (nkt),

n = number of objects,

k = number of cluster and t= number of iteration.

Method:

Step1: Randomly select k object from the data points, each of which initially represents a cluster mean or centre.

Step2: for remaining objects, an object is assigned to the cluster to which it is most similar, based on the (Euclidean or correlation) distance between the object and the cluster mean

Step3: Then compute the new mean for each cluster

Step4: repeat this process until no change in cluster mean this process is used to minimize the squared-error criterion of a set k clusters.

This Clustering method we use in our project to mine the library data

B. Working Steps:

First we Pre-process the data using Advance ETL Processor. This tool is used for-

Data filtering

Data sources typically contain large amounts of data. Reports usually need only a specific subset of data that meets certain conditions. You can select specific records to use in a report by using filters. For example, rather than get information about all customers, you can create filters to select customers in a certain region or customers with a certain credit rank. You can also design filters that provide the report user with the opportunity to specify the filter conditions when the report runs. This section discusses creating filters for which you specify the conditions at design time.

Data Cleaning

Data cleaning tasks include record matching, reduplication, and column segmentation which often need logic that go beyond using traditional relational queries. This has led to development of utilities for data transformation and cleaning.

Data Integration

Data integration involves combining data residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of

situations, which include both commercial (when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example) domains.

Data Normalization

Data normalization is the process of reducing data to its canonical form. For instance, Database normalization is the process of organizing the fields and tables of a relational database to minimize redundancy and dependency

Data Reduction

Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts.

In Library data we clean and filter the unnecessary attribute and record.

For k-mean clustering we used software PASW Modeller and analysis). This software can do data clustering, visualization, and analysis for numeric (e.g. gene expression data) as well as sequence data. This is a web-based tool.

Using PASW Modeller one can do hierarchical and K means clustering.

Algorithm of PASW:

Input: Data file, value of k, Distance (Euclidean or Pearson correlation)

Output: cluster data according to similarity

Step1: upload data file

Step2: Select clustering process (k-mean)

Step3: Define value of k

Step4: Select distance (correlation)

Step5: Get cluster according to value of k

1) **Results:**

In this paper we mine the library data of DAVV and find some results which can be helpful for the university.

Following results we have find:

1. No. of members in different departments.
2. Categorized the members as per the designation.
3. No of books issued and available in library.
4. Categorized the member according to their caste in different department.

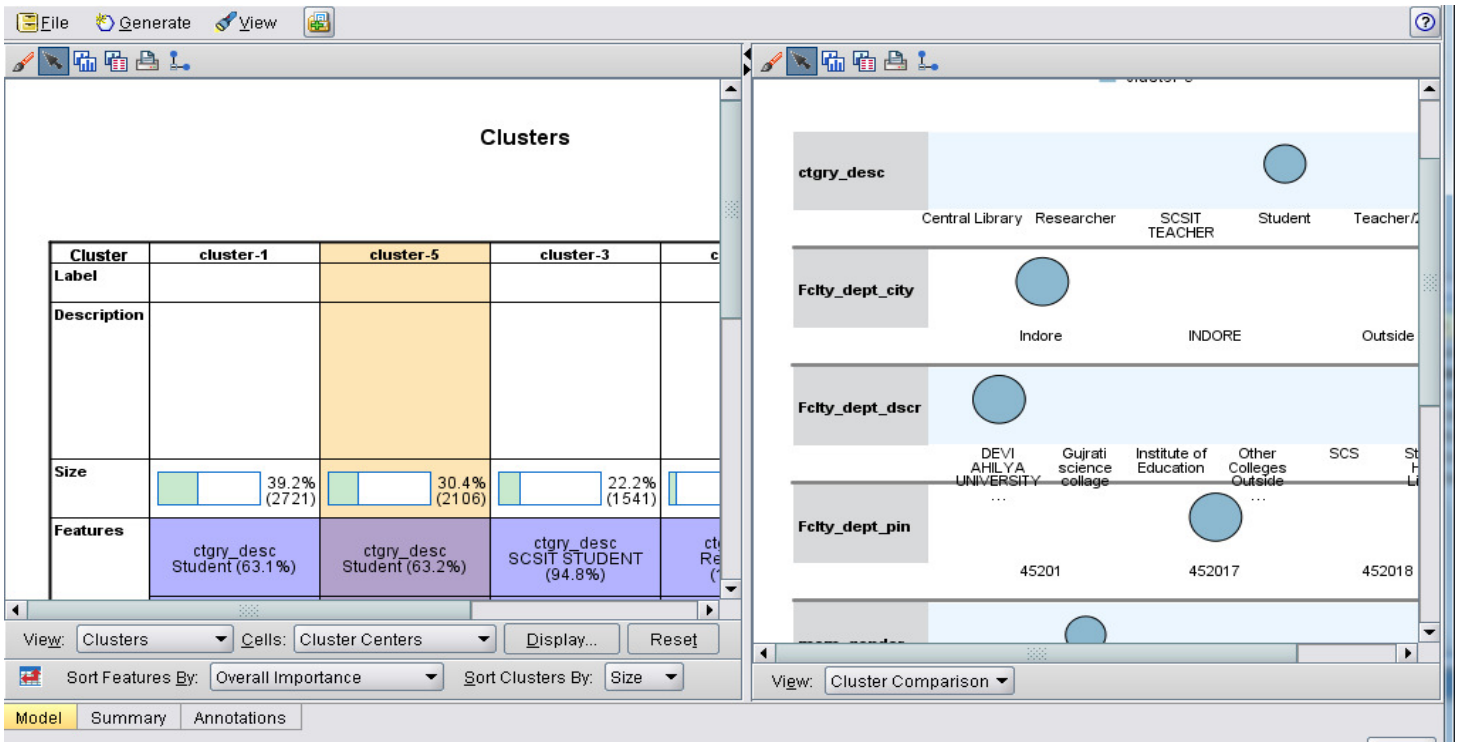

Fig 1.Cluster of Library's Members data by Category wise, department wise etc.

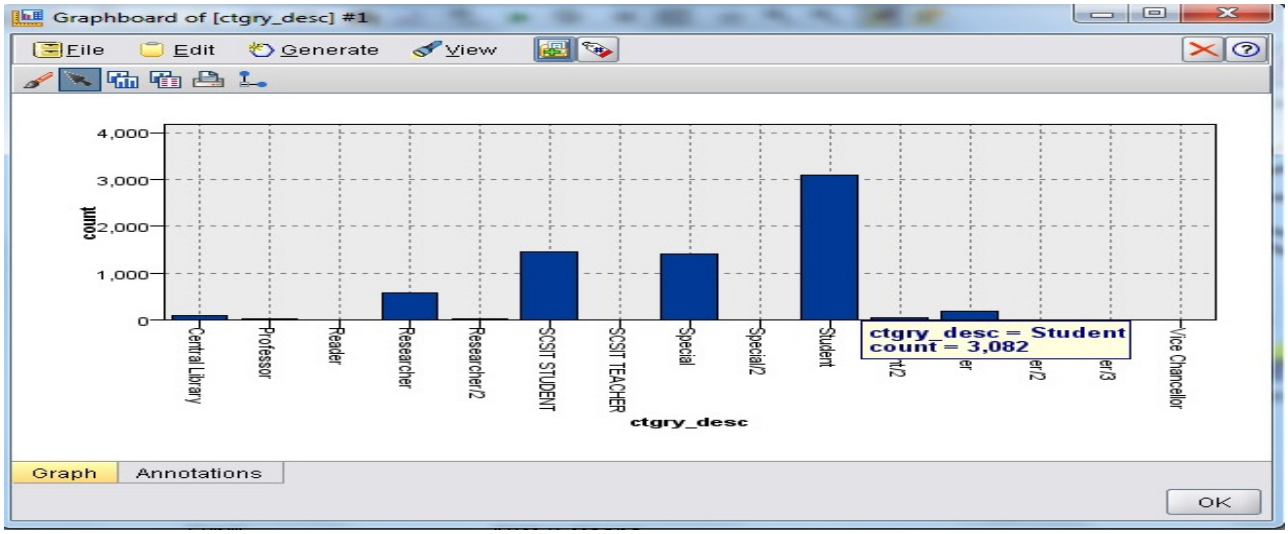

Fig 2.Graph Data of Library's Member by Category wise

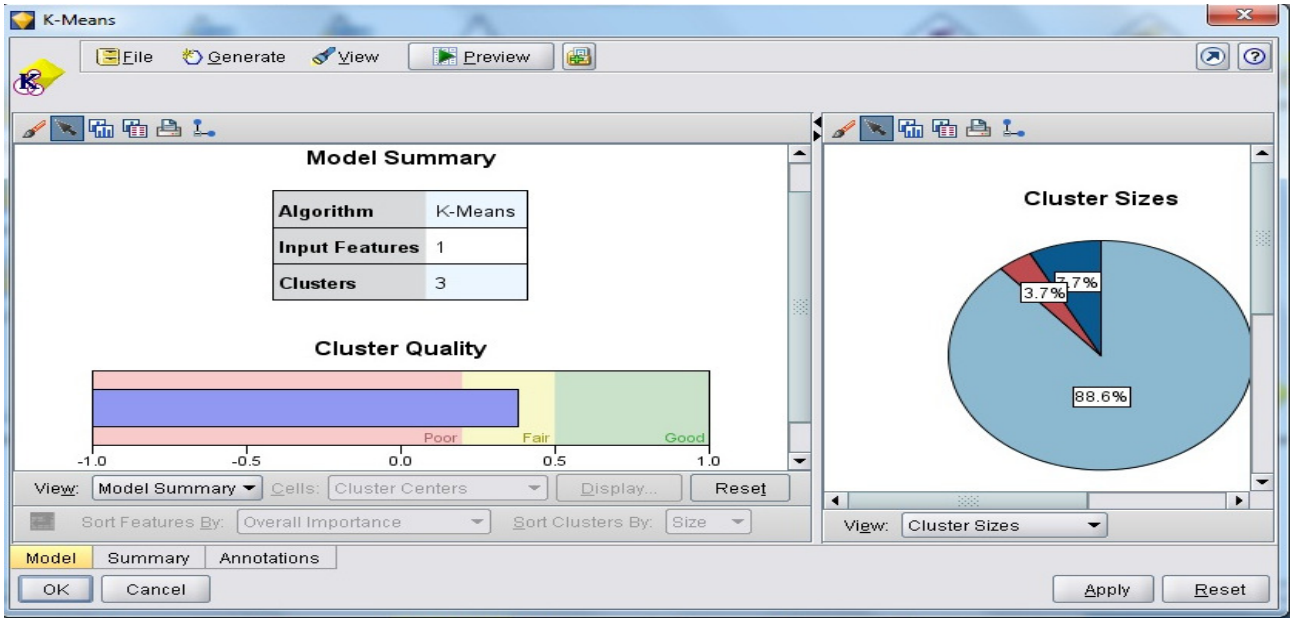

Fig. 3 Clusters of Library Member's Data by Member type (Caste type)

IV. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

This research examined and analyzed teachers' online behaviours in the context of a digital library tool, the Instructional Architect. First, an educational data mining approach, clustering, was applied to identify different groups of IA teacher users according to their diverse online

behaviours. A user model consisting of nine features was identified and fed into a LCA model, clustering IA teacher users into three groups, labelled *key brokers, insular classroom practitioners,* and *inactive islanders*.

Second, a triangulation study examined relationships between teachers' profile data and their usage patterns. This analysis showed strong relationships between teachers' characteristics and their online behaviours as described by user clusters. Specifically, teachers with more teaching experience were more likely to be *key brokers*, and those with less teaching experience were more likely to demonstrate ineffective use of the IA. Teachers who were more comfortable with technology were more likely to be *key brokers* and were least likely to be *insular classroom practitioners*. Such results show that effective usage of the Instructional Architect requires both pedagogical knowledge (gained through experience teaching) and technological knowledge. This finding helps to predict which kinds of teachers are more likely to adapt technology tools such as digital libraries, and more importantly, how to help teachers become more effective digital libraries users.

Three areas are proposed for future work. First, although LCA is alleged to outperform K-means, no competing clustering algorithm has been implemented to justify this choice. Secondly, previous work showed that greater use of the IA occurs in geographical areas where teacher professional development workshops using the IA
have been conducted [11] (Khoo et al., 2008; Xu, Recker, & Hsi, 2010). This suggests that workshop participants have a higher chance of becoming *sticky* users. Therefore, teachers who participated in such workshop can be singled out for detailed nalysis, as their distribution among clusters is predicted to be different. Finally, the third stage of KDD, evaluation and interpretation, could be conducted in a more comprehensive fashion. For example, the survey information filled out by workshop participants could be used to triangulate the clustering results, providing evidence for why and how the teachers like and dislike the IA.

Despite the current challenges, the field of educational data mining is making progress towards standardizing its procedures for tackling educational problems. This research shows that teachers' use of online resources can be studied by productively using web usage data and employing data mining approaches to investigate digital library problems in innovative ways.

REFERENCES

[1] Barker, L. J. (2009). Science teachers' use of online resources and the digital library for earth system education. *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 1-10).

[2] Baker, R. S. J. D., & 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.

[3] Borgman, C., Abelson, H., Dirks, L., Johnson, R., Koedinger, K., Linn, M. & Szalay, A. (2008).

[4] Carlson, B., & Reidy, S. (2004). Effective access: Teachers' use of digital resources (research in progress). *OCLC Systems and Services, 20*(2), 65-70.

[5] Choudhury, S., Hobbs, B., & Lorie, M. (2002). A framework for evaluating digital library services. *D-Lib Magazine, 8.*

[6] Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Web Web. *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence* (pp.558-567).

[7] Durfee, A., Schneberger, S., & Amoroso, D. L. (2007). Evaluating students' computer-based learning using a visual data mining approach. *Journal of Informatics Education Research, 9,* 1-28.

[8] Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco, CA: Kaufmann.

[9] Xu, B., & Recker, M. (in press). Understanding teacher users of a digital library service: A clustering approach. *Journal of Educational Data Mining, 3*(3).

[10] Hübscher, R., Puntambekar, S., & Nye, A. H. (2007). Domain specific interactive data mining.

[11] Khoo, M., Pagano, J., Washington, A. L., Recker, M., Palmer, B., & Donahue, R. A. (2008). Using web metrics to analyze digital libraries.

[12] Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 175-198).

[13] Maull, K. E., Saldivar, M. G., & Sumner, T. (2010, June). *Online curriculum planning behavior of teachers.*

[14] Pahl, C., & Donnellan, D. (2002). Data mining technology for the evaluation of web-based teaching and learning systems. In M. Driscoll & T. Reeves (Eds.), *Proceedings of the E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 747-752).

[15] Perrault, A. M. (2007). An exploratory study of biology teachers' online information seeking practices. *School Library Media Research*, *10*.

[16] Recker, M. (2006). Perspectives on teachers as digital library users: Consumers, contributors, and designers. *D-Lib Magazine, 9*(3).

[17] Recker, M., Walker, A., Giersch, S., Mao, X., Palmer, B., Johnson, D., Robertshaw, B. (2007). A study of teachers' use of online learning resources to design classroom activities. *New Review of Hypermedia and Multimedia*, *13*(2), 117-134.

[18] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

[19] Sumner, T., & CCS Team. (2010). Customizing science instruction with educational digital libraries. *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp.353-356).

[20] Tanni, M. (2008). Prospective history teachers' information behaviour in lesson planning. *Information Research, 13*(4).

[21] Wang, W., Weng, J., Su, J., & Tseng, S. (2004, October). *Learning portfolio analysis and mining in SCORM compliant environment.* [22] [22] Xu, B., & Recker, M. (in press). Understanding teacher users of a digital library service: A clustering approach. *Journal of Educational Data Mining, 3*(3).