Understanding Teacher Users of a Digital Library Service: A Clustering Approach

BEIJIE XU
Utah State University
beijie.xu@aggiemail.usu.edu
and
MIMI RECKER
Utah State University
mimi.recker@usu.edu

Department of Instructional Technology & Learning Sciences
Utah State University
Logan, Utah USA 84322-2830

This article describes the Knowledge Discovery and Data Mining (KDD) process and its application in the field of educational data mining (EDM) in the context of a digital library service called the Instructional Architect (IA.usu.edu). In particular, the study reported in this article investigated a certain type of data mining problem, clustering, and used a statistical model, latent class analysis, to group the IA teacher users according to their diverse online behaviors. The use of LCA successfully helped us identify different types of users, ranging from window shoppers, lukewarm users to the most dedicated users, and distinguish the isolated users from the key brokers of this online community. The article concludes with a discussion of the implications of the discovered usage patterns on system design and on EDM in general.

Keywords: Educational Data Mining, Educational Web Mining, Clustering, Latent Class Analysis, Digital Libraries, Teacher Users

1. INTRODUCTION

Increasingly, education and training are delivered beyond the constraints of the classroom environment, and educational digital libraries and their associated services are making major contributions to these changes [Choudhury et al. 2002]. With the rapid growth of e-learning environments and information networks, researchers as well as stakeholders need to ensure efforts and resources expended on the development of digital libraries are worthwhile in terms of their impact on targeted users.

Teachers, of course, are a primary intended audience of educational digital libraries. Yet little is known about the impact of these novel tools and services on the wide range of teachers and their resulting instructional practices. As these tools can be engineered to capture fine-grained footprints of user activities, opportunities exist to apply emerging educational data mining strategies to analyze web usage data so as to better understand digital libraries' teacher users.

This article focuses on a particular digital library service, called the Instructional Architect (IA.usu.edu), as a test bed for investigating how the Knowledge Discovery and Data Mining (KDD) process in general, and clustering methods in particular, can help identify the diverse teacher user groups and their characteristics. As will be described in more detail below, the IA is an educational digital library service that supports teachers in authoring and sharing instructional activities using online resources [Recker et al. 2006; 2007]. Currently, the IA has over 5,500 registered

This material is based upon work supported by the National Science Foundation under Grants No. 840745. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank the members of the IA research group, especially Bart Palmer, and our dedicated IA users.

teachers. This study applies the KDD process to the usage data from teachers who registered in 2009. The study involved defining a user model, examining the teacher groups that emerge through a cluster analysis, and making inferences about these clusters.

The article is structured as follows. The literature review describes the generic Knowledge Discovery and Data Mining (KDD) process, several existing clustering studies in the field of educational web mining, and empirical studies of teachers' uses of online resources. This is followed by a brief introduction to the Instructional Architect service. We then describe our clustering approach, specifically Latent Class Analysis, starting from data collection and selection, to data analysis, interpretation and inference. We conclude with the implication of this study on the design of such systems, on educational data mining more generally, and on how to improve and complement clustering approaches.

2. LITERATURE REVIEW

2.1 Educational Web Mining

There is growing interest in using data mining in the evaluation of web-based educational systems, making educational data mining (EDM) a rising and promising research field [Romero and Ventura 2007]. Data mining is the discovery and extraction of implicit knowledge from one or more large collection of data [Pahl and Donnellan 2002; Romero and Ventura 2007]. 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 [Baker and Yacef 2009].

Increasingly educational learning environments, including educational digital library services, are accessed through the Web, thereby enabling a low-cost mechanism for collecting users' fine-grained behavior in real-time, and thus leaving behind a massive amount of data to analyze. Web mining, in response to this phenomenon, is a particular category of data mining problem that seeks to discover implicit patterns from usage of web documents and services [Chen and Chau 2004]. This study contributes to the field of educational web mining by investigating how to apply data mining to a particular online digital library service.

2.2 Knowledge Discovery and Data Mining

Web mining typically follows the standard KDD process, entailing: 1) data cleaning and integration, 2) selection and transformation, 3) application of data mining algorithms, 4) evaluation and presentation [Han and Kamber 2005; Witten and Frank 2005]. Often the first two phases are combined and called data preprocessing [Cooley et al. 1997; Romero and Ventura 2007].

In general, web mining serves two purposes: description and prediction. Description aims at finding human-interpretable patterns that describe the data. Prediction analyzes the existing data, and discovers relationships among the variables, in order to use such information to predict the unknown or future values of similar variables. Our study was intended to cluster Instructional Architect users in order to better understand typical teachers' online behaviors; as such, it falls into the description domain.

2.3 Clustering in Educational Web Mining

The increasing availability of educational datasets and the evolution of data mining algorithms have made educational web mining a major interdisciplinary area, lying between the fields of education and information science. Romero and Ventura [2007] summarized web mining work as falling into the following categories: 1) clustering, classification and outlier detection, 2) association rule mining and sequential pattern mining and visualization, and 3) text mining. Among the large volume of literature related to each of these areas, this article focuses on clustering methods, and some representative clustering studies of educational datasets are reviewed in the following paragraphs.

Clustering is an unsupervised learning model for grouping physical or abstract objects in the case when there is neither a predefined number of clusters nor pre-labeled instances. Clustering algorithms normally group data based on two measures: the similarity between the data objects within the same cluster (minimal intra-cluster distance), and the dissimilarity between the data objects of different clusters (maximal inter-cluster distance).

Hübscher et al. [2007] used K-means and hierarchical clustering techniques to group students who have used CoMPASS, an educational hypermedia system that helps students understand relationships between science concepts and principles. 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 preprocessing, K-means and hierarchical clustering algorithms were used to find student clusters based on the structural similarity between navigation matrices.

Durfee et al. [2007] analyzed the relationship between student characteristics and their adoption and use of computer-based educational technology using factor analysis and self-organizing map (SOM) techniques. Survey responses to questions regarding user demographics, computer skills, and experience with a particular computer-based training software were collected from over 400 undergraduate students. In order to reduce the dimensionality of the dataset, the researchers first used factor analysis to group 28 variables into 8 orthogonal factors. They then used SOM, a two-dimensional representation of input space by identifying the borders between clusters to cluster and visualized the datasets into the eight individual feature planes. These feature planes were then combined into one landscape of hexagons of different shades and border colors. By visually analyzing the similarity and difference of the shades and borders, four resulting student clusters were identified in the end. Finally, a t-test on performance scores supported the clustering decisions. That is, student performances between the groups determined by SOM based on learner characteristics were significantly different.

Wang et al. [2004] combined sequential pattern mining with a clustering algorithm to study students' learning portfolio. The authors first defined each student's sequence of learning activities as 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 (LS). The discovered patterns were considered as variables of a feature vector. For each learner, the value of bit *i* was set as 1 if the pattern *i* was a subsequence of original learning sequence, 0 otherwise. After the feature vectors were extracted, a clustering algorithm called ISODATA was used to group users into 4 clusters.

Lee [2007] proposed to assess student knowledge and infer important knowledge states (mastery levels) in an integrated online environment using SOM K-means and principle component analysis (PCA). SOM K-means

involves two steps: the first step is to generate a self-organizing map using student data; the second step is to use K-means algorithm to cluster the map into predefined number of clusters. PCA was used to identify significant feature vectors. A test consisting of 20 items associated with different learning concepts was collected from 90 students. Subsequently, SOM K-means was used to identify student clusters, with each cluster's centroid as a representation of that cluster's knowledge states. Applying PCA upon the cluster centroids helped identify two significant feature vectors – two important knowledge mastery levels. Comparisons with other algorithms showed that applying PCA over SOM K-means could reveal more significant feature vectors (knowledge states) than PCA on the original data set.

Our literature review only identified one clustering study investigating teachers' use of an educational digital library service. In this study, a clustering approach was applied to discover patterns in teachers using an online curriculum planner [Maull, Saldivar, and Sumner, 2010]. This study first abstracted user sessions and selected 27 features for clustering experiments, and then used K-means and Expectation-Maximum 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 on top cluster features or cluster sizes.

Other clustering studies exist 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 e-learning environments, as follows. 1) A user-model must be carefully defined according to the topic to be studied. Navigational path, online performance, user characteristics, and a user's prior knowledge are potential choices of user features. 2) Clustering is a generic definition of a certain type of data mining method. Researchers can select the algorithms appropriate for their studies; however, different approaches may produce different results. 3) Other data mining methods such as rule discovery, dimensionality reduction, and filling in missing values can be incorporated with clustering algorithms to achieve a better grouping effect. 4) As an indispensible component of the KDD process, evaluation of the clustering results should be conducted if at all possible.

2.4 Teachers' Uses of Educational Digital Libraries

The term "digital library" commonly refers to the electronic extension, augmentation, enhancement and integration of functions of a traditional library [Borgman, 1999]. As the functionality of digital libraries has continued to evolve beyond search and archive, they have become a type of information network that serves a community of users who integrate, create, and repurpose a collection of information resources stored in multimedia repositories [Borgman 1999; Malik and Jain 2006; Zia 2001]. Educational digital libraries are a particular type of digital libraries, which provides and facilitates the access, retrieval, and use of a wide range of web resources for supporting teaching and learning. In the second area of our review, we focused on studies of teachers' perception and uses of educational digital libraries to help better situate our study.

Though researchers do not appear to have reached consensus on whether teachers underuse [e.g., Perrault 2007] or adequately use [e.g., Carlson and Reidy 2004] digital library resources, studies have similar findings on how teachers utilize web resources to enrich their teaching experience. The most frequently mentioned ways of using

web resources are lesson planning, curriculum planning [Carlson and Reidy 2004; Perrault 2007; Sumner and CCS Team 2010], and looking for examples, activities and illustrations to complement textbook materials [Baker, 2009; Sumner and CCS Team, 2010; Tanni, 2008]. Less frequently mentioned ways are getting background reading, becoming acquainted with teaching topics [Sumner and CCS Team, 2010; Tanni, 2008], networking to find out what other teachers do [Recker, 2006], and conducting research [Khoo, 2006]. It should be noted that some teacher information seeking activities are intertwined, as characterized in the following excerpt from Sumner and Marlino's [2004] article:

> Teachers planning a lesson on plate tectonics can browse a concept map on "Processes that Shape the Earth" to better understand the role of plate tectonics in relationship to other Earth processes, to locate resources that take this process-perspective, or to access research information on common student misconceptions, etc. (p. 173)

The above quote demonstrates that within one search session, teachers can complete multiple information seeking tasks, such as background reading, finding useful resources, and conducting research on student learning.

It is worth noting that sometimes teachers prefer to construct their own digital materials and activities instead of employing "pre-packaged" web resources [Pattuelli, 2008; Tanni, 2008]. Rather than use these resources as they are, they "remix" resources to better integrate and tailor them to their individual students' ability [Sumner and CCS Team, 2010].

The studies we reviewed mostly relied on traditional research methods such as interview [e.g., Aivazian et al. 2003; Baker 2009; Perrault 2007; Recker et al. 2007; Sumner and CCS Team 2010; Tanni 2008], observation [Baker 2009; Recker et al. 2007], focus group [e.g., Carlson and Reidy 2004; Shreeves and Kirkham 2004; Sumner et al. 2003], self-report and reflection [e.g., Baker, 2009; Shreeves and Kirkham, 2004], and surveys using qualitative or quantitative measures [e.g., Aivazian et al. 2003; Carlson and Reidy 2004; Khoo 2006; Perrault 2007; Recker et al. 2007; Sumner et al. 2003]. Study foci ranged from teachers' information seeking practices [e.g., Perrault 2007; Tanni 2008] to their information use [e.g., Carlson and Reidy 2004; Tanni 2008], from teachers' attitudes toward and perceptions of using digital libraries and the broader Internet in instructional planning [e.g., Recker et al. 2007; Sumner and CCS 2010], to the motivation and barriers to digital library adoption and usage [e.g., Baker, 2009; Sumner and CCS 2010].

Similar in spirit to EDM, web metrics analyses have been applied in studies of digital library usage. Asunka et al. [2008] analyzed the Gottesman Libraries' server log. They analyzed popular search terms, investigated the relative popularities of the various electronic resources and services, and the relationship between the access to library resources and the location of their links on the homepage. In another series of studies, Khoo et al. [2008] reviewed the use of web metrics in four digital libraries projects: the Instructional Architect, the Library of Congress, the National Science Digital Library (NSDL), and Teachers' Domain. Results focused on how to track users' search and page viewing habits, how session length may be used to understand resource/collection use, and how to track users' geographic locations. For example, by examining session length, the authors noted there are several kinds of Teachers' Domain users. One group tended to be referred to the website with very specific queries, while another group came to the site, browsed, then selected from the resource offerings. In another example,

session data and usability testing for the National Science Digital Library website suggested that most users accessed the NSDL in very straightforward ways: they entered a search term and clicked on the desired result.

3. THE INSTRUCTIONAL ARCHITECT

This research is set within the context of the Instructional Architect (IA.usu.edu), an educational digital library service developed for supporting authoring of simple instructional activities using online resources in the National Science Digital Library (NSDL.org) and on the Web [Recker et al. 2006; 2007]. With the IA, teacher users are able to search, select, sequence, annotate and reuse online learning resources to create instructional activities, called *IA projects* (also referred to as "projects" for simplicity), which can be kept private, made available to only teachers' students, or to the wider Web. Figure 1 shows an example of a simple IA project created by one of our teacher users. Using the IA, this teacher created the layout and text, along with links to online resources discovered in the NSDL. As can be seen, an IA project has the author's screen name and title on the top, followed by a brief overview, and then the project content with resource links embedded, as illustrated in Figure 1.

3.1 Service

To use the IA, a teacher must first create a free IA account, which provides exclusive access to his/her saved resources and projects. As part of the registration process, the teacher completes a profile indicating subjects and grades taught, teaching experience, and level of information literacy.

After a teacher logs in, the IA offers two major usage modes: resource management and project management. In the resource management mode, teachers can search for and store links to NSDL resources, add links to online resources, or copy other people's IA projects. Figure 2 shows the list of resources stored by a teacher named "D. Schuehler", the author of the project in Figure 1. Those highlighted resources have been used in at least one of her projects.

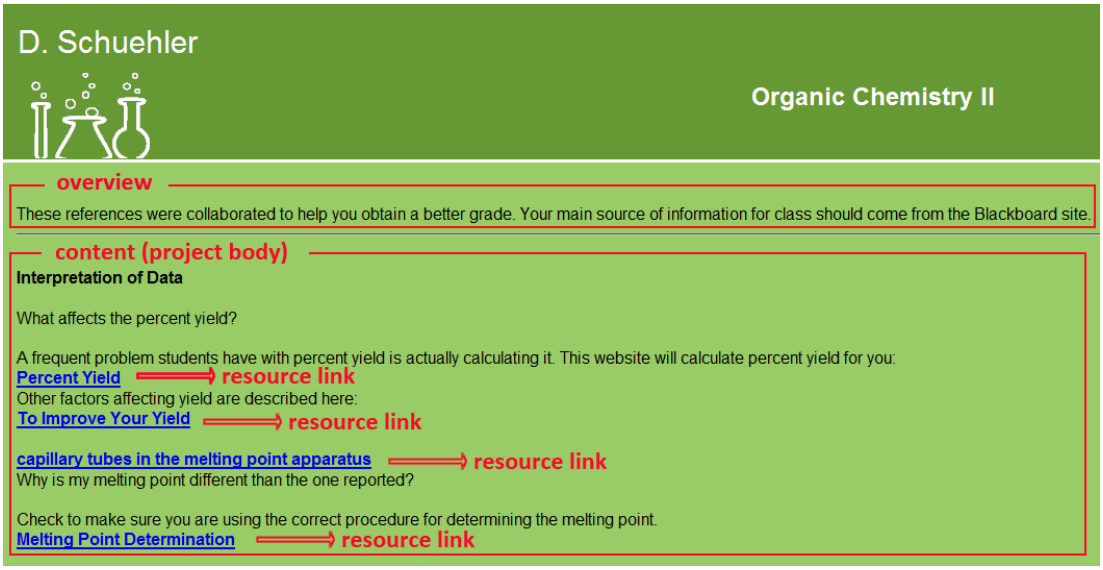

Fig. 1. An IA project named "Organic Chemistry II" created by D. Schuehler.

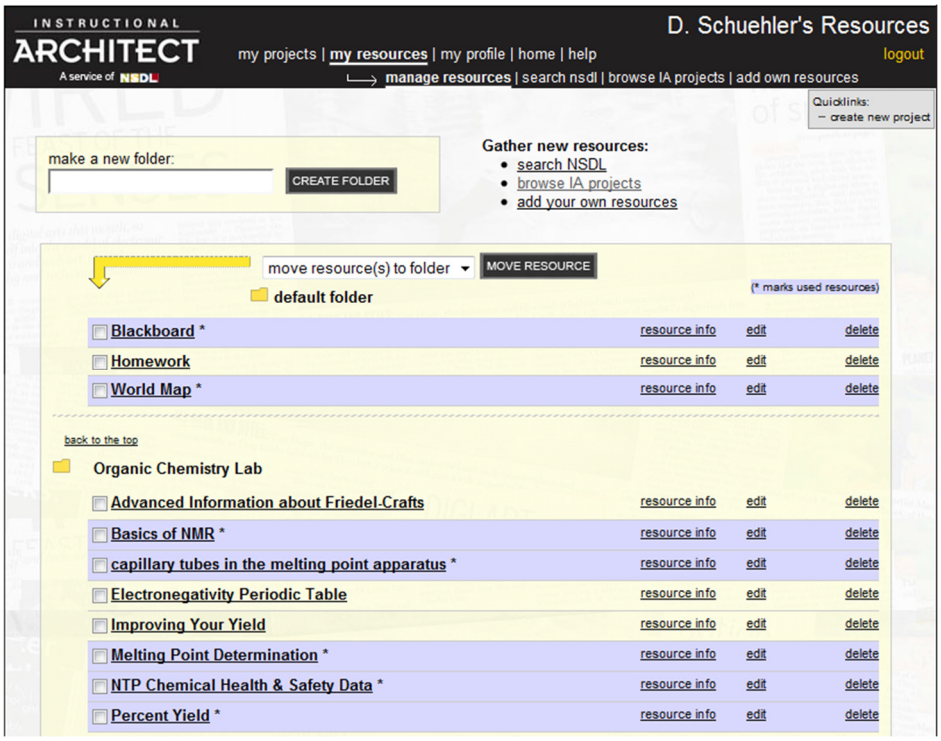

Fig. 2. Searching for NSDL resources inside the IA.

With the IA's user-friendly project management and creation interface, a teacher only need to enter an IA project's title, overview, and content, and the system will generate a webpage dynamically. The teacher's resource collections are listed on the left, and can be added to the project by clicking on the arrows behind (see Figure 3). When a project is generated on request, those resources will be converted to hyperlinks. JavaScript and HTML are supported, which means dynamic objects such as multimedia, blogs, and RSS can be included. Teachers can add basic metadata to describe their project, such as subject area, grade level, and core curriculum standard, and these metadata are used to support search and browse of public IA projects.

Once completed, a project can be marked as public, student-view, or private. Anyone can visit a public project, students can access their teachers' student-view projects through their student accounts, and private projects are only viewable by the author. All public projects are saved under the Creative Commons' *free to share and free to remix* license. As noted above, a registered teacher can make a duplicate of any public project by clicking the "copy" button at the bottom of the webpage. In this way, the IA provides a service level for supporting a teacher community around creating, remixing, and sharing instructional resources and activities.

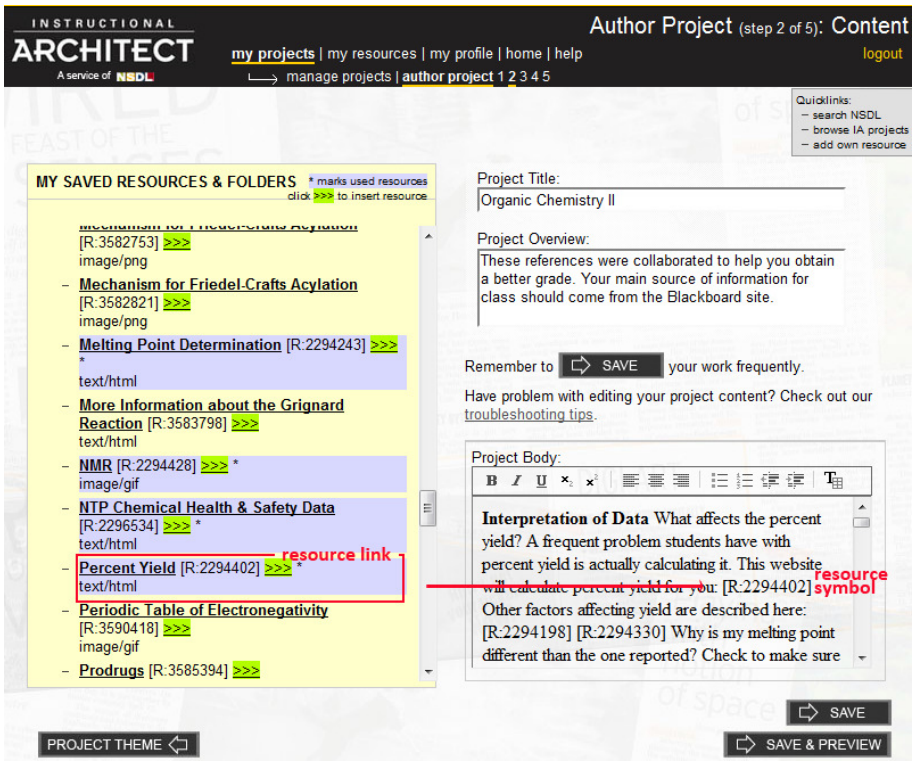

Fig. 3. Project creation interface.
Resources are listed on the left, and a user enters content on the right.

Figure 4 presents the data model for the Instructional Architect. As can be seen, teachers play a central role in this model, and are therefore the target of this study.

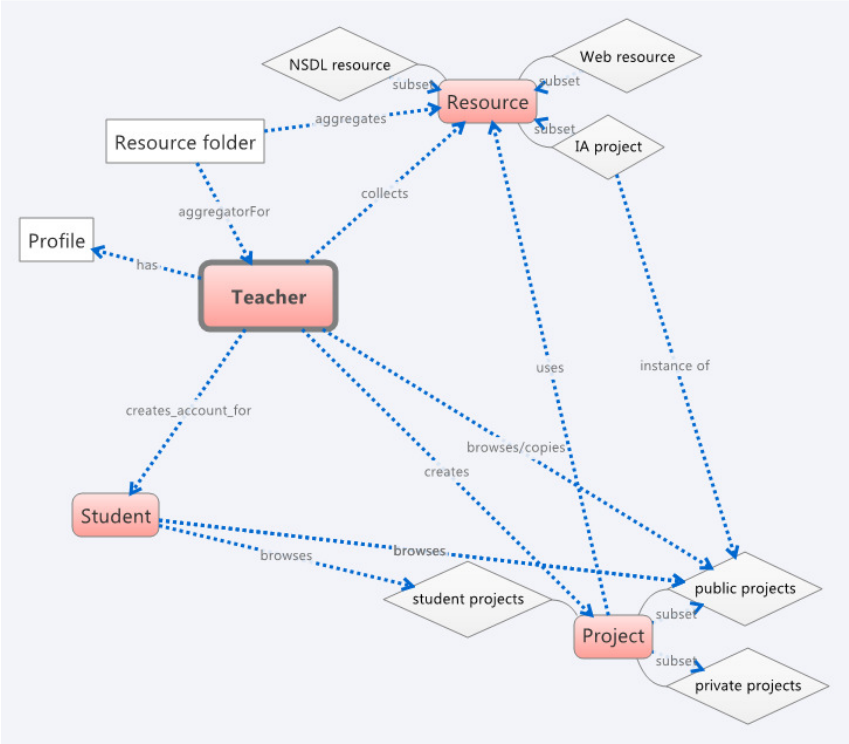

Fig. 4. The Instructional Architect's data model.

3.2 Usage

From 2002 to October 2010, over 5,500 teachers have registered with the IA, more than 12,000 IA projects have been created, and links to over 54,000 online resources have been added by teachers to the database. Since August 2006, public projects have been viewed over 1 million times. Compared with large-scale national digital libraries such as ERIC and the NSDL, the IA provides medium-sized datasets that are manageable for data mining in terms of magnitude, yet large enough to contain diverse usage patterns.

4. PURPOSE AND RESEARCH QUESTIONS

The previous section describes the wide range of activities a teacher can engage in within the context of the IA: searching for and selecting online resources, browsing, creating, and sharing instructional activities, etc. Yet little is known about these teacher users of the IA and their diverse behaviors. In addition, the literature review found that, despite the wealth of usage data collected by digital libraries, studies still generally relied on conventional research methods. In terms of EDM studies using clustering approaches, our review distilled key considerations, which informed our research questions below.

The purpose of this research is to use the IA as a test bed for investigating how to apply web mining methods – clustering approaches in particular – to better understand teachers' use of this digital library service. The following four research questions guided this study within the context of the IA:

1. What usage model best characterizes teachers' usage activities?
2. What usage patterns and clusters emerge when mining teacher usage data?
3. What inferences can be made about teachers' behaviors from the discovered usage patterns?
4. What are the implications of the discovered usage patterns for design?

5. METHODOLOGY

Web mining the IA dataset followed the three-phase KDD process – data preprocessing, applying data mining algorithms, and data post-processing. Thus, the KDD framework was used to build the research methodology.

5.1 Data Preprocessing: Data Sources

A powerful educational system should be powered by a multi-function database, which not only stores the instructional content, but also tracks all user interactions [Talavera and Gaudioso 2004]. The IA relational database serves such a purpose. In addition to information related to IA functionality, the database contains several tables built to store user traces. For example, a table called saved_projects stores all past versions of every IA project, providing an avenue to examine how teachers develop and shape their projects; a table called tracking_hits records any hit on an IA resource or an IA project, and stores the IP address, user ID, timestamp, session ID, referrer page, and target object (either an IA resource or IA project); the tracking_page_hits table stores similar information but on a finer-grained level – in addition to requests to IA resources and projects, it records almost every user click on an IA webpage. In this study, the IA database was taken as the only data source for analyzing teachers' resource and project management and navigation profiles.

5.2 Data Preprocessing: User Feature Space

As mentioned earlier, IA teachers were the focus of this study. In order to construct a comprehensive user-model, the major roles a teacher plays in the IA environment were first outlined, and then behaviors under each role were summarized, and, lastly, measurable metrics and features were defined to describe the behaviors under each category.

A teacher can assume three general roles in the IA environment: resource collection, project authoring and usage, and navigation. Data from these three roles were included in the feature space for representing a teacher's online behavior, and are explained next.

Role I – Resource collection and usage. Behaviors in this role include: collecting resources from the NSDL; storing links to favorite IA projects and other web resources; organizing the collected resources into folders; and embedding resources into projects. Four related metrics were:

1. *Number of resources collected*. The total number of resources collected regardless of resource origin.
2. *Number of resource folders*. The number of folders reflects the diversity of a teacher's interests and how well s/he organizes them.
3. *Resource usage rate.* This metric captures a teacher's use of online resources in projects.
4. *Number of resources per project.* This metric measures how many resources a teacher would like to use in explaining one set of instructional concepts.

Role II – Project authoring and usage. Behaviors in this role include: creating projects; copying projects; editing projects; choosing different publishing options; and implementing projects as measured by hits on the project. Related metrics were:

1. *Number of projects*. Because private projects are inaccessible to anyone but the author, only public and student projects are counted when measuring teachers' productivity using the IA and their contributions to this community.
2. *The percentage of each type of projects.* A preliminary analysis showed that 24% of the teachers created only private projects, while 29% never kept a project private. Project type reflects a teacher's motivation in creating a project and its target audience; for instance, whether there are any student-viewable projects intended solely for classroom use.
3. *The percentage of copied projects*. The ratio between copied and original projects indicates: 1) teachers' willingness to copy others teachers' projects, 2) the relative weight between being a consumer and a contributor in this community.

It is difficult to measure the quality of an IA project without examining its actual content. However, determining the quality of online content remains a "grand challenge" [Grimes, 2007] and it is virtually impossible to rate a project using text mining techniques, due to each project's unique context, possible occurrence of fractured and ungrammatical syntax, or occasional irregular spellings and abbreviations [Grimes 2007]. To compensate somewhat for this limitation, the following six indicators were used as a proxy to measure the quality of a project.

1. *Number of resources per project.*
2. *The amount of content per project (measured by words, excluding the text in resource links).*

3. *Ratio between the previous two.*
4. *Number of revisions.*
5. *Number of project visits.*
6. *Number of times the project was copied by other teachers.*

The first indicator, *number of resources per project,* fits under both role I and role II. The first four indicators measure the internal characteristics of a project, and the latter two measure project quality via its usage rate. The utility of these metrics was uncertain at the very beginning. Some were derived based on the authors' prior work developing a rubric for measuring quality in IA projects [Leary et al. 2009]. The authors had also noted that projects with less than 20 words or less than three resource links had little utility for the general user. Despite these uncertainties, we expected that the clustering algorithm would help reveal the usefulness of each metric. As discussed further below, clustering results indeed showed that not all metrics were useful. The number of project visits excludes authors' visits to their private projects, and external visits referred from other websites. The latter is excluded because links to some (but not all) IA projects are harvested into other digital libraries, including the NSDL, thereby inflating the number of visits to harvested projects. To remove this potentially confounding factor, only student visits and visits from IA users were included in this study.

Role III – Navigation. Behaviors in this role include: visiting and navigating through the IA website, browsing and copying other teachers' projects. Six related metrics were:

1. *Number of visit to the IA website.* Most web usage datasets show an underlying zipf (power-law) distribution [Nielson, 1997; Recker and Pitkow, 1996], with a few elements showing very high counts, most showing very low, and a medium number of elements in the middle. As can be seen in Figure 5, teachers' visits to the IA website follow such distribution.
2. *Number of project browses*. A teacher's *project visits* (see *Role II – Project authoring and usage*) refers to the number of visits to this teacher' projects, and *number of project browses* is his/her visits to other people's public projects. A histogram of project browses follows the zipf distribution too (see Figure 6). The above two parameters are used to define user stickiness.

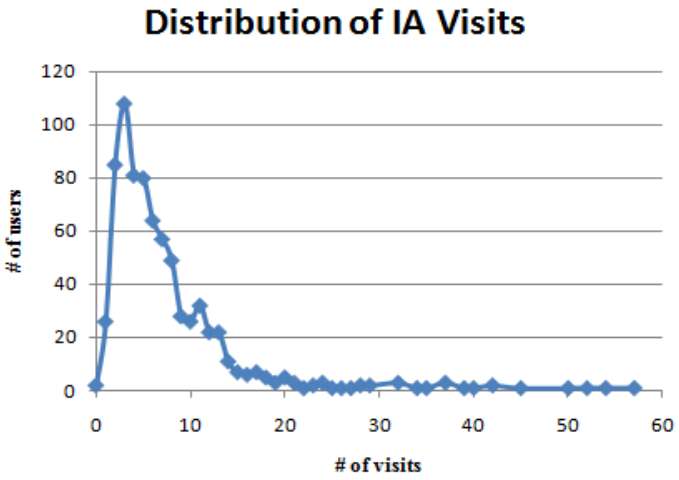

Fig. 5. Distribution of visits to the IA website.

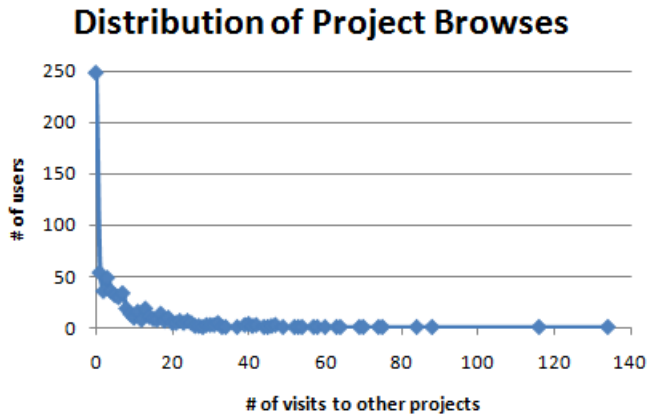

Fig. 6. Distribution of users' project browses over one year.

3. *The percentage of copied projects.*
4. *Visit length.* A session-level parameter that measures the length of a single visit in seconds.
5. *Visit depth.* Another session-level parameter that measures the number of hits / page views in a single visit.
6. *Duration between two visits.* A simple analysis indicates that 66% of the consecutive visits occur within a week. This may happen on several occasions, such as a revisit right after registration, an immediate revision to a new project, or a final check on a project to be released to students. Though it is impossible to deduce each visit's motivation, the duration (in hours) between two consecutive visits is taken as a descriptive feature for a user session.

The third indicator *the percentage of copied projects* fits under both role II and role III.

Some of the user features such as the number of visits and IA project types could be easily obtained, while some needed to be extracted from the ill-formed raw data and transformed into structures suitable for analysis. For example, all three session-level variables (*visit length*, *visit depth*, and *duration between two visits*) were derived from page hits information. Hits were first aggregated into user sessions, and then each session was analyzed to generate its length and depth, and finally compared with the earlier session by the same user to produce the time gap.

In summary, according to teachers' roles played in the IA, three categories, and 19 features were identified, and two of them were double-categorized. Table I summarizes the user feature space, including the categories, the features, their data sources and data preprocessing decisions.

Table I. User Feature Space

category	raw data	transformed data	data type
resources	*resource collection*		
	number of resources	number of resources collected	count
	number of folders	number of resource folders	count
	resource usage		
	project content	resource usage rate	continuous
	project content	*average number of resources per project	continuous
IA projects	*project authoring*		
	number of projects	number of projects	count
	project content	*average number of resources per project	continuous
	project content	average amount of content per project (measured by words)	continuous
	project content	resource / content (measured by words) ratio	continuous
	project history	average number of project revisions	continuous
	project originality	*percentage of copied projects	continuous
	project usage		
	publishing options	percentage of public projects	continuous
	publishing options	percentage of student projects	continuous
	publishing options	percentage of private projects	continuous
	transaction data	average number of project visits	continuous
	project originality	average number of project copied by others	continuous
navigation	*user stickiness*		
	transaction data	number of visits to the IA	count
	transaction data	number of project browses	count
	project originality	*percentage of copied projects	continuous
	navigation profile		
	transaction data	average seconds per visit	continuous
	transaction data	average depth per visit	continuous
	transaction data	average hours since previous visit	continuous

Note: *The *average number of resources per project* fits under both the resources and IA projects categories. The *percentage of copied projects* fits under both the IA projects and navigation categories.

5.3 Data Mining Algorithms: Clustering with Latent Class Analysis

This study used Latent Class Analysis (LCA) [Magidson and Vermunt 2004] to classify registered teacher users into groups. LCA is a model-based cluster analysis technique in that a statistical model (a mixture of probability distributions) is postulated for the population based on a set of sample data. LCA offers several advantages over traditional clustering approaches such as K-means: 1) for each data point, it assigns a probability to the cluster

membership, instead of relying on the distances to biased cluster means; 2) it provides various diagnostics such as common statistics, Log-likelihood (LL), Bayesian information criterion (BIC)[1] and p-value to determine the number of clusters and the significance of variables' effects; 3) it accepts variables of mixed types without the need to standardize or normalize them; and 4) it allows for the inclusion of demographic and other exogenous variables either as active or inactive factors [Magidson and Vermunt, 2004; Vermunt and Magidson, 2002].

Latent Class (LC) Modeling was first introduced by Lazarsfeld and Henry [1968] as a way of discovering latent attitudinal variables from dichotomous survey items. Goodman [1974] supplemented it nicely by extending the analysis to nominal/categorical variables, to deal with the formulation of *K* latent classes through the observation of *n* manifest variables, where both sets (latent and manifest) of variables could be polychotomous. The traditional LCA [Goodman, 1974] assumes that each observation belongs to only one of the *K* latent classes, and all the manifest variables are locally independent of each other (local independence). But the latest LCA model assumes a probability distribution. The basic structure of a LC model for continuous *y* variables is:

$$f(y_i) = \sum_{x=1}^{K} P(x) f(y_i|x),$$

where $f(y_i)$ is the distribution of a random manifest variable y_i, and $P(x)$ is the probability of belonging to latent class *x* without anything else unknown, $f(y_i|x)$ is the distribution of y within latent class *x*. Starting from this, the least restrictive model is obtained by assuming that all $\underline{y}$'s follow class-specific multivariate normal distributions, that is

$$f(y_i|x) = (2\pi)^{-K/2} |\Sigma_x|^{-1/2} \exp\{-\frac{1}{2}(y_i - \mu_x)' \Sigma_x^{-1} (y_i - \mu_x)\}.$$

In this model, each latent class has its own set of means and its own variance-covariance matrix Σ_x, leaving too many parameters to be estimated.

In recent years, LCA has been further developed to include the mixed scale type (nominal, ordinal, continuous, and count), and to allow for both complete and partial local dependence in order to accommodate more research situations [Magidson and Vermunt, 2004; Vermunt and Magidson, 2002]. To reduce the parameters and to restrict an LCA model, one can either set cluster-independent error variances and covariances to zero, or set some off-diagonal elements of the covariance matrix to zero, i.e., local independence between some or all pairs of manifest variables, as shown in the following equation:

$$f(y_i|x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\{-\frac{(y_i - \mu_x)^2}{2\sigma_x^2}\}$$

Finally, after an LC model is constructed, cases are assigned to the latent class that can help achieve the highest $f(y_i)$ [Magidson and Vermunt, 2004].

LCA uses the maximum likelihood method for parameter estimation. It starts with an EM algorithm and then switches to the Newton-Raphson algorithm [Minka, 2002; Ypma, 1995] when it is close enough to the final solution. In this way, the advantages of both algorithms, that is, the stability of EM and the speed of Newton-Raphson when it is close to the optimum solution [Vermunt and Magidson, 2005], are exploited.

[1] Both Log-likelihood and Bayesian Information Criterion are used to assess a model's fitness.

Table II summarizes different LCA models under all possible combination of the settings. Restrictions simplify a model and decrease the number of parameters; on the other hand, a full model is more flexible and has more degree of freedom [Walker, 1940].

Table II. Different Latent Class Analysis Models (for Continuous Variables).

	settings		level of restriction	number of parameters
	local independence	cluster independence		
1	All	yes	****	*
2	partial	yes	***	**
3	no	yes	***	**
4	all	no	**	***
5	partial	no	**	***
6	no	no	*	****

Note: The '*' represents the relative magnitude.

Finally, data post-processing was conducted through interpreting the discovered clusters, making inferences about project quality, and validating the inferences.

6. DATA ANALYSIS

The data from IA teacher users who registered in 2009 were used in this study. From this, one-time visitors and those who have never created any non-private IA projects were excluded. The data from the remaining 757 teachers (out of a total of 1164 registered during that period) were included.

Initially, all 19 features were entered into the latent class analysis as indicators (three in the resources category, nine in the IA project category, five in the navigation category, and two double-categorized). When continuous indicators (14 in this case) are used, the cluster module can be specified ranging from the most unrestricted to the most restricted models. As explained previously, with an unrestricted model, each cluster may have its own variance and a full covariance matrix; though flexible, it results in a large number of parameters to be estimated, which increases as the indicators and the number of clusters k increase. On the other hand, the most restrictive model is obtained if cluster-independent error variance and covariances, and local independence are forced. Though fewer parameters to be estimated, it relies on an unrealistic assumption.

An intermediate model was used to start the data analysis – class dependent variance and covariance and all indicators being locally independent (the 4th model in Table II), and set number of clusters equal three to eight (k = 3 ~ 8). The R^2 of each indicator was reported along with other parameters and values. R^2, also called coefficient of determination, is the proportion of the total variation of scores from the grand mean that is accounted for by the variation between the means of the group [Aron et al., 2009]; in terms of the LCA, it indicates how much variance of each indicator is determined by the difference between latent classes [Statistical Innovations, 2005]. Some indicators had an R^2 less than 10%, meaning little variance on such features was explained by a model. Such

indicators were removed from all models one by one. 10% was rather an arbitrary cut-off point; however, according to the observation, when there were more than ten indicators (user features), not every fraction of the user feature space has a discriminative power in describing group difference. Further, larger k tended to increase the R^2 values. This was because as *k* increased, data inside a cluster were more cohesive and shared less similar characteristics with data from other clusters, and each indicator contributed more in explaining group membership.

In a perfect world, a statistical model can predict all data points' membership with 100% accuracy. However, errors of prediction, also called residual [Cohen, 2001], always exist in a real world situation. The bivariate residual (BVR) in an LCA model is a local measure of model fit by assessing the extent to which the observed association between any pair of indicators are explained by a model [Statistics Innovation, 2005]. As each BVR corresponds to the amount of difference between the observed frequencies in a 2-way cross-tabulation of the indicators contrasted with those expected counts estimated under the corresponding LCA model [Magidson and Vermunt, 2004], a smaller BVR is preferred over a larger one.

The bivariate residuals (BVR) of several pairs of indicators were very large, attesting to the existence of several significant associations (local dependencies) among such pairs of indicators, and our original locally independent model fell somewhat short of correctly clustering users. However, enforcing local dependence between indicators with large BVR did not always increase model fit, which is measured by Bayesian Information Criterion (BIC) in the Latent Class Analysis. BIC is a posterior estimation of model fit based on comparing probabilities that each of the models under consideration is the true model that generated the observed data [Kuha, 2004]. BIC is often used by researchers in selecting a model of the best fit among a class of models with different numbers of parameters [e.g., Claeskens and Hjort, 2008; Nishida and Kawahara, 2005]. A model with a lower BIC value is preferred over a model with a higher value. BIC measure is widely used in Latent Class Analysis' model selection.

In summary, each pair of indicators that had BVR greater than 10 were checked one at a time, and set as locally dependent only when the new model returned a smaller BIC. It is worth noticing that the R^2 and BVR values kept changing in different settings. Removing indicators and forcing local dependence on certain pairs were not separate but iterative steps.

In the end, 13 indicators remained in the analysis. They were: *number of resources*, *number of projects*, *average number of resources per project*, *average amount of content per project (measured by words)*, *average number of project revisions*, *average number of project visits*, *average number of project copies*, *percentage of public projects*, *percentage of student projects*, *percentage of copied projects*, *number of visits to the IA*, *number of project browses*, and *average depth per visits.* Those indicators still encompass all three aspects of teachers' IA activities, one in the resources category, seven in the IA project category, three in the navigation category, and two double-categorized.

An increase in the number of clusters produced a smaller BIC, but when $k = 8$, one of the cluster only had nine teachers (1.2% of the total). A close examination of this cluster of users' IDs revealed that this tiny group was formed by taking a few cases from the smallest cluster in a similar model when $k = 7$, and moreover, it didn't exhibit very different characteristics from the seventh cluster of the 7-cluster model. Therefore, the 8-cluster model overestimated the number of clusters, and the final number of clusters was $k = 7$.

In order to set up the most parsimonious probability model, we made comparisons between models of different degree of restrictions (see Table III). Both Log-likelihood (LL) and BIC were used to assess a model's fitness. As shown in the equation below, opposite to BIC, higher LL indicates a better fit.

$$BIC_{LL} = -2LL + \ln(N) \times M,$$

where N is the sample size, and M is the number of parameters. The second model in Table III (class dependent variance-covariance matrices and local dependences between some but not all indicators) had the smallest BIC value, and was accepted as the final model. In addition, it had the second lowest classification error. Classification error is the probability that a modal assignment rule would fail to correctly classify users into their true clusters respectively.

Table III. Test Results for All 7-cluster Models

Model	Class dependence	Local dependence	LL	BIC	Number of parameters	Classification errors
1	x	x	-21060.644	44852.585	412	0.009
2	x	partial	-20668.951	42537.817	181	0.015
3	x		-21537.081	44852.585	160	0.019
4		x	-26550.538	54042.446	142	0.040
5		partial	-26452.164	53640.188	106	0.043
6			-26648.560	53999.833	106	0.044

7. RESULTS AND INTERPRETATION

7.1 User Clusters

Being a probability model, the Latent Class Analysis produced several probability tables, such as the probability of a case belonging to a certain cluster, and the distribution probabilities of an indicator within a certain cluster. For simplicity's sake, the characteristics of a teacher group were induced based on the indicators' cluster-wise average values. Table IV shows the final clustering results. The clusters are arranged in order of their size, as shown in the first two rows (percent and number). The values underneath are the cluster's mean values for each corresponding indicators.

Cluster 1 (N = 280, 36.8%): isolated islanders. Most of the teachers in this group created projects with more content ($u_{content} = 207$), but only embedded a few resources ($u_{resources\_used} = 3.78$). In particular, only 12 out of the 280 teachers in cluster 1 had projects with more than 10 resource links. Teachers in this cluster seldom browsed and never copied ($u_{percentage\_copy\_projects} = 0$) other teachers' projects. Moreover their own projects were rarely visited and never copied ($u_{project\_copies} = 0$) by others. If the IA is viewed as a social community, its teacher users sharing content with one another, then teachers in cluster 1 are identified as isolated islanders in this IA community.

Cluster 2 (N = 103, 13.7%): lukewarm teachers. Teachers in this group did not view many other projects ($u_{project\_browses} = 0.90$). Though cluster 2 was the most productive group ($u_{number\_of\_projects} = 5.27$), most of their projects were characterized by little content, few resource links, and rare revisions. Teachers in this group always made their projects available to their students ($u_{percentage\_student\_projects} = 0.99$) and were also willing to share them with the public audience ($u_{percentage\_public\_projects} = 0.88$), however, the IA community did not appear to value them much, as they were rarely copied ($u_{project\_copies} = 0.02$). As such, teachers in cluster 2 were labeled as lukewarm teachers.

Table IV. Latent Class Analysis Result.

			cluster 1	**cluster 2**	**cluster 3**	**cluster 4**	**cluster 5**	**cluster 6**	**cluster 7**
		%	36.8	13.7	12.3	12.0	11.1	10.4	3.8
		N	280	103	93	91	83	78	29
indicators	range	mean							
Resource category									
number of resources	0~299	12.57	6.22	12.65	9.53	4.83	12.23	22.40	81.75
Project category									
number of projects	0~10	2.15	1.24	5.27	1.58	0.00	3.57	2.06	4.51
*avg. num. of res. per project	0~44	3.89	3.78	2.02	4.57	0.00	2.94	9.37	9.35
avg. amount of content	0~2843	151.93	206.95	20.58	166.42	0.00	67.26	371.90	168.88
avg. num. of project revisions	0~28	5.62	3.54	0.64	2.81	0.00	1.64	6.08	3.04
avg. num. of project visits	0~272	2.67	1.13	4.36	3.57	0.00	3.84	16.32	53.68
avg. num. of project copied	0~6.33	0.075	0.00	0.02	0.50	0.00	0.00	0.00	0.28
percentage of public projects	0~1	0.55	0.61	0.88	0.59	0.00	0.69	0.43	0.50
percentage of student projects	0~1	0.61	0.66	0.99	0.48	0.00	0.77	0.65	0.59
*percentage of copied projects	0~1	0.14	0.00	0.00	0.46	0.18	0.16	0.25	0.54
Navigation category									
num. of visits to the IA	0~57	7.45	5.37	8.23	5.66	4.22	8.61	11.19	26.93
number of project browses	0~134	8.36	2.76	0.90	6.99	3.79	30.91	10.66	35.98
average depth per visit	2.5~23	35.69	28.96	35.14	39.24	24.94	57.66	45.66	45.36

Notes:

a). **average number of resource links per* project belong to both the resource category and the IA project category; *percentage of copied projects* belongs to both the IA project category and the navigation category.

b). *project visits* measures the number of times a certain IA project has been visited by anyone except by the author, and *project browses* measures the number of peer projects a teacher has visited.

c). *project copies* measures the number of times a project has been copied by anyone except by the author, and *percentage of copy projects* measures the ratio of non-original projects among this teacher's entire collection of IA projects.

d). Since private projects are generally tentative tryouts, they are ignored when measuring the average quality of a teacher's projects. Related indicators are *number of projects*, *average number of resources per project, average amount of content per project, average number of project revisions, average number of project visits,* and *average number of project copies*.

e). Local dependence is set between the following pairs of indicators: *percentage of student projects* and *percentage of public projects*, *percentage of student projects* and *average number of resource links per project*, *percentage of student projects* and *average amount of content per project*.

Cluster 3 (N = 93, 12.3%): goal-oriented brokers. Though teachers in this group did not visit the IA a lot, they tended to borrow ideas from other users' projects. In particular, 46% of their projects were adapted from others. Maybe by viewing and digesting peer projects, they had a better sense of a project's quality. Their projects were relatively verbose ($u_{content}$ = 166.42) and used a fair amount of resources ($u_{resources\_used}$ = 4.57). Perhaps because they were not often listed on the first page of returned results in search and browse, their projects were not visited a lot by other teachers ($u_{project\_visits}$ = 3.57). However, 38.6% of them had been copied and adapted by others, suggesting their projects were well received. Group 3 were not the stickiest users judging from their visit frequency (u_{visits} = 5.66). Nevertheless, they made best of each visit, consuming quality projects and producing valued work in return. Those goal-oriented teachers were therefore considered brokers that knit the IA community together.

Cluster 4 (N = 91, 12.0%): window shoppers. This group of teachers had never contributed to the IA community because they only created a few private projects, perhaps just for practice or fun. Not surprisingly, they were rare visitors compared with other groups. Recall that since all teachers without any project authoring activity or repeated visits were excluded from this study, members in cluster 4 should not be considered the least active IA users. They browsed others projects, but chose not to make their own projects visible to the public ($u_{percentage\_public\_projects} = 0$), not even to their students ($u_{percentage\_student\_projects} = 0$). Considering their lurking characteristics, they represented the window shoppers in this community.

Cluster 5 (N = 83, 11.1%): beneficiaries. Like cluster 2, teachers in this group were willing to share their work with the public, but their projects were seldom visited ($u_{project\_visits} = 3.84$) or copied by their peers in return ($u_{project\_copies} = 0$). Unlike cluster 2, this group was more active, spending a lot of time searching for and browsing existing projects ($u_{depth} = 57.61$, $u_{project\_browses} = 30.91$). They produced more in-depth work than cluster 2, characterized by longer content, more resource links, and more revisions. It appears that teachers in this group had learned a few things from the peers but were not able to produce quality projects to contribute back to the community, and thus are considered as consumers and beneficiaries at this stage.

Cluster 6 (N = 78, 10.4%): classroom practitioners. Judging from the number of embedded resource links ($u_{resources\_used} = 9.37$) and length of content ($u_{content} = 371.90$), the teachers in this group appeared to have put in lots of efforts into authoring projects. This group's projects received the highest number of visits ($u_{project\_visits} = 16.32$). However, most of the hits came from their student accounts ($u_{student\_visits} = 23.70$), and only a few were from other teachers ($u_{peer\_visits} = 1.93$). Similar to cluster 3, it is conjectured that their projects were deeply buried in the list of projects of similar topics returned by the IA search engine, and thus not easily discoverable. But even when occasionally their projects were visited by other teachers, they had never been copied. The text length of the projects suggests these projects were tailored for a specific context and group of students, and thus not easily adaptable. Given the fact that teachers in this group appeared to have designed their projects to meet their very specific instructional needs, they are labeled as classroom practitioners.

Cluster 7 (N = 29, 3.9%): dedicated sticky users. Very similar to cluster 3, teachers in this group serve as brokers: they both consumed others' work by copying project ($u_{percentage\_copy\_projects} = 0.54$) and contributed back to this community ($u_{project\_copies} = 0.28$). They did not appear to be as goal-driven as cluster 3, as teachers in this group reported unusually high visits ($u_{visits} = 26.93$), dedicated enormous time in viewing peer projects ($u_{project\_browses} = 35.98$) and collecting resources ($u_{resources\_collected} = 81.96$) though the majority was not utilized in project authoring ($u_{resource\_usage\_rate} = 11\%$). In sum, this group exhibited two characteristics: dedicated to this community, and stickiest behaviors, and are therefore labeled accordingly.

7.2 Factors Influencing Project Quality

In order to extend the previous analyses and to specifically focus on teachers that create IA projects that were copied and adapted by others users, three particular clusters – cluster 2, 3, and 7 – were put into closer examination. These teachers represented 29.8% of the studied users, or 225 people in total. Projects created by teachers in cluster 3 had the highest probability of being copied and those in cluster 2 had the least chance, and those in cluster 7 were in

between. We examined these three groups of teachers, seeking to understand whether there was any teacher behavior that might help increase the chance of creating valued projects. Here, the assumption is that a project that was copied and adapted by other teachers was valued, and hence a quality project.

Figure 7 plots teachers ($N_{cluster2}$ = 103, $N_{cluster3}$ = 93, $N_{cluster7}$ = 29) along two dimensions: the average content length per project, and average number of resources per project. Since the data were skewed, a log transformation was used to make the data points more evenly distributed on the plot. The same procedure was applied to generate Figure 8 as well. If Figure 7 is segmented into four even tiles, 95% teachers in cluster 2 fall into the lower left tile, while 80% of the teachers in cluster 3 and 7 belong to the upper two tiles, and a few on the lower right. This indicates that, in general, projects from teachers in cluster 3 and 7's exceeded those of teachers in cluster 2 either in length (upper left tile), or in the number of embedded resource links (lower right tile), or both (upper right tile). Examining Figure 8, teachers in cluster 7 have gathered a much larger pool of resources than the other two groups, which presumably made it easier for them to choose the appropriate web resources to accomplish their instructional objectives.

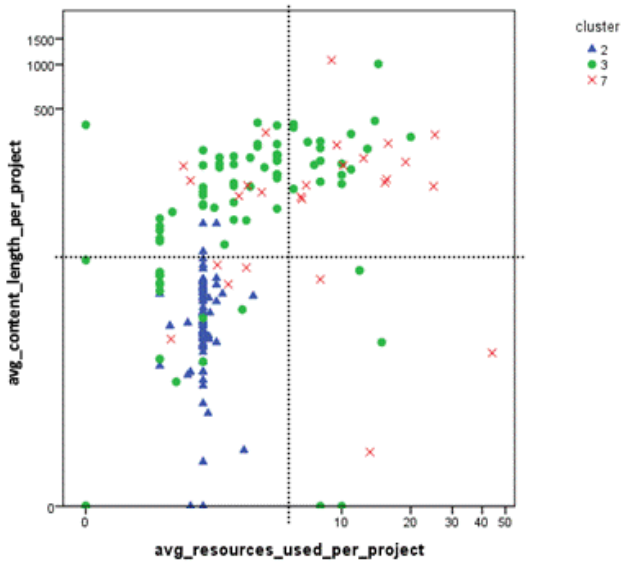

Fig. 7. A plot of teachers in cluster 2, 3, and 7 in terms of project content length and number of resources per project (log transformed).

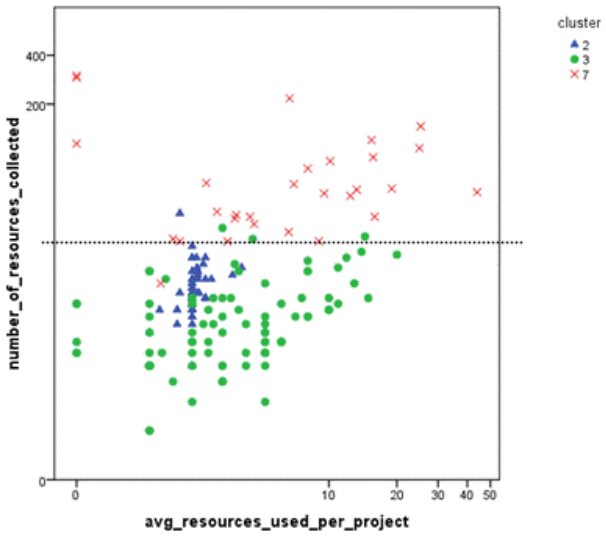

Fig. 8. A plot of teachers in cluster 2, 3, and 7 in terms of number of resources collected and number of resources used per project (log transformed).

Given the fact that projects in clusters 3 and 7's were frequently viewed and copied, and surpassed those in cluster 3 in length and in number of resource links, it can be surmised that content-rich and resource-rich are two essential characteristics of high quality IA project.

The previous clustering analysis suggests that teachers in cluster 2 have never copied a project, and have conducted less project browsing activities than the other two groups (see Table IV). To conduct a fine-grained comparison of the browsing activity, the number of project browses was segmented into four levels using the following procedure. The teachers with zero project browse were first singled out and assigned to the lowest level (the far left), and then the mean and standard deviation for the remaining were calculated after applying a log transformation (to reduce skewness). Finally, the remaining teachers were categorized into three levels – one standard deviation below the mean, one standard deviation above the mean, and those in the middle. Figure 9 plots each cluster's distribution of teachers falling into each level. Judging from the figure, teachers in cluster 7 indeed have viewed more projects than the other two groups, with 55% of them falling into the right end of this distribution, and less than 10% with no or small amount of project browses. On the other hand, more than 75% of the teachers in cluster 2 have never visited other teachers' projects, and none of them was one or more standard deviations above the mean. This analysis provides further evidence that cluster 2 represented the lukewarm teacher group, while cluster 7 represented the sticky one in terms of the magnitude of project browses. It also suggests that engaging in browsing behavior seems to be a precursor to creating valued IA projects.

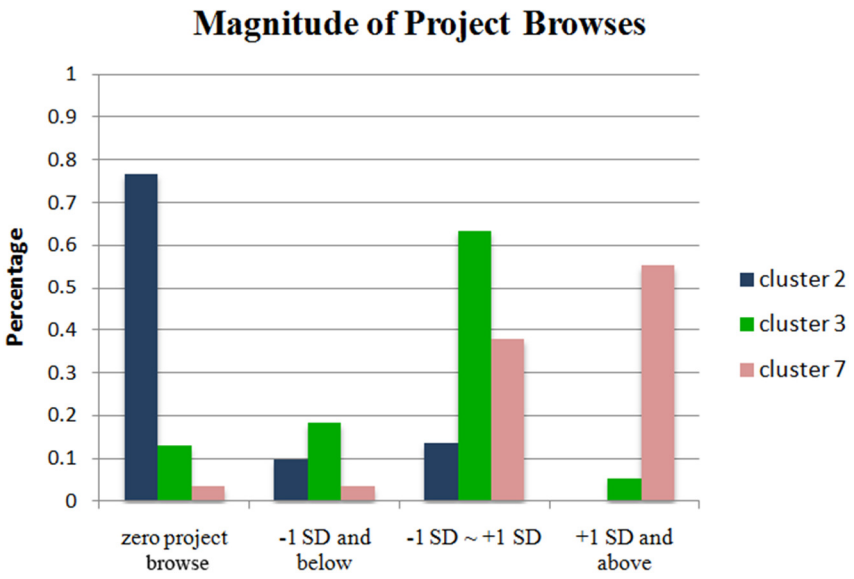

Fig. 9. A plot of teachers in cluster 2, 3, and 7 in terms of the magnitude of project browses.

Cluster 3 and cluster 7, representing approximately 16% of the community, exemplify the ethos of reciprocal contributions and hence can be seen as the backbones of the IA community.

Two inferences have been made through the visual examination and comparison between clusters 2, 3, and 7: 1) that project features (text, and resource links) and 2) teacher's browsing behavior can affect the outcome of IA projects. Statistical analyses were used to determine the validity of our inferences.

Claim 1: On the project-level, content-richness and resource-richness are two essential characteristics of high quality IA projects.

Since a teacher could publish more than one project, and claim 1 is about the IA project's properties, all public projects (N = 1280) created by teachers (instead of the teachers themselves) were selected as the sample dataset. Those projects were assigned to group 1 if copied by other teachers and to group 2 if otherwise. Projects' *number of resource links* and the *content length (measured by words)* were set as the dependent variables to be compared. The descriptive statistics are listed in Table V.

Table V. Descriptive Statistics of the Project-level Dataset in Supporting Claim 1

	N	Variable	Min	Max	Mean	Std. dev
Group 1	74	Content length	0	1071	138.59	162.19
		Resource links	0	50	5.51	6.83
Group 2	1206	Content length	0	2843	108.08	236.03
		Resource links	0	55	3.44	4.57

Note: group 1 = have been copied; group 2 = have not been copied.

Since both measures were positively skewed, and the data was not evenly distributed between the two groups, the assumptions of the parametric independent sample t-test were violated. In this case, the Mann-Whitney U Test was adopted as an alternative.

The Mann-Whitney U Test indicated that both *content length* (U = 32393, p < 0.00) and *number of resource links* (U = 34466, p < 0.00) had significant influence on the project's popularity, which was measured by whether a project had been copied by other teachers. This result helps confirm the claim that text-richness and resource-richness are two essential characteristics of high quality IA project.

Claim 2: On the teacher-level, engaging in browsing behaviors seems to be a precursor to creating valued IA projects.

All teachers who have public projects (N = 503) were selected as the sample dataset. The teachers who have projects copied by others were assigned to group 1, to group 2 if otherwise. Teachers' browsing behavior (the number of peer projects browsed) was set as the dependent variables. The descriptive statistics are listed in Table VI. The Mann-Whitney U Test indicated that there was no significant difference between the two groups (U = 12495, p = .98). Therefore, this result failed to associate the creation of valued IA projects with teachers' browsing behavior.

Table VI. *Descriptive Statistics of the Teacher-level Dataset in Supporting Claim 2*

	N	Min	Max	Mean	Std. dev
Group 1	74	0	1071	9.02	12.21
Group 2	447	0	134	9.29	14.63

Note: group 1 = have been copied; group 2 = have not been copied.

The statistical analyses supported the first inference drawn from the clustering results, but rejected the other. Therefore, we conclude that teachers should be encouraged to embed more useful resources and add elaborations in order to create higher quality projects. On the other hand, though copying one another's projects reflects teachers' collaborative activities in the community [Xu and Recker, 2010], there is no converging evidence regarding whether such activities indeed help improve teachers' abilities in identifying high quality IA projects and the knowledge of how to build them.

8. CONCLUSION AND DISCUSSION

This article described the KDD process and its application to the field of educational data mining in the context of a digital library service, the Instructional Architect. In particular, we explored a certain type of data mining problem, clustering, and used a statistical model, latent class analysis, to group the IA teacher users according to their diverse online behaviors. The LCA successfully helped us to identify seven different types of users, ranging from window shoppers and lukewarm users to the most dedicated users, and to distinguish the isolated users from the key brokers of the community. This section discusses implications of this study for the design of systems such as the IA, and future EDM studies.

8.1 Implication for Design of the IA

Results from this study showed that in the IA community, consumers outnumber contributors. Of the 757 users under study, most were isolated islanders, window shoppers, lukewarm users, or beneficiaries. Only 122 users were key brokers or sticky users, comprising 16.1% of the selected cases, and 10.5% of the users in the study. While "free

riding" is a common phenomenon in peer production communities [López-Pintado, 2008; Wasko et al. 2009], sustainability of these communities can be enhanced by supporting the critical mass of active members [Wasko et al. 2009].

Within the IA, to continue to sustain and even grow the number of critical contributors, its system design might consider incentive schemes to encourage more knowledge sharing. For example, the interface could display the number of times a project has been viewed or copied, in order to publically acknowledge popular authors. Popular projects could be included in the showcase section of the IA. The search functionality could include a recommender engine in order to suggest similar projects and encourage sharing. Finally, as an incentive for peer contribution, the IA system could support user commenting and rating to encourage iterative improvements to projects.

Although some users readily share their projects with the public, sometimes the project is of dubious quality. For example, teachers in cluster 2 (*Lukewarm teachers*) made the majority of their projects public, however, their projects appear to be of low quality, as they were not viewed by many and never copied. A vetting process could be implemented in the IA system, providing suggestions to project authors on how to make improvements to their projects before making them public. Such vetting could include a checklist identifying important elements in a high quality project.

Our analysis of project quality revealed that it is very important to have a pool of useful resources as the foundation of a high quality project. The current IA collaboration scheme only allows users to share public projects; they cannot share their resource collection unless used within public projects. As such, there is no way to copy an individual resource collection over without copying the entire project itself. Given the importance of resources, the IA interface could be modified to allow users to share their resource collections.

Finally, and more futuristically, LCA results could be used to offer personalized messages to members of each cluster. For example, teachers in cluster 7 (*Dedicated sticky users*) could be prompted to make their projects public. These projects could also be automatically featured in the project showcase section of the IA. Teachers in Cluster 1 *(Isolated islanders)* could be reminded of the browse and copy features of the IA, functionality they seldom use.

During the inference of clusters' characteristics, we found out that some seemingly good projects (for instance, projects created by users in cluster 3 and cluster 6) have not caught due attentions because they were not at the top among the returned results, and not easy to be discovered. The IA projects browsing interface could be improved to address this problem. For example, allow users to filter search results by content length and number of resource links; and rank projects in order of their popularities, etc.

8.2 Implications for Educational Data Mining

Results from this study suggest several implications for future EDM studies. First, the development of a user feature space should start by defining the behavioral / navigational categories supported by the system's functionality, and then enumerating the user features that could fit into each category. When there is redundancy in the initial user feature space, the removal of redundant features relies on the inherent nature of the clustering algorithm, but should at the same time remain the backbone of the original categories.

Second, most current web-based educational applications have an inherent built-in social structure. Explicit and implicit linkages between users (e.g., the browsing and copying behavior of users in the IA) are potentially fruitful areas for exploration.

Finally, while latent class analysis has been widely used in psychology and medical studies, it has not exploited to its full potential in educational data mining. This study provides an example on how to use LCA in addressing educational problems, and has promise for future research.

8.3 Limitations and Potentials

While we believe our approach has utility, it still has plenty of room for improvement. First, the current study aggregated the project-level information to the user level, using average values to represent a user's project-related characteristics. Though such generalization provides an overarching picture of a user, it, however, glosses over the details of individual IA projects. As such, analyzing individual project-related features could prove to be fruitful. For example, it could help better understand individual teachers' project authoring habits, their likes and dislikes about IA projects, and perhaps most importantly, helps us advise IA users on how to create higher quality projects. Based on the above reasoning, conducting an LCA analysis on a deeper level – moving from grouping users to understanding individual projects' qualities – could be an important direction for future research.

Second, though latent class analysis is alleged to outperform k-means, no competing clustering algorithm has been implemented to justify the choice of algorithm. In addition, as discussed in the literature review, other data mining methods could be incorporated with clustering algorithms to achieve a better grouping effect. At this stage, our study is limited to using a statistical latent class model to analyze IA usage, but in the future, other methods such as association rule mining and sequential pattern mining could be utilized as well.

Finally, the third stage of KDD, evaluation and interpretation, could be conducted in a more comprehensive fashion. For example, our previous work showed that greater use occurs in geographical areas where teacher professional development workshops using the IA have been conducted [Khoo et al. 2008; Xu et al. 2010]. This suggests that workshop participants have a higher chance of becoming sticky users. Therefore, teachers with workshop history can be singled out for analysis, and their distribution among clusters is expected to be different. Finally, we plan to triangulate our data mining results with other more conventional data sources, including data from users' registration profiles and surveys.

Despite the current challenges, the field of educational data mining is making progress towards standardizing its procedures for tackling educational problems. As online learning environments continue to generate a data deluge of massive and longitudinal datasets, and data mining algorithms continue to evolve, opportunities to explore this rich territory are flourishing.

REFERENCES

AIVAZIAN, B. I., GEARY, E., KHOO, M., SUMNER, T., AND IRETON, S. 2003. Serving K-12 education with DWEL. *Knowledge Quest 31*, 21-22.

ARON, A., ARON, E. N., AND COUPS, E. J. 2009. *Statistics for psychology (5th ed.)*. Pearson Education, Upper Saddle River, New Jersey.

ASUNKA, S., CHAE, H. S., HUGHES, B., AND NATRIELLO, G. 2008. Understanding academic information seeking habits through analysis of web server log files: The case of the teachers college library website. The *Journal of Academic Librarianship 35*, 33–45.

BAKER, L. J. 2009. Science teachers' use of online resources and the digital library for earth system education. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, Austin, Texas, USA, 1-10.

BAKER, R. S. J. D., AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining 1*, 3-17.

BORGMAN, C. L. 1999. What are digital libraries? Competing visions. *Information Processing and Management 35*, 227-243.

CARLSON, B., AND REIDY, S. 2004. Effective access: Teachers' use of digital resources (research in progress). *OCLC Systems & Services 20,* 65 – 70.

CHEN, H., AND CHAU, M. 2004. Web mining: Machine learning for web applications. In *Annual Review of Information Science and Technology 38*, C. BLAISE, Eds. Information Today, Inc, Medford, NJ, 289-329.

CHOUDHURY, S., HOBBS, B., AND LORIE, M. 2002. A framework for evaluating digital library services. *D-Lib Magazine 8*.

CLAESKENS, G., AND HJORT, N. L. 2008. *Model selection and model averaging*. Cambridge University Press, New York, NY.

COHEN, B. H. 2001. *Explaining psychological statistics.* John Wiley & Sons, Inc, New York, NY.

COOLEY, R., MOBASHER, B., AND SRIVASTAVA, J. 1997. Web mining: Information and pattern discovery on the World Wide Web. Paper presented at *the 9th IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA.

DURFEE, A. SCHNEBERGER, S. AND AMOROSO, D. L. 2007. Evaluating students computer-based learning using a visual data mining approach. *Journal of Informatics Education Research 9,* 1-28.

GOODMAN, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*, 215-231.

GRIMES, S. 2007. The grand challenge for text mining. Retrieved from, http://www.informationweek.com/news/software/bi/showArticle.jhtml?articleID=228900459.

HAN, J., AND KAMBER, M. 2006. *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann Publishers, San Francisco, CA.

HÜBSCHER, R., PUNTAMBEKAR, S., AND NYE, A. H. 2007. Domain specific interactive data mining. In *Proceedings of Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling*, Corfu, Greece, 81-90.

KHOO, M. 2006. *NSDL user survey 2006.* Retrieved from, http://www.ischool.drexel.edu/faculty/mkhoo/docs/nsdl_06_user_survey.pdf.

KHOO, M., PAGANO, J., WASHINGTON, A. L., RECKER, M., PALMER, B., AND DONAHUE, R. A. 2008. Using web metrics to analyze digital libraries. In *Proceedings of the Joint Conference on Digital Libraries*, New York, 375-384.

KOUTRI, M., AVOURIS, N., AND DASKALAKI, S. 2004. A survey on web usage mining techniques for web-based adaptive hypermedia systems. In *Adaptable and Adaptive Hypermedia Systems,* S. Y. CHEN, AND G. D. MAGOULAS, Eds. IRM Press, Hershey, PA, 125-149.

KRIEGEL, H. P., BORGWARDT, K. M., KRÖGER, P., PRYAKHIN, A., SCHUBERT, M., AND ZIMEK, A. 2006. Future trends in data mining. *Data Mining and Knowledge Discovery 15*, 87-97.

KUHA, J. 2004. AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33*, 188-229.

LAGOZE, C., VAN DE SOMPEL, H., NELSON, M., AND WARNER, S. 2002. The open archives initiative protocol for metadata harvesting. Retrieved from http://www.openarchives.org/OAI/openarchivesprotocol.html

LAZARSFELD, P. F., AND HENRY, N. W. 1968. *Latent structure analysis*. Boston: Houghton Mifflin.

LEARY, H., GIERSCH, S., WALKER, A., AND RECKER, M. 2009. Developing a review rubric for learning resources in digital libraries. ITLS Faculty Publications.

LEE, C. 2007. Diagnostic, predictive and compositional modeling with data mining in integrated learning environments. *Computers & Education 49*, 562-580.

LÓPEZ-PINTADO, D. 2008. The spread of free-riding behavior in a social network. *Eastern Economic Journal 34*, 464-479.

MAGIDSON, J., & VERMUNT, J., K. 2004. Latent class models. In *The Sage Handbook of Quantitative Methodology for the Social Sciences,* D. KAPLAN, Eds. Sage Publications, Thousand Oaks, CA, 175-198.

MALIK, M., AND JAIN, A. K. 2006. Digital library: Link to e-learning. *DRTC - ICT Conference on Digital Learning Environment*. Bangalore, India.

MAULL, K. E., SALDIVAR, M. G., AND SUMNER, T. 2010. Online curriculum planning behavior of teachers. In *Proceedings of the 3rd International Conference on Educational Data Mining*, Pittsburgh, PA, 121-130.

Minka, T. P. 2002. *Beyond Newton's method.* Retrieved from http://research.microsoft.com/en-us/um/people/minka/papers/minka-newton.pdf.

NIELSON, J. 1997. Zipf curves and website popularity. Retrieved from http://www.useit.com/alertbox/zipf.html

NISHIDA, M., AND KAWAHARA, T. 2005. Speaker model selection based on the Bayesian information criterion applied to unsupervised speaker indexing. *IEEE Transactions on Speech and Audio Processing 13*, 583-592.

PAHL, C., AND DONNELLAN, D. 2002. Data mining technology for the evaluation of web-based teaching and learning systems. Paper presented at *the E-Learn 2002 World Conference on E-Learning in Corporate, Goverment, Healthcare, & Higher Education*, Montreal, Quebec, Canada.

PATTUELLI, M. C. 2008. Teachers' perspectives and contextual dimensions to guide the design of N.C. history learning objects and ontology. *Information Processing and Management 44*, 635–646.

PERRAULT, A. M. 2007. An exploratory study of biology teachers' online information seeking practices. *School Library Media Research 10.*

RECKER, M., AND PITKOW, J. 1996. Predicting document access in large, multimedia repositories. *ACM Transactions on Computer-Human Interaction 3*, 352-375.

RECKER, M. 2006. Perspectives on teachers as digital library users: Consumers, contributors, and designers. *D-Lib Magazine 12*. Retrieved from http://www.dlib.org/dlib/september06/recker/09recker.html.

RECKER, M., WALKER, A., GIERSCH, S., MAO, X., HALIORIS, S., PALMER, B., JOHNSON, D., LEARY, H., AND ROBERTSHAW, M. B. 2007. A Study of teachers' use of online learning resources to design classroom activities. *New Review of Hypermedia and Multimedia 13*, 117-134.

ROMERO, C. AND VENTURA, S. 2007. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications 33*, 135-146.

SHREEVES, S. L., AND KIRKHAM, C. M. 2004. Experiences of educators using a portal of aggregated metadata. *Journal of Digital Information 5*.

STATISTICAL INNOVATIONS. 2005. Tutorial 1: Using Latent GOLD® 4.5 to estimate LC cluster models. Retrieved from http://www.statisticalinnovations.com/products/latentgold_v4.html.

SUMNER, T., AND CCS TEAM. 2010. Customizing science instruction with educational digital libraries. In *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries*, Gold Coast, Queensland, Australia, 353-356.

SUMNER, T., KHOO, M., RECKER, M., AND MARLINO, M. 2003. Understanding educator perceptions of "quality" in digital libraries. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries,* Houston, Texas, 269-279.

SUMNER, T., AND MARLINO, M. 2004. Digital libraries and educational practice: A case for new models. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tucson, Arizona, 170-178.

TALAVERA, L., AND GAUDIOSO, E. 2004. Mining student data to characterize similar behavior groups in unstructured collaboration spaces. Paper presented at *the Workshop on Artificial Intelligence in CSCL, 16th European Conference on Artificial Intelligence*, Valencia, Spain.

TANNI, M. 2008. Prospective history teachers' information behaviour in lesson planning. *Information Research 13.*

VERMUNT, J., K., AND MAGIDSON, J. 2002. Latent class cluster analysis. In *Applied Latent Class Analysis,* J. HAGENAARS AND A. MCCUTCHEON, Eds. Cambridge University Press, New York, NY, 89-106.

VERMUNT, J., K, AND MAGIDSON, J. 2005. *Technical guide for Latent GOLD 4.0: Basic and advanced.* Statistical Innovations Inc, Belmont, MA.

WALKER, H. M. 1940. Degrees of freedom. *Journal of Educational Psychology 31*, 253-269.

WANG, W., WENG, J., SU, J., AND TSENG, S. 2004. Learning portfolio analysis and mining in SCORM compliant environment. Presented at *the 34th ASEE/IEEE Frontiers in Education Conference*, Savannah, GA.

WASKO, M. M., TEIGLAND, R. AND FARAJ S. 2009. The provision of online public goods: Examining social structure in an electronic network of practice. *Decision Support Systems 47*, 254-265.

WITTEN, I. H., AND FRANK, E. 2005. *Data mining: Practical machine learning tools and techniques* (2nd ed.). Morgan Kaufmann, San Francisco, CA.

XU, B., AND RECKER, M. 2010. Peer production of online learning resources: A social network analysis. Poster presented at *the third Annual Conference on Educational Data Mining*, Pittsburgh, PA.

XU, B., RECKER, M., AND HSI, S. 2010. The data deluge: Opportunities for research in educational digital libraries. In *Internet Issues: Blogging, the Digital Divide and Digital Libraries*, C. M. EVANS, Eds. Nova Science Publishers, Hauppauge, NY.

YPMA. T. J. 1995. Historical development of the Newton-Raphson Method. *SIAM Review 37*, 531-551.

Zia, L. L. 2001. Growing a national learning environments and resources network for science, mathematics, engineering, and technology education. *D-Lib Magazine 7*.