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 in the context of a digital library service called the Instructional Architect (IA.usu.edu). In particular, we investigated a certain type of data mining problem, clustering, and used a statistical model, latent class analysis, to group the IA 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.

Keywords: Educational Data Mining, Educational Web Mining, Clustering, Latent Class Analysis

---

1. INTRODUCTION

Advances in technology have resulted in major changes in the world of education. Increasingly, education and training are delivered beyond the classroom environment, and digital libraries and their associated services have made major contribution to these changes inside and outside formal education. The level of interest regarding digital libraries has been growing [Choudhury et al. 2002] as a great number of institutions consider them an opportunity for formal and informal learning and knowledge sharing. As digital library services become a fundamental part of e-learning environments and information networks, researchers, as well as stakeholders, need to ensure efforts and resources expended on digital libraries are worthwhile in terms of their impact on users.

Evaluation of digital libraries has adopted diverse approaches and taken on many forms [Choudhury et al. 2002]. In particular, provided with abundant user footprints and data mining strategies, researchers can analyze fine-grained web usage data so as to understand digital library's users and their usage patterns. This article uses a particular digital library service, called the Instructional Architect (IA.usu.edu), as a test bed for investigating how to use Knowledge Discovery and Data Mining (KDD) process in general, and clustering methods in particular, to help identify the diverse digital library user groups and their characteristics.

We first review the generic Knowledge Discovery and Data Mining (KDD) process, followed by a description of several clustering studies in the field of educational web mining. We then provide a brief introduction to the Instructional Architect service. This is followed by descriptions of our clustering strategies, starting from data collection and selection, to data analysis and interpretation. We conclude with thoughts on improving and complementing the clustering strategies, and on the implication of this study.

2. LITERATURE REVIEW

2.1 Educational Web Mining

There is growing interest in data mining and 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" [Educational Data Mining, n.d.].

Increasingly educational learning environments, including educational digital libraries, 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 investigates how to apply data mining to a particular online digital library service, thereby contributing to the field of educational web mining.

2.2 Knowledge Discovery and Data Mining

Web mining by and large follows the standard KDD (Knowledge Discovery and Data Mining) process, entailing: 1) data cleaning and integration, 2) selection and transformation, 3) application of data mining algorithms, 4) evaluation and presentation [Han and Kamber 2002; Witten and Frank 2005]. Often the first two phases are combined and called data preprocessing [Cooley et al. 1997; Romero and Ventura 2007]. These phases are described next.

2.2.1 Phase I – Data Preprocessing. Raw data are generally far from being ready to be ingested by a mining algorithm, as there can be missing entries, irrelevant information, or the need to integrate data from different sources before using them. Thus, as the first step of knowledge discovery, data preprocessing, is very critical in ensuring that the data are in a suitable shape and can produce valid results. This step includes data cleaning, path completion, data integration, data selection, and data transformation, described next.

To a large degree, educational data mining originates from the analysis of transaction logs of student-computer interaction [Baker and Yacef 2009]. Usually, a web server log stores all requests sent from the clients, which means irrelevant information such as company logo, place-holder graphics, or spider hits taint the transaction logs [Cooley et al. 1999]. *Data cleaning*, the most intensive step in data preprocessing, focuses on removing noise and inconsistent data from the data source [Han and Kamber 2002].

Most web browsers cache the pages that have been requested in the past in order to reduce response time, making the web server unaware of repeated page requests [Cooley et al. 1997; Koutri et al. 2004; Sheard et al. 2003; Weischedel and Huizingh 2006]. *Path completion* focuses on recovering missing pieces of transaction logs due to page caching.

Data mining algorithms generally require a homogeneous dataset – data originating from a single source. However, sometimes information from a single source is insufficient for data mining, and there is a need to refer to

different sources to develop a more comprehensive picture of the topic at hand. *Data integration* entails the combination of data from multiple autonomous and heterogeneous resources [Halevy et al. 2006; Han and Kamber 2002; Romero and Ventura 2007]. It is another central step in data preprocessing for knowledge discovery [Kriegel et al. 2006].

Usually not all of the information obtained from the raw dataset is useful for data mining; *data selection* focuses on choosing a set of user-related variables – a feature vector – to represent a particular user activity.

Finally, to further reduce the complexity of the user features, researchers often convert data into forms appropriate for mining. Binning, aggregation, smoothing, generalization, normalization and attribute constructions are the commonly used *data transformation* methods [Han and Kamber 2002].

2.2.2 Phase II – Applying Data Mining Algorithms. From a data source point of view, web mining can be divided into two categories: web usage mining and web content mining [Koutri et al 2004]. Web usage mining aims to discover meaningful usage patterns, and transaction data are usually generated from users' interaction with a system, for example, document references and user visits. Web content mining, on the other hand, aims to extract useful information from the web documents themselves, and the data used are usually the textual content or document metadata.

In general, web mining serves two purposes: description and prediction. Description aims at finding human-interpretable patterns that describe the data; clustering, association rule mining, sequential pattern discovery all belong to this type. Educational data mining, as one of data mining techniques' applications, adopts almost all types of algorithms. 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; classification, regression, anomaly detection algorithms fall into this category.

2.2.3 Phase III – Interpretation and Post-processing. There is no universal standard for evaluating data mining results, and a widely divergent set of results are possible even when using the same dataset after going through different preprocessing procedures and web mining algorithms. In addition, the interpretation of results is highly problem dependent. Just as neither statistical p-values nor effect sizes make sense unless contextualized and accompanied with appropriate explanations, the same applies to data mining results. Discovered patterns are not very useful unless there are mechanisms and tools to help analysts better understand them [Cooley et al. 1997]. Interpretation techniques are drawn from a number of fields such as statistics, data visualization, and usability studies. Though statistical analysis software (e.g., SPSS, LatentGold) and web analytics tools (e.g., Google Analytics) have the ability to visually display the analysis results, it is usually up to the researcher to interpret and present the discovered patterns.

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 related web mining work 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 studies 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 respectively 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 is collected in the form of navigation events, where each event consists of a timestamp, a student name, and a science concept. After preprocessing, K-means and hierarchical clustering algorithms are 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 is 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* is 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.

There are many 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 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 all candidate 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.

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 simple authoring of 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 project pages, called *IA projects* (also referred to as "projects" for simplicity), which can be kept private, or made available to only their students, or to the wider Web. Figure 1 shows an example of a simple IA project created by one of our teacher users. The teacher created the layout and text, along with links to online resources discovered in the NSDL or on the Web.

Fig. 1. An IA project named "Classifying Triangles" created by Mr. Herring.

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: resources management and project management. In the resources management mode, teachers can search for and store links to NSDL resources within the IA context (see Figure 2). In addition, teachers can name and add non-NSDL resource links, or save other people's projects to their own collection too.

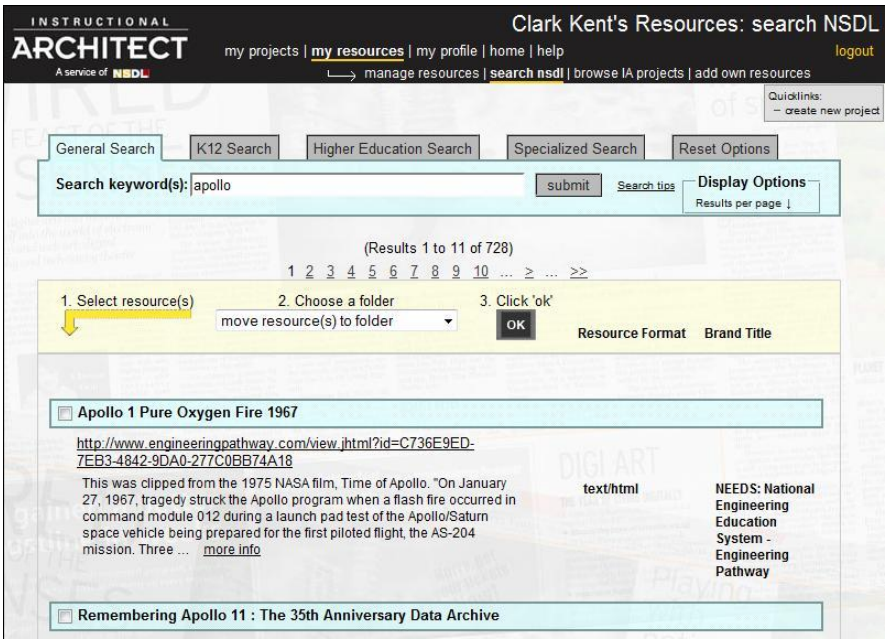

Fig. 2. Searching for NSDL resources inside the IA.

With the IA's user-friendly project authoring interface, teachers only need to enter an IA project's title, overview, and content, and the system can generate a webpage dynamically upon request. The teacher's resource collections are listed on the left, and can be added to the project with a single click (see Figure 3). 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 projects. A project can be marked as public, student-view, or private. Anyone can visit a public project, and only students have exclusive access to their teachers' "student-view" projects, and private projects are only viewable by the owner. All public projects are saved under the Creative Commons' *free to share and free to remix* license. Any 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 of instructional resources and activities.

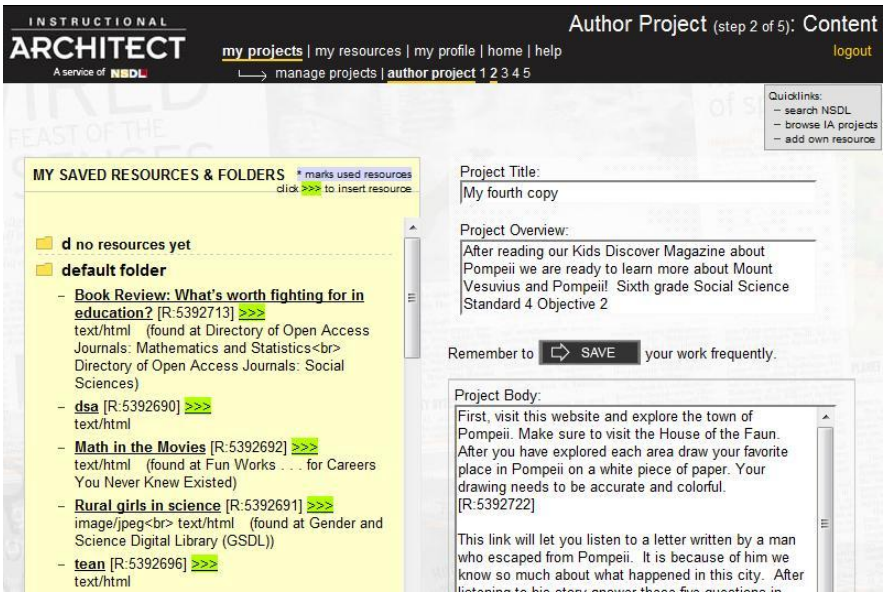

Fig. 3. Project creation interface.

Resources are listed on the left, and a user enters content on the right.

Figure 4 presents the data model for the Instructional Architect. Teachers play a central role in this model, and are therefore the targets of this study.

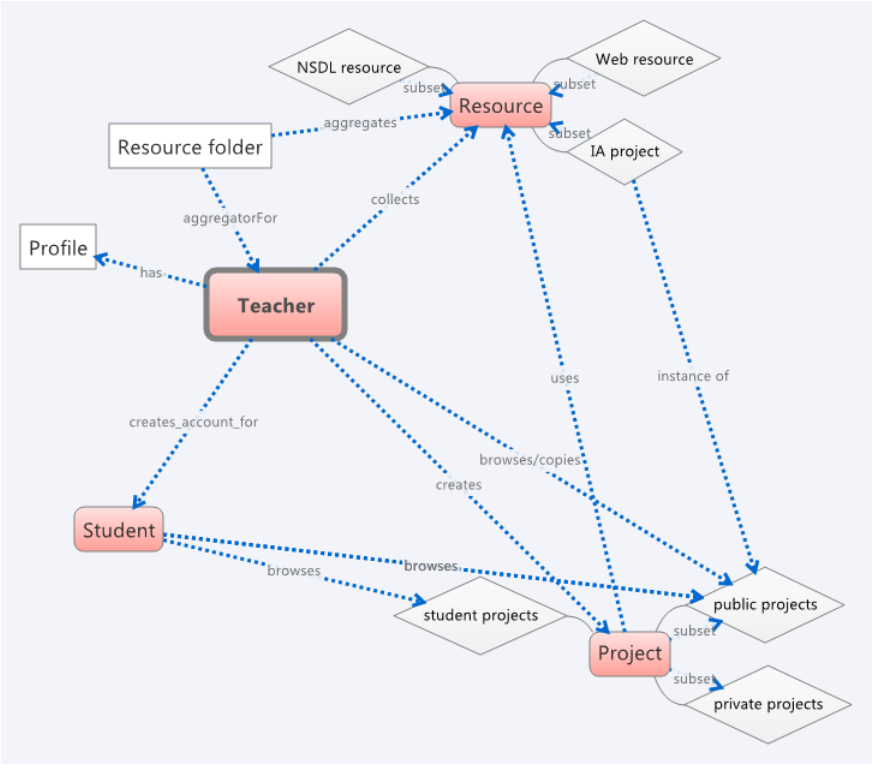

Fig. 4. The Instructional Architect's data model.

3.2 Usage

From 2002 to February 2010, over 4,800 teachers have registered with the IA, more than 9,900 IA projects have been created, and 44,600 external online resources have been added to the database. Since August 2006, public projects have been viewed over 708,300 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 IA serves a wide range of teachers and students, including preschool, K-12, and higher education. In addition, a teacher can conduct a wide range of activities within the IA: searching for and sequencing online resources, browsing and publishing projects, etc. Yet little is known about IA users and their diverse behaviors within the IA. As such, the purpose of this research is to use the IA as a test bed to investigate how to apply web mining methods – clustering approaches in particular – to better understand teachers' use of this digital library service. The following two research questions are addressed in this study:

1. What are the considerations in aggregating and selecting metrics that are able to meaningfully characterize IA teachers and their usage patterns?
2. What usage patterns emerge from mining these data?

5. METHODOLOGY

Web mining the IA datasets 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. The major considerations in using clustering methods are highlighted in this study in response to the lessons learned in the literature review.

5.1 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 purposes. 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 every IA project's past versions, providing a window to examine how the 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, 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.

Though much prior research addresses issues in collecting data from a web server log [e.g., Koutri et al. 2004; Pahl and Donnellan 2002], we still prefer using the IA's database as the primary data source, due to the fact that it 1) provides a more comprehensive picture of teacher activities, 2) makes it easier to identify individual teachers given

the existence of unique user IDs and session IDs, 3) eliminates the records of unwanted requests such as images, and 4) presents better-formed columned information for data cleaning and extraction.

5.2 User Feature Space

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

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

5.2.1 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; embedding resources into projects. Three related metrics are:

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.

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

1. *Number of projects*. Because private projects are inaccessible to anyone but the author, only public and student projects were counted when measuring teachers' productivity using the IA authoring tool and their contributions to this community.
2. *The percentage of each type of projects.* A previous analysis showed that 24% of the teachers only created private projects, while 29% never kept a project private. Therefore, project type reflects a teacher's motivation in creating a project and its target audience; for instance, are there any student-only projects intended 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 or 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 circumvent this problem, we used the following 6 indicators as a proxy to measure the quality of a project.

1. *Number of resources per project.*
2. *Number of words per project (excluding the text in resource links).*
3. *Ratio between the previous two.*
4. *Number of revisions.*
5. *Number of project hits.*
6. *Number of times the project was copied by other teachers.*

The first four measured the internal characteristics of a project, and the latter two measured the quality through implicit external ratings.

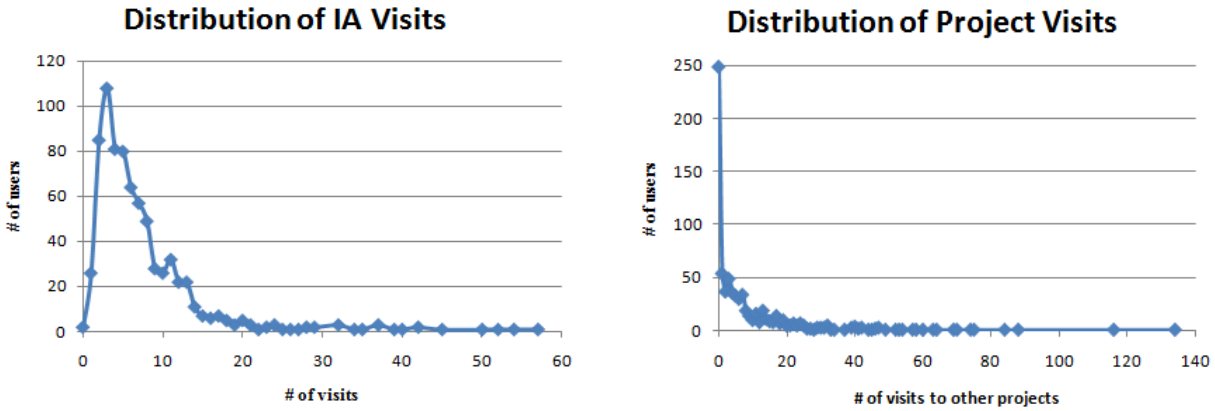

Fig. 5. a) distribution of visits to the IA website; b) distribution of visits to IA projects over 1 year.

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

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 5a, teachers' visits to the IA website follow this distribution.
2. *Number of project visits.* Here we defined *project hits* (see 5.2.2) as the number of visits to each of the teacher's projects, and *project visits* as that teacher's visits to other people's public projects. A histogram of project visits follows the zipf distribution too (see Figure 5b). We used the above two parameters to define *user stickiness*.
3. *Visit length.* A session-level parameter that measured the length of a single visit in seconds.
4. *Visit depth.* Another session-level parameter that measured the number of hits / page views in a single visit.
5. *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 we cannot deduce each visit's motivation, we included the duration (in hours) between two consecutive visits as a descriptive feature for a user session.

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.

Table I summarizes the user feature space, including the categories, the 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
IA Projects		*Project Authoring*	
	Number of projects	Number of projects	Count
	Project content	Average number of resources per project	Continuous
	Project content	Average number of words per project	Continuous
	Project history	Resource / words ratio	Continuous
	Project content	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
	Project originality	Average number of project hits	Continuous
	Transaction data	Average number of project copies	Continuous
Navigation		*User Stickiness*	
	Transaction data	Number of visits to the IA	Count
		Navigation Profile	
	Transaction data	Number of project visits	Count
	Transaction data	Average seconds per visit	Continuous
	Transaction data	Average depth per visit	Continuous
	Transaction data	Average hours since previous visit	Continuous

5.3 Clustering

This study used Latent Class Analysis (LCA) [Magidson and Vermunt 2002] 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 a 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 them; and 4) it allows for the inclusion of demographics and other exogenous variables either as active or inactive factors [Magidson and Vermunt 2004; Vermunt and Magidson 2002]

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 IA projects were excluded. The data from the remaining 757 teachers (out a total of 1164 registered during that period) were included.

Initially, all 19 variables (3 in the resources category, 11 in the IA project category, and 5 in the navigation category) were entered into the latent class analysis as indicators. When continuous indicators are used, the cluster module can be specified ranging from the most unrestricted to the most restricted models. 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, if we assume all clusters share the same variance and all covariances equal zero (locally independent), we get the most restrictive model, requiring less parameters but relying on an unrealistic assumption.

We started with an intermediate model – class dependent variance and off-diagonal elements of the covariance matrix being zero, and set number of clusters equal three to eight ($k = 3 \sim 8$). Some indicators had an R^2 less than 0.1, meaning little variance on such features was explained by a model. Thus, indicators with less discriminative power were removed from all models one by one. We observed that larger k tended to increase the R^2 values. This is because as k increases, data inside a cluster are more cohesive and share less similar characteristics with data from other clusters, which leads to each indicator contributing more in explaining group membership.

The bivariate residuals (BVR) of several pairs of indicators were very large, suggesting that a model of diagonal covariance matrix fell somewhat short of explaining the association between the variables. We checked each pair of indicators with BVR >=10 one at a time, and set them as locally dependent only when the new model returned a smaller Bayesian Information Criterion (BIC). BIC is often used by researchers to determine the model of best fit [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. It is worth noticing that because 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 (see Table III).

An increase in the number of clusters produced a smaller BIC, but when $k = 8$, one of the cluster only had 9 teachers (1.2% of the total). A close examination of indicators' mean values and cluster size revealed that this tiny group was formed by taking a few cases from two of the larger clusters in a similar model when $k = 7$. Moreover, this group didn't exhibit very distinctive characteristics. Therefore, we believed that this model overestimated the number of clusters, and we set the final number of cluster as $k = 7$.

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

In order to set up the most parsimonious probability model, we made comparisons between models of different degree of restrictions (see Table II). Model 2 (class dependent variance-covariance matrices and local dependences between some but not all variables) had the smallest BIC value, and was used as the final model.

Table II. 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

6. RESULTS AND INTERPRETATION

6.1 Results

Table III shows the final clustering results. The first two rows show the size of each cluster (percent and number). The values under each cluster are the cluster's mean values for each corresponding indicators.

6.2 Interpretation

6.2.1 User Clusters. Based on each variable's distribution and its average value, we induced the characteristics of each teacher group. In addition, we extracted more details from the database, and associated them with the users of each cluster, thereby providing a more comprehensive interpretation for teacher clusters. The resulting clusters and their interpretation are described next.

Cluster 1 (36.8%): Isolated islanders.

Most of the teachers in this group create projects with a higher number of words ($u_{\text{words}} = 207$), but only embed a few resources ($u_{\text{resources\_used}} = 3.78$). In particular, only 12 out of the 280 teachers in cluster 1 have projects with more than 10 resource links. Teachers in this cluster seldom browse and never copy ($u_{\text{percentage\_copy\_projects}} = 0$) other teachers' projects. Moreover their own projects are rarely visited and never copied ($u_{\text{project\_copies}} = 0$) by others. If the IA is viewed as a learning community, with its teachers sharing content with one another by browsing and copying peer projects, then teachers in cluster 1 are identified as isolated islanders in this IA community.

Cluster 2 (13.7%): Lukewarm teachers.

Teachers in this group do not view many other projects ($u_{\text{project\_visits}} = 0.90$). Though cluster 2 is the most productive group ($u_{\textit{number\_of}\_\text{projects}} = 5.27$), most of their projects are characterized by little content, few resource links, and rare revisions. Teachers in this group always make their projects available to their students ($u_{\text{percentage\_student\_projects}} = 0.99$) and are also willing to share them with the public audience ($u_{\text{percentage\_public\_projects}} = 0.88$), however, the IA community doesn't appear to value them much, as they are rarely copied ($u_{\text{project\_copies}} = 0.02$). As such, teachers in cluster 2 are labeled as lukewarm teachers.

Table III. LCA Results of Clustering IA Users into Seven Clusters

				Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
Cluster size		%		36.9	13.7	12.3	12.0	11.1	10.4	3.8
		N		280	103	93	91	83	78	29
Category	Indicators Name	Range	Mean							
Resource	Number of resources	0~299	12.57	6.22	12.65	9.53	4.83	12.23	22.40	81.75
IA Project	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. num. of words per project	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 hits	0~271.75	2.67	1.13	4.36	3.57	0.00	3.84	16.32	53.68
	Avg. num. of project copies	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	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 visits	0~134	8.36	2.76	0.90	6.99	3.79	30.91	10.66	35.98
	Average depth per visit	2.5~231	35.69	28.96	35.14	39.24	24.94	57.66	45.66	45.36

Notes:

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

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

c) 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 number of words per project*, *average number of project revisions*, *average number of project hits*, and *average number of project copies*.

d) 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 number of words per project*.

Cluster 3 (12.3%): Goal-oriented brokers.

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

Cluster 4 (12.0%): Window shoppers.

This group of teachers has never contributed to the IA community because they only create a few private projects, perhaps just for practice or fun. Not surprisingly, they are rare visitors compared with other groups. Recall that since all teachers without any project authoring activity or no repeated visits were excluded from this study, members in cluster 4 should not be considered the least active IA users. They browse others projects, but choose not to make their own projects visible to the public. ($u_{\text{percentage\_public\_projects}}$ = 0), not even to their students ($u_{\text{percentage\_student\_projects}}$ = 0). Considering their lurking characteristics, they represent the window shoppers in this community.

Cluster 5 (11.1%): Beneficiaries.

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

Cluster 6 (10.4%): Classroom practitioner.

Judging from the number of embedded resource links ($u_{\text{resources\_used}}$ = 9.37) and length of content (u_{words} = 371.90), the teachers in this group appear to have put in lots of efforts into authoring projects. This group's projects receive the highest number of visits ($u_{\text{project\_hits}}$ = 16.32). However, most of the hits come from their student accounts ($u_{\text{student\_hits}}$ = 23.70), and only a few are from the general public ($u_{\text{public\_hits}}$ = 1.93). We conjecture that like cluster 3, 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 are visited by other teachers, they have never been copied. The text length of the projects suggests these projects are tailored for a specific context and group of

students, and thus not easily adaptable. Given the fact that teachers in this group appear to have designed their projects to meet their very specific instructional needs, they are labeled as classroom practitioners.

Cluster7 (3.9%): Dedicated sticky users.

Very similar to cluster 3, teachers in this group serve as brokers: they both consume others' work by copying project ($u_{\text{percentage\_copy\_projects}} = 0.54$) and contribute back to this community ($u_{\text{project\_copies}} = 0.28$). They do not appear to be as goal-driven as cluster 3, as teachers in this group report unusually high visits ($u_{\text{visits}} = 26.93$), dedicate enormous time in viewing peer projects ($u_{\text{visits\_to\_other\_projects}} = 35.98$) and collecting resources ($u_{\text{resources\_collected}} = 81.96$) though the majority is not utilized in project authoring ($u_{\text{resource\_usage\_rate}} = 11\%$). In sum, this group exhibits two characteristics: dedicated to this community, and stickiest behaviors, and are therefore labeled accordingly.

6.2.2 Comparison between Three Particular Clusters. To extend the previous analyses and to specifically focus on teachers that create IA projects copied and adapted by others users, we more closely examined three particular clusters – cluster 2, 3, and 7. These teachers represent 29.8% of the studied users, or 225 people in total. Projects created by teachers in cluster 3 have the highest probability of being copied and those in cluster 2 have the least chance, and those in cluster 7 are 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 them 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 6 plots teachers ($N_{\text{cluster2}} = 93$, $N_{\text{cluster3}} = 93$, $N_{\text{cluster7}} = 29$) along two dimensions: the average number of words per project, and average number of resources per project. Since the data were skewed, we used a log transformation to make the data points more evenly distributed on the plot. The same procedure was applied to generate Figure 7 as well. If Figure 6 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 exceed 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 7, teachers in cluster 7 have gathered a much larger pool of resources than the other two groups, which presumably makes it easier for them to choose the appropriate web resources to accomplish their instructional objectives.

Given the fact that projects in clusters 3 and 7's are frequently viewed and copied, and surpass those in cluster 3 in length and in number of resource links, we can surmise that text-rich and resource-rich are two essential characteristics of high quality IA project.

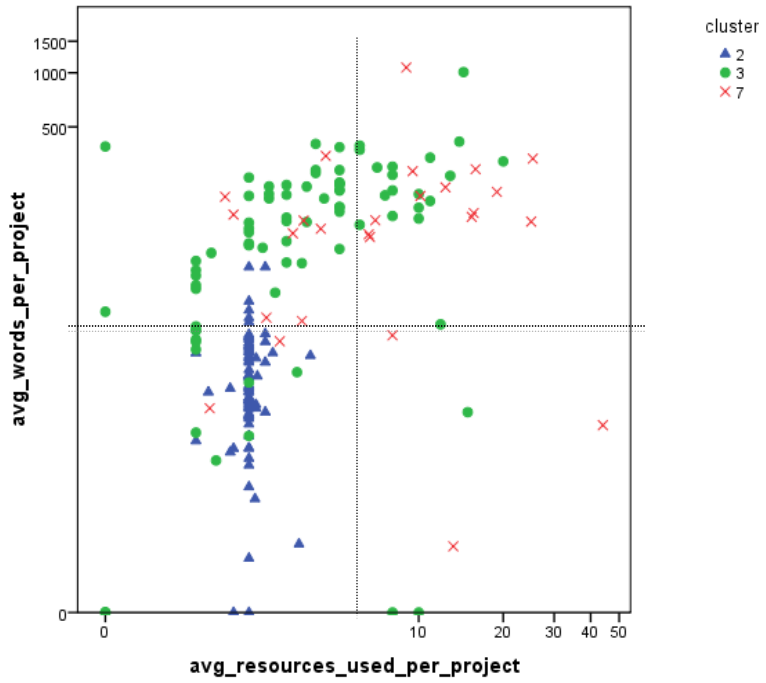

Figure 6. A plot of teachers in cluster 2, 3, and 7 in terms of number of words per project and number of resources per project (log transformed).

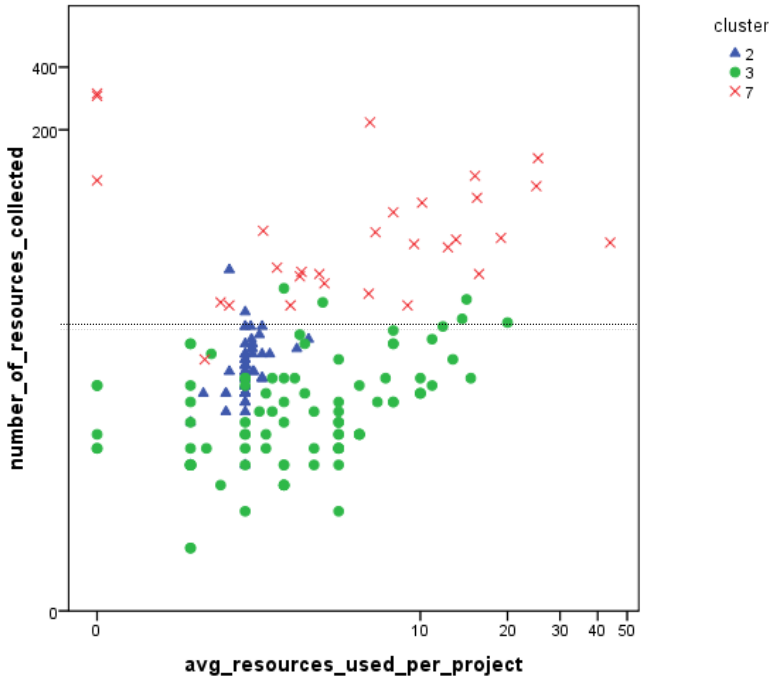

Figure 7. A plot of teachers in cluster 2, 3, and 7 in terms of number of resources collected and number of resources used per project (log transformed).

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 III). To conduct a fine-grained comparison of the browsing activity, we segmented the number of project visits into four levels (see Figure 8) using

the following procedure. We first took out the teachers with zero project visits and assigned them to the lowest level, and then calculated the mean and standard deviation for the remaining after applying a log transformation (to reduce skewness). Finally, we categorized the remaining teachers into three levels – one standard deviation below the mean, one standard deviation above the mean, and those in the middle, and finally plotted the percentage of teachers falling into each level. Figure 8 reveals that 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 visits. On the other hand, more than 75% of the teachers in cluster 2 have never visited other teachers' projects, and none of them is one or more standard deviations above the mean. This analysis provides further evidence that cluster 2 represents the lukewarm teacher group, while cluster 7 represents the sticky one in terms of the magnitude of project visits. It also suggests that engaging in browsing behavior seems to be a precursor to creating valued IA projects.

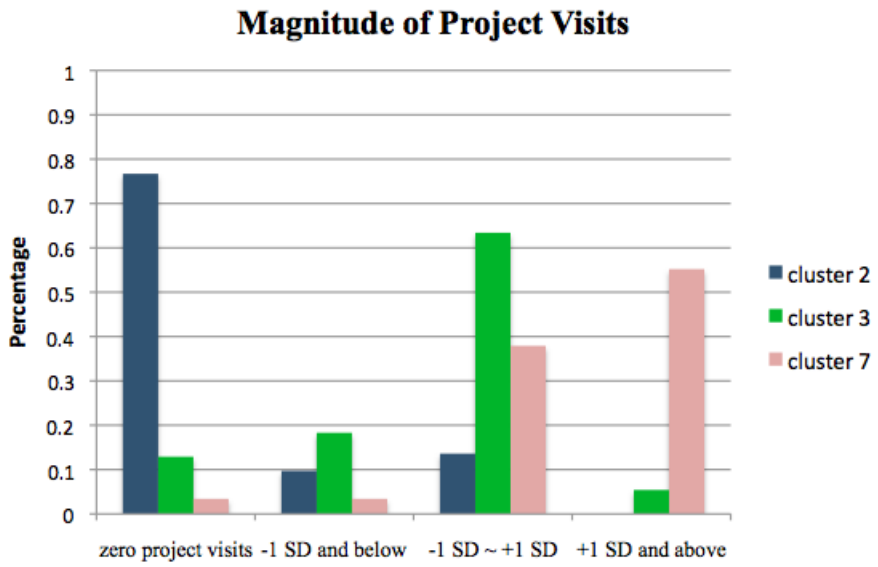

Figure 8. A plot of teachers in cluster 2, 3, and 7 in terms of the magnitude of project visits.

Public IA projects are published under the most liberal Creative Commons license, and borrowing and copying ideas is encouraged, as it helps build higher quality work. 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.

7. DISCUSSION

This article described the KDD process and its application in 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 distinguish the isolated users from the key brokers of the community.

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, 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.

Third, the third stage of KDD, evaluation and interpretation, could be conducted in a more comprehensive fashion. For example, our previous work showed that greatest use occurs in areas where IA has conducted teacher professional development workshops [Khoo et al. 2008; Xu et al. 2010], which means workshop participants have a higher chance of becoming sticky users than others. Teacher with workshop history can be singled out for analysis, and their distribution among clusters are expected to be different than the rest. Further, we plan to expand our methodology from conventional data mining approach to triangulating the findings with other more conventional data sources, such as data from usability studies, field studies, surveys, focus groups, and interviews.

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.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. 840745 & 0840738. 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.

REFERENCES

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.

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.

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.

COOLEY, R. MOBASHER, B., AND SRIVASTAVA, J. 1999. Data preparation for mining world wide web browsing pattern. *Knowledge and Information Systems 1*, 5-32.

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.

EDUCATIONAL DATA MINING. n.d. Retrieved February 15, 2010, from the International Working Group on Educational Data Mining, http://www.educationaldatamining.org/

GRIMES, S. 2007. The grand challenge for text mining. Retrieved from, http://intelligent-enterprise.informationweek.com/blog/archives/2007/04/the_grand_chall.html;jsessionid=XVSYIGQTUSHJ5QE1GHPSKHWATMY32JVN.

HALEVY, A., RAJARAMAN, A., AND ORDILLE, J. 2006. *Data integration: The teenage years.* In *Proceedings of the 32nd International Conference on Very Large Data Bases*, Seoul, Korea, 9-16.

HAN, J., AND KAMBER, M. 2002. *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., 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.

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

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 Oakes, CA, 175-198.

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.

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 February 15, 2010, 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.

SHEARD, J., CEDDIA, J., HURST, J., AND TUOVINEN, J. 2003. Determining website usage time from interactions: Data preparation and analysis. *Journal of Educational Technology Systems 32*, 101-121.

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.

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.

WEISCHEDEL, B., AND HUIZINGH, E. K. R. E. 2006. Website optimization with web metrics: A case study. In *Proceedings of the 8th International Conference on Electronic Commerce*, Fredericton, New Brunswick, Canada, 463-470.

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.

WITTEN, I. H., AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann, San Francisco, CA.

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.