A computational approach for understanding and predicting the behavior of educators using an online curriculum planning tool

Ogheneovo Dibie
University of Colorado
Institute of Cognitive Science
Dept. of Computer Science
Boulder, Colorado USA
ogheneovo.dibie@colorado.edu

Keith Maull
University Corporation for
Atmospheric Research
Boulder, Colorado USA
maull@ucar.edu

Tamara Sumner
University of Colorado
Institute of Cognitive Science
Dept. of Computer Science
Boulder, Colorado USA
sumner@colorado.edu

ABSTRACT

TODO: Working on this, still very preliminary...

In this paper, we present a computational approach for understanding and predicting the behavior of Earth Science educators using an online curriculum planning tool incorporating digital library resources. This paper expands on prior work on understanding educators' adoption and use of digital library resources [11], by extending the use-diffusion based methodology for characterizing use and understanding the trends that occur within user behaviors. . Furthermore, we show that a user's long-term behavior can be predicted to a certain extent from small window of time.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

ACM proceedings, LaTeX, text tagging

1. INTRODUCTION

In today's classroom, educational digital libraries play an important role of supporting educators in customizing their instruction to meet the diverse needs of students[18]. They provide educators with high quality resources to supplement classic teaching materials such as textbooks and teaching guides. Also, digital libraries provide learners with avenues for informal learning–learning outside the classroom–as a way of augmenting knowledge and experiences gained in the classroom [10]. Given the importance of educational digital libraries to learning and teaching, it is necessary to improve our understanding of how resources within these libraries are used. This knowledge would be beneficial in providing library providers and managers with information on how to support and improve the information needs of library patrons.

Understanding technology use spans far beyond the realm of digital libraries. In marketing science there has been extensive research into technology adoption and use. A theoretical framework for understanding technology use is the use-diffusion framework proposed by Shih and Venkatesh [16].Use-diffusion characterizes the uses of a product in two dimensions: the frequency of use and variety of use [16]. For example, following the use-diffusion model, the usage patterns of Facebook users can be characterized based on the amount of time spent on the Facebook platform and the various types of actions performed during that time. Other theories of use consider a user's content preference in addition to usage frequency and variety [5]. We discuss the use-diffusion as well as other paradigms of technology use in greater detail in section 2.

Within the realm of digital libraries, recent research has focused on computational approaches for determining the behaviors of educators. Inspired by theories of technology use from Shih & Venkatesh and Ram & Jung [16, 13], Maull et al. [11] developed a use diffusion-based methodology for understanding how a group of middle and high school Earth Science educators used an online curriculum customization tool incorporating digital library resources [11]. They used computational techniques such as clustering and clickstream entropy to produce user typologies [11]. A user typology refers to the categorization of aggregate user behaviors into distinct user types [5]. Each user type gives an overview of the manner of use of members of that type [5]. For example a user of type *power user* is one who spends a lot of time on a product and also uses it in a variety of contexts. Similar to the work of Maull et al. [11], Xu et al. studied the different user types that emerged from teacher use of a educational digital library service, the Instructional Architect [19].

While, current usage theories give insight into user types, they do not account for the evolution of such behavior. For example, a user of the type *power user* (spends a lot of time on a product and exercises most of its features) after a year

of using a product, may not have exhibited this behavior from first use. We hypothesize that user behaviors as described by user types are not static but dynamic. In fact, Shih and Venkatesh posit a similar hypothesis in elaborating on user types they discover [16]. In this research, we aim at shedding light on the evolution of a user type by understanding how the patterns of use that describe it change over time and what frequent patterns can be observed within it. Using a market-basket analysis type approach to analyze the usage patterns within a user type, we get a better understanding of the unique patterns of use that exist within that type. This would help in understanding the correlations that occur within specific user types. For example, when a user of a type *power user* spends a high or low amount of time on the system, what actions are they likely to be performing etc.

Furthermore, current research on digital library use is mostly retrospective, i.e. a user's behavior is determined after a set period of use. Studies [11, 19] were based on a year of use. We can however go a step further by predicting a user's behavior based on knowledge of a user typology. This could be especially useful for new users of the system as this knowledge can help system providers/managers with information on how to better support their needs based on predicted behaviors. In our domain (educational digital library resources), this could mean implementing better professional development (PD) training or accessibility to resources. For example, if after a couple months of usage, a user's behavior is predicted as likely being of the type *limited use* (spends very little time on the system and likely to discontinue use), efforts could be made to encourage use or at least understand why the product is barely being used. In this paper, we develop computational models for predicting the long-term behaviors of users from small windows of time. In this paper, we posit long term behavior as a user's membership in a user type at the end of our observation period–one academic year. This knowledge would be beneficial in providing library administrators/managers with timely information on what system changes/user training would improve the use of their platforms.

*We state final results here*I state my results here and paper organization....

2. RELATED WORK

This research draws upon theoretical models and computational approaches of determining technology use. We examine these theories in this section

2.1 Theoretical models of technology use

Technology use research follows from a rich history of work in technology adoption. Technology adoption occurs when a user decides an innovation is of utility and decides to use it [17]. Most research on technology adoption is focused on factors that influence a user's decision to adopt an innovation. They include Roger's innovation diffusion theory where the five "adopter categories", a user's propensity to adopt an innovation and the spread of that innovation within a social system are examined [14]. Other models of technology adoption include the concerns based model which focuses on an an individual's specific reasoning behind adopting an innovation [6, 7] and the technology acceptance model [9] which takes into account the influence of an individual's self efficacy and expertise in deciding whether to adopt a technology or not [9].

However technology adoption models do not account for how a technology is being used. Knowledge of how a product is used or not being used is key for manufacturers in making improvements/changes to the product. This need has led to the rise of theories that explain actual technology use.

A key theory on technology use is the theory of use-diffusion. Originally proposed by Ram and Jung[13], it measures technology use on two dimensions: frequency and variety. Usage frequency refers to how often a product is being used while variety refers to the different applications or contexts within which a product is being used. For example, usage frequency for an Ipad will be based on how long it is being used while usage variety encompasses the various applications it is used for e.g. games, word processing, camera etc.

The work of Ram and Jung was expanded on by Shih and Venkatesh [16]. Through self-report questionnaires and a dairy study [1]of individuals using a household technology, they discover a user typology of four distinct user types: *intense use, non-specialized use, specialized use, and limited use.* Users of type *intense use* are characterized by high frequency and a high variety of use. Users of type *specialized use* are characterized by a high frequency but a low variety of use. An example of this could be an administrative assistant who spends a lot of time on a computer but only uses it for processing word documents. Users of type *limited use* are characterized by low frequency and variety of use. Users of type *Non-specialized use* are characterized by a high variety of use and low frequency.

2.2 Computational approaches for determining user typologies

Recent research has focused on extending theoretical foundations on user typologies to educational contexts. Rather than determining user types solely based on self-reported usage as is the case in [16, 13], these computational approaches generate a typology of user behavior via clustering of actual usage. Clustering is a data mining technique for automatically grouping related items into bins. 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) [8]. Xu [19] examined the use of clustering techniques to generate fine grained user typologies within a web based instructional tool known as the Instructional Architect (IA). The IA is an educational digital library service designed to facilitate the creation of simple instructional projects using web resources from the National Science Digital Library (NSDL) and the web in general[9]. Projects created can be kept as private, shared with just students and other teachers or made publicly available on the web. Three

[1]A diary study is a study method common in the fields of anthropology, psychology and Human Computer Interaction (HCI) in which users log their interactions with a product. It provides a great way for researchers to context longitudinal data on a user's experience with a product.

clusters of users were discovered namely: key brokers, insular classroom practitioners and ineffective islanders [19]. Key brokers frequently browsed projects created by other users in IA and their own projects also received a lot of attention from other users; their public projects were of especial high quality [19].Insular classroom practitioners did not create high quality projects because their projects were characterized by very little content and links to external resources. They had little interest in viewing projects created by other IA users and most of the projects they created were limited to use with students in their classrooms. Ineffective islanders were characterized by publishing a single project of supposedly good quality but these projects were not shared with neither students nor the public [19].

Working with the Curriculum Customization Service (CCS)– an online curriculum planning tool incorporating digital library resources [18], Maull et al. [11] develop a typology of user behaviors observed in the CCS. This typology is inspired by a use-diffusion based methodology and characterizes use based on frequency and variety type metrics observable through server usage logs. Frequency based metrics include the number of session, hours spent, and variety based metrics include areas of the CCS that were accessed such as interactiveresources, publisher materials, and user-contributed resources. We provide detailed explanation of these features in section 4.

Our work is based on the same context as Maull et al [11] .This research extends the work of Maull et al. [11] use-diffusion based methodology of understanding technology use in two keys ways:

1. This work provides a deeper understanding of how usage features that characterize user types trend within the time period of the observed user type. Furthermore, this work introduces a market-basket analysis of usage features to understand how usage features correlate with each other within each user type. This answers the question of when a user is doing X, what else are they likely to be also doing.

2. This works explores computational models to predict an educator's behavior from small windows of time. It aims to answer the question of can we predict a user's type at the end of a year from an early window of usage?

2.3 Predicting User Behavior

Predicting user behavior on the web has been of major research interest in recent times. This is primarily due to the need to improve user experience. Research studies have looked into the correlation between the online activities of users in order to understand how one influences the other [1]. For example, this could be the likelihood a user purchasing an item after reading reviews or a blog post about it [1]. Other research have explored online user behavior at a much lower level of granularity such as the likelihood of a user clicking on a sponsored search result [2] [4]. Attenberg et al. [4] show that a user's click through behavior on sponsored search results can be predicted with a decent degree of accuracy based on the user's search query,result and prior behavior. User behavior on sponsored search result is determined by the number of additional clicks made and time spent on sponsored results.

In this paper, we aim at exploring user behavior prediction in the context of an online planning tool incorporating digital library resources. We aim at identifying the most predictive features of a user's eventual user type and the earliest window of time within which a user's behavior can be accurately predicted.

3. RESEARCH QUESTIONS

The specific research questions to be addressed in this paper are as follows:

1. Does the usage pattern which describe a user type remain the same or does it vary and if so how?
   In this question we examine how the usage patterns (in terms of usage features) that describe a user type change from time to time

2. What are the frequent usage patterns that can be observed within a user type?
   This question gets at what usage features go together within a particular user type. It would provide a better understanding of what users of a particular type are likely doing when doing something else.

3. How well can computational models predict a user's eventual user type from smaller time windows?
   We examine a set of classifiers for predicting the user type of users at the end of a school year from earlier time frames such as the first month of use and first semester of use. We aim to discover what time frame and usage features are best predictive of a user's user type

4. RESEARCH CONTEXT

The research questions outlined in this proposal will be examined in the context of an instructional curriculum planning tool called the Curriculum Customization Service (CCS) [18]. The CCS is a web based instructional planning tool which provides primarily middle and high school Earth Science teachers with access to digital versions of their class room text book, curriculum-relevant high quality, digital library resources and community-contributed resources. These resources are used in a multitude of pedagogical scenarios: from lesson planning to in-class projection and demonstration to customizing instruction to meet the needs of a diverse group of learners [15]. Educators use the CCS as a supplement to other instructional aids such as text books and teaching guides. Deployed since the fall of 2009, the CCS is now being used in six school districts across Colorado and Utah. The CCS is instrumented to capture the click actions of users. Click actions being tracked include all unique page elements such as clicks on links, toggles, tabs, and buttons. The aggregate data collected from the clicking activity of users is collectively known as clickstream. Clickstream data is useful in not only useful in showing a user's navigational path i.e. pages clicked on during a visit to a website, and

[2]Sponsored search results are advertisements displayed by search engines next to the actual search results returned. They are major revenue source for most web search engines.

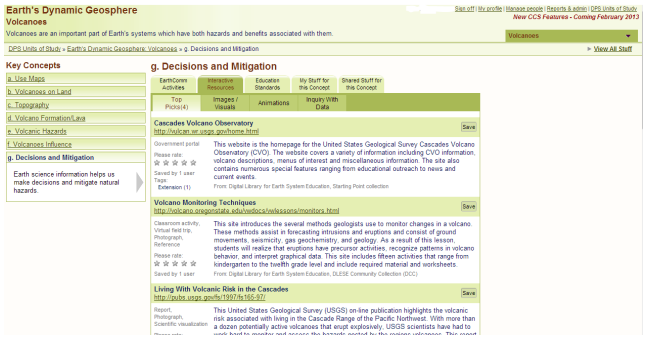

Figure 1: Snapshot of the CCS platform highlighting its main capabilities: Publisher materials (EarthComm Activities tab), Digital library resources (Interactive Resources tab), personalization capabilities(My Stuff tab) and community features (Shared stuff tab). The Interactive Resource tab is opened in this figure showing the top recommended digital resources

click actions performed within a specific page (i.e. page elements that were clicked on within a particular page), but also a user browsing behavior [12]. Clickstream analysis has proved to be very useful in developing user typologies from online usage as illustrated by Maull et al. [11] with the CCS and xu et al.[19] with the Instructional Architect. The clickstream data under analysis for this research includes information such as the user identity (user id), time stamp, specific system components clicked on (within the CCS this could be an interactive resource from a digital library such as DLESE or a user-contributed resource), session length. A series of clicks performed over a period of time constitutes a session and a session is delineated by 45 minutes of inactivity or by the user logging out.

4.1 Data

Our analysis is based on the clickstream of a large urban school district with 80 science educators who used the CCS during the 2011-2012 school year. About 20,000 click actions were registered in members of this district

4.1.1 Feature selection

Our study in this research is based on six usage features highlighted in table 1. These features cover the four core areas of the CCS which are: (1) Digital library resources, (2)Publisher material, (3)Community contributed resources (4)Personalization

Digital Library resources: A core aim of the CCS is to enable users customize instruction with high-quality digital library resources. The CCS is designed to automatically pull educational resources from digital libraries such as the Digital Library for Earth System Education (DLESE) and align these resources to a curriculum and educational standards. Within the CCS' interface, these resources are collectively referred to as interactive resources(IR resources). IR resources are sub-grouped into four categories depending on its type. They are : Top Picks, Animations, Images/Visuals and Inquiry with Data.

Table 1: CCS usage features that characterize a user type

Num	Feature
1	Total number of sessions (frequency)
2	Total number of hours spent on the site (variety)
3	Total "My Stuff" activity (variety)
4	Total activity within publisher material (variety)
5	Total Shared stuff activity (variety)
6	Total Interactive Resources activity (variety)

Publisher Materials: Publisher materials are digitized versions of primarily paper-based instructional materials. Publisher materials include a digitized version of the student textbook, and supplemental teaching materials such as instructional support materials, teaching tips and embedded assessments. Access to digitized version of classic teaching materials within the CCS makes curriculum planning and in-class usage more accessible and convenient.

Personalization: The CCS provides a host of personalization features which we refer to as **My Stuff**. As users go through the CCS platform, My Stuff features they can save publisher materials, digital library and community-shared resources to their My Stuff area. In addition to saving other system components, users can also upload their own resources to My Stuff and retrieve as needed at a later time.

Community Features: The CCS promotes social behavior such as sharing resources among users. Within each section of the curriculum, users are able to share resources such as documents, links and rich media with others. Other users who find these resources useful can use or save these resources in My Stuff for later use.

5. CHARACTERIZING USER TYPES

Before proceeding with our analysis on understanding and predicting the evolution of user types, we introduce a framework for characterizing the user types of a typology. We follow a similar approach to generating a user typology as Maull et al [11].The expectation maximization (EM) clustering algorithm is applied to the usage logs of all users during the months of September - May to generate a typology of user behaviors. Typically, the school calendar runs from August-June, but usage in August and June are partial as school begins in late August and ends in early June. Thus, we omit them from our analysis. We examine usage through six features which cover the main components of the CCS platform. These features are highlighted in table 1 with the feature type i.e. whether this a frequency or variety based feature indicated in parenthesis.

In prior computational approaches for generating user typologies, user types discovered via clustering are labeled in a manual ad-hoc fashion [11, 19]. There are no set rules for mapping a cluster to a specific user type. In this work we introduce an equal frequency binning discretization based approach for mapping clusters to a specific user type. Discretization based framework has successfully been used in defining feature values that describe user type archetypes in prior studies [5, 3]

The discretization works as follows:

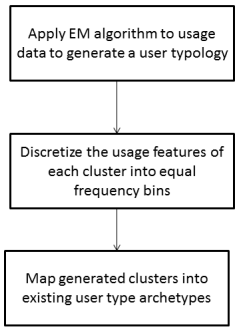

Figure 2: Process flow of labeling a cluster with a user type

Table 2: Discretized feature values from cluster results of Maull et al. [11]

Feature	1	2	3	4	5
Sessions	Low	Mid-High	High	High	High
Hours	Low	Mid-High	High	High	High
IR Activity	Low	Low-mid	High	Mid	Low
My Stuff Activity	Low	Mid-High	High	Mid	High
Shared Stuff Activity	Low	High	High	High	High
Publisher Activity	Low	Low	High	Mid	Low

1. For each usage feature, we perform equal frequency discretization on its continuous value into three bins: low, medium and high. Using a naive binning approach such as splitting a feature range into thirds can lead to large frequency skews within a single bin. Equal frequency binning avoids this and provides a distribution-dependent notion of levels [8].

2. We use the user types developed by Maull et al. [11] and their discretized user types as archetypes for characterizing usage as illustrated in table 2

3. The EM algorithm is run against the total usage to generate unique clusters. The output of EM clustering is a set of clusters(user types) with the cluster mean value for each feature. The usage feature value for each cluster is discretized into low, mid and high bins.

4. We attempt to map the discretized feature values to each of the user type archetypes from Maull et al. [11]. If a mapping to a user type does not exist, we introduce a new label that describes the cluster based on its feature values.

In table 2, Cluster 1 refers to the **limited use** "uninterested adopter" category. Members of this category have very low overall use of the CCS. Cluster 2 refers to the **specialize use** "Interactive Resource specialists." Members of this category make heavy use of interactive resources relative to other system features. Cluster 3 refers to the **Intense use** "Ardent Power User" category. Members of this group feature heavy robust use of all system features. Cluster 4 is the **Non-Specialized Use**"Moderate Generalist" category. They exhibit moderate overall system use. Cluster 5 is the **Specialized Use** "Community Seeker Specialist" category. They make heavy use of shared stuff and interactive resources relative to publisher materials

Table 3: Output of clustering experiments

Feature	1	2	3
Sessions	Low(4.048)	high(97.2)	high (26.5)
Hours	Low (0.649)	High (17.802)	High (3.946)
IR Activity	Mid (3.867)	High (80.199)	High (18.275)
My Stuff Activity	Low(0.781)	High (10.2)	High (9.645)
Shared Stuff Activity	Low(1.6524)	High(21.6)	High(39.6)
Publisher Activity	Low(6.438)	High (162)	High (40.045)
N	32	5	43

5.1 User typology

After clustering usage for all users we get a user typology as shown in table 3

There were three clusters generated from our user type experiments. We describe them as follows:

Cluster 1: With all it's features with a value of *Low* apart from interactive resources with a value of mid, this cluster is closest to the user archetype of limited use "uninterested adopter." An introspection into the login frequency of members of this type, reveals that a good portion of users stopped using the CCS after the first few months of the school year.

Cluster 2: The usage pattern of cluster 2 with all feature values having a value of high is most closely correlated to that of the power user archetype.

Cluster 3: This cluster of users also have a high rating on average in all usage features. However, the values of each of these features are in the low range of the high value bin. On average users in this clusters are not as intensely engaged with the CCS platform as the power users in cluster 2. We would categorize these users as Community Seeker specialists, because they feature higher use of shared stuff compared to users of the other two clusters (in terms of mean number of clicks).

6. USAGE PATTERN TRENDS

We explore the trend in usage pattern on a feature-by-feature basis. In this analysis we look into how the values of each usage feature per user type changes on three levels: semester to semester, every two months and month to month. This would give a good idea of which usage features remain relatively stable (in terms of use) and which vary per user type. This knowledge would be particularly useful in detecting strong features that define a user type. For example, if the

Table 4: Usage feature values of during the spring semester of the 'Limited use' user type

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Fall semester	Low	Low	mid	mid	Low	Low
Spring semester	Low	Low	mid	mid	Low	Low

Table 5: Bi-monthly usage feature values of the 'Limited use' user type

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Sept-Nov	Low	Low	mid	mid	mid	mid
Nov-Jan	Low	mid	mid	Low	mid	Low
Jan-Mar	Low	mid	mid	mid	Low	mid
Mar-May	Low	Low	mid	Low	Low	mid

power user type always has a high value in publisher material activity, this might point to the fact that a high value in publisher material activity is a strong predictor of a 'power user' user type.

6.1 Limited use user type

6.1.1 Semester-Semester trends

According to table 4, all usage feature values remain the same across both semesters

6.1.2 Bi-monthly trends

6.1.3 Monthly trends

In this user type, the hours spent and publisher activity features seem to be very stable and hence good predictors of membership within —-. The other features number of sessions, shared stuff activity, interactive resource activity tend to vary quite a bit from month to month.

6.2 Community Seeker user type

6.2.1 Semester-Semester trends

As highlighted in table 7, within the community seeker user type, activity in my stuff, shared stuff and publisher material remained at a high level on average. However, frequency based features $hours\&numberofsessions$ moved from the high-mid and mid-high level ranges respectively. This might indicate that users logged spent less time across more sessions and more time across less sessions in the fall and spring semesters respectively. Also use of interactive resources dropped from high to mid from the fall to spring semesters

Table 6: Monthly usage feature trends in the limited user type

	Num Sess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Sept	Low	mid	mid	Low	mid	Low
Oct	Low	mid	mid	Low	Low	mid
Nov	mid	mid	mid	Low	mid	Low
Dec	Low	mid	high	mid	mid	mid
Jan	mid	mid	mid	Low	mid	Low
Feb	mid	mid	Low	Low	mid	Low
Mar	mid	Low	mid	mid	mid	mid
Apr	Low	mid	mid	Low	mid	Low
May	Low	mid	Low	Low	Low	Low

Table 7: Usage feature values of during the spring semester of the 'Community seeker' user type

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Fall semester	high	mid	high	high	high	high
Spring semester	mid	high	mid	high	high	high

Table 8: Bi-monthly usage feature values of the 'Community Seeker' user type

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Nov-Jan	high	mid	mid	mid	mid	mid
Jan-Mar	mid	mid	mid	mid	mid	high
Mar-May	mid	mid	mid	high	mid	high

6.2.2 Bi-monthly trends

6.2.3 Monthly trends

All usage features apart from interactive resource activity tend to vary quite a bit from month to month. Though, further introspection indicates that users in this category rank in the high range of mid interactive resource activity.

6.3 Power user user type

Within the 'power user' user type, usage features were examined in a semester-semester, bi-monthly and monthly basis.

6.3.1 Semester-Semester trends

All usage features trend at a high level for members of the power user category during the fall and spring semester on average. This is similar to the trend overall during the entire school year. We examine these features on a deeper level to discover if any differences in the values of these features can be observed on a bi-monthly or monthly level.

6.3.2 Trends every two months

Considering 2-month time frames as illustrated in table 10, all usage features maintain a value of high apart from My Stuff and Shared Stuff activities during the school year.

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Sept	mid	mid	mid	mid	mid	high
Oct	mid	mid	mid	mid	mid	high
Nov	high	high	high	high	mid	high
Dec	high	mid	mid	mid	mid	high
Jan	high	high	high	high	high	mid
Feb	mid	mid	high	high	mid	high
Mar	high	mid	high	high	mid	mid
Apr	mid	mid	high	high	mid	high
May	mid	high	mid	mid	high	mid

Table 9: Usage feature values of during the spring semester of the 'Power User' user type

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Fall semester	high	high	high	high	high	high
Spring semester	high	high	high	high	high	high

Table 10: Bi-monthly usage feature values of the 'Power User' user type

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Sept-Nov	high	high	high	high	high	high
Nov-Jan	high	high	high	mid	high	high
Jan-Mar	high	high	high	high	high	high
Mar-May	high	high	high	mid	high	mid

My stuff activity dips between time periods of September-November and NovemberJanuary and again from between time periods JanuaryMarch and MarchMay. My Stuff and Shared stuff activity dips to a Low value range from a high-value range during the time of periods of JanuaryMarch and MarchMay.

6.3.3 *Monthly trends*

From table 11, it is clear that the number of sessions, hours and my stuff activity varies remain constantly high across all months. This indicates that a value of high in these three features is a strong predictor of membership within this cluster. The values of interactive resource activity, my Stuff activity and shared stuff Activity vary quite a bit within the school year.

7. FREQUENT PATTERNS OF USE

The aim of this study is to understand what actions of members of a specific user type usually go together. As an example, we would like to understand what members of the 'power user' user type are likely doing when they spend a high amount of time on publisher materials. This would give a good understanding of what usage feature correlations occur within a user type.

Table 11: Usage features month-month in the Power User user type

	Num Ssess	Hrs	IR Actv	MS Actv	Pub Actv	Shared Stuff
Sept	high	high	high	mid	high	high
Oct	high	high	high	high	high	mid
Nov	high	high	mid	mid	high	mid
Dec	high	high	high	Low	high	mid
Jan	high	high	high	high	high	high
Feb	high	high	Low	mid	high	mid
Mar	high	high	high	mid	high	high
Apr	high	high	mid	Low	high	Low
May	high	high	Low	Low	high	mid

We take a market basket analysis approach to understanding the frequent patterns that occur within each user type. Market basket analysis is a data mining approach for understanding consumer behavior through their transaction patterns [8]. It seeks to uncover meaningful and interesting associations in customer purchase data. A canonical example of the use of market basket analysis is a grocery store analyzing consumer purchases to understand what items shoppers usually purchased together. Upon discovery that beer and diapers occur frequently in consumer purchases, the store may decide to place these items next to each other or have a group sale of both items together to further encourage this shopping behavior.

To implement a market basket analysis on user behavior, we consider each member of a user type's monthly usage as a 6-item transaction (a transaction can also be referred to as an itemset) of the form {num_session, hours_spent ir_actv, shared_stuff, my_stuff, pub_stuff } representing each of the usage features we analyze. Each item has a discretized value of either high, medium or Low as described in section 5. Finally, we run the Apriori frequent pattern mining algorithm on all transactions within a user type to find frequent patterns that are interesting. Apriori is a seminal algorithm introduced by Agrawal and Srikant [2] for performing fast frequent itemset minning in large transaction databases to discover association rules of the form $\mathbf{X} \Longrightarrow \mathbf{Y}$ where $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{I}$. $\mathbf{I}$ is an itemset of the form $\mathbf{I} = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \}$ within the set of transactions $\mathbf{D}$. The association rules generation process of the Apriori algorithm can be summarized in two steps:

1. A minimum support is applied to itemsets(transactions) to find all *frequent itemsets* in the set of transactions. The support, $\mathbf{supp}(\mathbf{X})$ of an itemset X, is defined as the proportion of itemsets in the data set which contain the itemset.
   $supp\,(X) = \frac{number of transactions which contain the itemset X}{total number of transactions}$
2. A minimum confidence constraint is applied to the frequent itemsets to form rules. The confidence of a rule is the probability of finding the antecedent(RHS) of a rule given it's precedent(LHS).
   Formally, $conf(X \Longrightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$. For example a rule $\mathbf{milk, diapers} \Longrightarrow \mathbf{beer}$ with a support of 65% indicates that the rule is correct for 65% of all transactions containing milk and diapers. A rule is identified as *strong* (i.e. a likely occurrence) if it meets the minimum support and confidence levels.

In our analysis we set a minimum support and confidence level of 0.9 for strong rules. We do this to generate the most likely rules and constrain the size of frequent items generated by the Apriori algorithm.

7.1 Frequent patterns within the Power User user type

We ran the Apriori algorithm on the set of all transactions of all users in the power user type and the ten strong rules (rules meeting the minimum support of 0.5 and minimum confidence of 0.9 thresholds) were generated. They are as follows:

1. $num\_sessions = high \;\; hours = high \longrightarrow pub_activity = high \;\; conf$: (0.96)
2. $pub\_activity = high \longrightarrow num\_sessions = high \;\; conf$: (0.94)
3. $num\_sessions = high \longrightarrow pub\_activity = high \;\; conf$: (0.94)
4. $pub\_activity = high \longrightarrow hours = high \;\; conf$: (0.94)
5. $hours = high \;\; pub\_activity = high \longrightarrow num\_sessions = high \;\; conf$: (0.93)
6. $num\_sessions = high \;\; pub\_activity = high \longrightarrow hours = high \;\; conf$: (0.93)
7. $num\_sessions = high \longrightarrow hours = high \;\; conf$: (0.9)

As a reminder, members of this user type are categorized by a usage pattern: { } As expected, most rules show strong correlations between usage features with a 'high' rating. Rules (2) & (3) indicate an interesting correlation between My Stuff Activity and Publisher Activity. It indicates that per month usage of the CCS, when most users have a high frequency of use (spend alot of time on the CCS relative to everyone else), they are likely to be spending most of their time on publisher materials rather than My Stuff activity.

7.2 Frequent patterns within the Community seeker user type

On average, members of this group did not exhibit consistent usage of the CCS compared to the power user group. A small proportion of users did not use the CCS during every month under analysis.

Only one strong association rule was generated from running Apriori on the set of transactions from this dataset:

1. $Hours = Low \;\; Pub\_Actv = mid \;\; Shared\_stuff\_Actv = Low \longrightarrow IR\_Actv = Low \;\; conf$: (0.93)

This pattern is interesting as users within this group are more inclined to use shared stuff resources. However, when they spend a Low amount of time on the system, they are more likely to be accessing publisher material than shared stuff materials.

7.3 Frequent patterns within the limited use user type

This cluster represents the sets of limited users i.e., users whom had a Low frequency and variety of use. Although this group features a large of users (n=32), there were only 70 transactions recorded during the 9 month study period, an average of 2.19 transactions per user. There were many users within this group who either registered use only in the month of September and stopped using the system or had very sparse use throughout the school year.

Three association strong association rules were discovered for members of this group. They are:

Table 12: Fall & Spring semester predictions

	Fall	**Spring**	**Fall *bf***	**Spring *bf***
Naive Bayes	71.43%	68.25%	78.57%	61.9%
J48	75.57%	68.25%	75.57%	68.25%
SVM	61.43%	66.67%	78.57%	68.25%
Bagging	74.29%	61.9%	74.29%	61.9%

1. $Hours\_spent = Low \;\; Shared\_stuff\_activity = Low \longrightarrow My\_Stuff\_Activity = Low \;\; conf$: (0.96)
2. $Publisher\_activity = Low \;\; Shared\_stuff\_activity = Low \longrightarrow My\_Stuff\_Activity = Low \;\; conf$: (0.96)
3. $Interactive\_Resource\_Act = Low \;\; Shared\_stuff\_activity = Low \longrightarrow My\_Stuff\_Activity = Low \;\; conf$: (0.91)

Though not particularly interesting, these rules re-emphasize the fact that members of this user type had very Low system usage.

8. PREDICTING USER BEHAVIOR

This study would look at machine learning classifiers for predicting a user's behavior from small windows of time. The goal of this study is two fold:

- Determine the earliest window of time that provides the best prediction of a user's eventual class
- Determine the usage feature(s) that are most predictive of a user's eventual type

We consider the task of predicting a user's eventual type by considering features in the semester, bi-monthly and monthly basis. For each time window, we perform classification with four algorithms: Naive Bayes, SVM, J48 Decisions Trees and Bagging. The accuracy of the best algorithm is reported here. Furthermore, we determine the most predictive features of a user type per time window automatically using the attribute selection filter in WEKA [3] We also perform classification tasks using just this set of features.

8.1 User type prediction per semester

On a semester-semester basis, we discovered that the fall semester is best predictive of a user's type at the end of the school year. Using all the usage features, we can predict the user type with an accuracy of 75.57%. Using WEKA's attribute selection filter, we discovered that the number of sessions, hours spent and publisher material activity are the most predictive features of a user's type. Using these features alone improved prediction to an accuracy of 78.57%. The results of our classification experiments with different classifiers are shown in table 12

8.2 User type prediction bi-monthly

We whittled our prediction time window to time chunks of two months of use. Thus, we considered the probability of predicting a user's type from usage in the following two

[3]www.cs.waikato.ac.nz/ml/weka/

month chunks: September-November, November-January, January-March and March-May. We discovered that usage during March-May are best predictive of a user's type. Our experiments show that we can predict a user's type with an accuracy of about 84.15% considering all usage features. Running the classification task with the auto-selected best features did not boost the classification accuracy in comparison to using all the usage features.

8.3 User type prediction per month

Finally, we considered the task of predicting a user's type on a month by month basis. Thus given a user's usage for a month, which month is best predictive of the user's eventual type. Running the classification tasks with all usage features as described in table 1, usage for the month of April is the most predictive of a user's eventual type given with an accuracy of 71.25%. This is not surprising as our results in section 8.2 shows that the time window of March-May as the most predictive of a user's type.

However, as in our previous analysis, using the best features of the number of sessions, shared stuff activity and publisher material activity (this feature subset was selected by WEKA's CfsSubsetEval function), we discovered that the month of October provides a higher prediction accuracy of 72.5%.

9. LIMITATIONS

10. DISCUSSION AND CONCLUSION

11. REFERENCES

12. ADDITIONAL AUTHORS

Additional authors: John Smith (The Thørväld Group, email: `jsmith@affiliation.org`) and Julius P. Kumquat (The Kumquat Consortium, email: `jpkumquat@consortium.net`).

13. REFERENCES

[1] E. Adar, D. S. Weld, B. N. Bershad, and S. S. Gribble. Why we search: visualizing and predicting user behavior. In *Proceedings of the 16th international conference on World Wide Web*, pages 161–170. ACM, 2007.

[2] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[3] S. Angeletou, M. Rowe, and H. Alani. Modelling and analysis of user behaviour in online communities. In *The Semantic Web–ISWC 2011*, pages 35–50. Springer, 2011.

[4] J. Attenberg, S. Pandey, and T. Suel. Modeling and predicting user behavior in sponsored search. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1067–1076, New York, NY, USA, 2009. ACM.

[5] P. B. Brandtzæg. Towards a unified media-user typology (mut): A meta-analysis and review of the research literature on media-user typologies. *Computers in Human Behavior*, 26(5):940–956, 2010.

[6] F. F. Fuller. Concerns of teachers: A developmental conceptualization. *Teaching Effectiveness: Its Meaning, Assessment, and Improvement*, page 175, 1975.

[7] G. E. Hall. The concerns-based approach to facilitating change. *Educational Horizons*, 57(4):202–08, 1979.

[8] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[9] P. Legris, J. Ingham, and P. Collerette. Why do people use information technology? a critical review of the technology acceptance model. *Information & management*, 40(3):191–204, 2003.

[10] M. H. Marchionini Gary. The roles of digital libraries in teaching and learning. *Communications of the ACM*, 38(4):67–75, 1995.

[11] K. E. Maull, M. G. Saldivar, and T. Sumner. Understanding digital library adoption: A use diffusion approach.

[12] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4):579–595, 2004.

[13] S. Ram and H.-S. Jung. The conceptualization and measurement of product usage. *Journal of the Academy of Marketing Science*, 18(1):67–76, 1990.

[14] E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[15] M. G. Saldivar. *Teacher Adoption of a Web-based Instructional Planning System*. PhD thesis, Boulder, CO, USA, 2012. AAI3508034.

[16] C.-F. Shih and A. Venkatesh. Beyond adoption: development and application of a use-diffusion model. *Journal of Marketing*, 68(1):59–72, 2004.

[17] E. T. Straub. Understanding technology adoption: Theory and future directions for informal learning. *Review of Educational Research*, 79(2):625–649, 2009.

[18] T. Sumner and C. Team. Customizing science instruction with educational digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 353–356. ACM, 2010.

[19] B. Xu. *Understanding Teacher Users of a Digital Library Service: A Clustering Approach*. PhD thesis, Utah State University, 2011.