A computational approach for understanding and predicting user behavior

Ben Trovato*
Institute for Clarity in Documentation
1932 Wallamaloo Lane
Wallamaloo, New Zealand
trovato@corporation.com

G.K.M. Tobin†
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld‡
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

ABSTRACT

TODO: Still 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' use of digital library resources [maull], by introducing a methodology for characterizing and understanding the patterns of use that characterize user behaviors. Furthermore, we show that a user's long-term behavior can be predicted to 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 [8]. 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 [6]. 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.Use-diffusion characterizes the uses of a product in two dimensions: the frequency of use and variety of use [venkatesh]. 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 site and the various types of actions performed during that time. Other theories of use consider a user's content preference in addition to frequency and variety [Brandtaegz]. We discuss the use-diffusion as well as other paradigms of technology use in greater detail in section 4.

Within the realm of digital libraries, recent research has focused on computational approaches for determining the usage behaviors of educators. Inspired by theories of technology use from Shih and Venkatesh and Ram and Jung [venkatesh,jung], Maull et al.[maull] 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 [5]. They used computational techniques such as clustering and clickstream entropy to produce user typologies [5]. A user typology refers to a collection of aggregated user behavior over a period of time. Each behavior within a typology is referred to as a user type. A user typology gives insight into the breath and depth of a product use across all user types. Similar to the work of Maull et al. [5], Xu et al. studied the different unique user typologies that developed from teacher use of a educational digital library service, the

*Dr. Trovato insisted his name be first.
†The secretary disavows any knowledge of this author's actions.
‡This author is the one who did all the really hard work.

Instructional Architect [6].

While, current usage theories give insight into aggregate user behavior, they do not account for the evolution of such behavior. For example, a user categorized as a âĂIJpower userâĂİ (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 a user types are not static but dynamic. In fact, Shih and Venkatesh posit a similar hypothesis in elaborating on user types they discover. [3]. In this research, we aim at understanding how the patterns of use that describe a user type change over time and what frequent patterns can be observed within a particular user behavior. Using a market-basket analysis type approach to analyze the usage patterns of users 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 specific user type spends a high/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 [maull, xu] 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 is could be especially useful for new users of the system as this knowledge can help system providers/managers with information on how to influence their behavior in one way or the other. In our domain (educational digital library resources), this could mean implementing better professional development (PD) training or accessibility to resources. In this paper, we develop computational models for predicting the long-term usage behaviors of users from small windows of time. 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....

2. RELATED WORK

This research draws upon theoretical models and computational approaches of determining technology use. We examine these thoeries 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[ETStraub]. 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 [rogers]. Other models of technology adoption include the concerns based model which focuses on an an individual's specific reasoning behind adopting an innovation [fuller, hall] and the technology acceptance model [TAM] which takes into account the influence of an individual's self efficacy and expertise in deciding whether to adopt a technology or not [TAM ref]

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[jung], 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. Through self-report questionnaires and a dairy study [1]

The work of Ram and Jung was expanded on by Shih and Venkatesh [venkatesh]. Using the use-diffusion methodology to describe the behavior of users 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, 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) [13]. Xu [9]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 clusters of users were discovered namely: key brokers, insular classroom practitioners and ineffective islanders [9]. Key brokers frequently

[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.

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 [7].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 [9].

Working with the Curriculum Customization Service (CCS)– an online curriculum planning tool incorporating digital library resources [ccs], Maull et al. [maull] develop a typology of user behaviors observed in the CCS. This typology is inspired by a use-diffusion methodology and characterizes use based on the frequency and variety type metrics observable through server logs. These metrics include the a user's number of session, hours spent, and variety based metrics include areas of the CCS that were accessed such as Interactive Resources, Publisher material, shared stuff and my stuff. We provide detailed explanation of these features in section 4.

Our work is based on the same context as Maull et al.[3].This work furthers Maull et al. [maull] use-diffusion based methodology of understanding technology use in two 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. Thus 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 say the first month?

2.3 Predicting User Behavior

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 introduce a framework for characterizing user types and examine how the usage patterns 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?
   In this paper, 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 with time frame and class of classifier is best predictive of a user's eventual user type

4. RESEARCH CONTEXT

The research questions outlined in this proposal will be examined in the context of an instructional planning tool called the Curriculum Customization Service (CCS) [2]. 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 [10]. 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 [12], and 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. [maull] with the CCS and Xu et al. [8] 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 sessionâĂŤand a session is delineated by 45 minutes of inactivity or termination of the session by logging out.

Include a diagram of the CCS here highlighting it's main features

4.1 Data

Our analysis in this research is based on click stream logs of educators' use of the CCS during the 2011-2012 school year. These data collected spans educators from six separate school district. 48,142 click actions were registered by 174 users across all six school districts. Table xx shows the break down of users per school district

Data for this analysis will be partitioned into three buckets: District 1 , District 2 and all other districts. The reason for this partition is that district 1 contains the majority of users in the system and users from district 2 do not have full access to all system features. Currently, users from dis-

Table 1: Number of users per district during the 2011-2012 school year

District	Number of users
District 1	80
District 2	41
District 3	9
District 4	11
District 5	7
District 6	27

Table 2: Data buckets for analysis

Bucket	Number of users
Bucket 1	80
Bucket 2	41
Bucket 3	54

trict 2 do not have access to publisher materials and thus their usage cannot be equally be categorized with usage from other school districts. All other school districts feature very small number of users. Having a small (n) makes it hard to properly run clustering algorithms to properly analyze system usage [clust]. With an input size of (n = 5 users) for instance, it is difficult for most clustering algorithms to find enough differences in the data input to perform correctly [cite]

Table xx illustrates the data division for each bucket:

5. HOW USAGE PATTERNS TREND WITHIN A USER TYPE

The goal of this study is to understand how usage patterns trend within a particular user type. We follow a similar approach to generating a user typology as Maull et al [maull]. For all users within a specific data bucket, we apply the expectation maximization (EM) clustering algorithm to generate a user typology of behaviors that are observable. These behaviors are characterized by five system features which cover both usage frequency and usage variety. These features are highlighted in table xx with a paranthesis to indicate the feature type i.e. frequency or variety below:

In prior computational approaches for generating user typologies, user types discovered via clustering are labeled in a manual ad-hoc fashion [maull][xu]. There are no set rules for mapping a cluster to a specific user type. In this work we introduce an equal frequency discretization framework for mapping clusters to a specific user type. Discretization frameworks have successfully been used in developing user type archetypes by [Brantegadzt] and Angeletou et al. [Angeletou] in characterizing the behaviors of users in an online discussion forum.

Table 3: 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 user-contributed, "My Stuff" content activity (variety)
4	Total activity within publisher material
5	Total Shared stuff activity
6	Total Interactive Resources activity

The discretization framework works as follows:

1. For each usage feature, we discretized usage into three bins: low, mid and high. We use the user types developed by Maull et al. [maull] as archetypes for characterizing usage. These features characterizing these user types are their discretized values are illustrated in table xx

2. Usage within each bucket is clustered to generate a typology of user behaviors

3. The usage features for each cluster is discretized into low, mid and high. These discretized features allow for mapping of user types to one of the typology archetypes.

4. Finally, for each user behavior, I'll produce a chart showing how the different usage features trend within the timeline of that behavior. Thus for a type power user characterized by the usage feature tuple of {hrs, sessions, ir_actv, my_stuff, shared_stuff, pub_stuff } with values {high, high, high, high, high, high}. I'll look to see what these values are like on a month to month basis.

I'll insert a figure showing the mapping process

insert table showing the discretized usage feature of each user type illustrated by Maull et al.

5.1 User typology of bucket A

After clustering usage for all users in bucket A during we get the following user typology and set of user types:

5.2 User typology of bucket b

5.3 User typology of bucket c

6. FREQUENT PATTERNS OF USE WITHIN A USER TYPE

This study involves performing a market basket analysis on the usage feature values that characterize a user type. This analysis would generate a set of association rules to illustrate the relationship between usage features that characterize a particular user type. The analysis is performed in the following steps:

1. For each member of a user type, the discretized usage feature values for each month of use are considered to be a single transaction in the form: {num_session, ir_actv, shared_stuff, my_stuff, pub_stuff }

2. Running the Apriori algorithm on all transactions within a particular user type, we discover the frequent patterns and association rules that occur within a specific type

7. 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:

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

8. LIMITATIONS

9. DISCUSSION AND CONCLUSION

10. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

11. ADDITIONAL AUTHORS

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

12. REFERENCES

[1] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[2] J. Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.

[3] M. Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.

[4] M. Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.

[5] L. Lamport. *LaTeX User's Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

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

[7] S. Salas and E. Hille. *Calculus: One and Several Variable*. John Wiley and Sons, New York, 1978.

[8] C. T. Sumner Tamara. Customizing science instruction with educational digital libraries. 2010.

12.1 References

Generated by bibtex from your .bib file. Run latex, then bibtex, then latex twice (to resolve references) to create the .bbl file. Insert that .bbl file into the .tex source file and comment out the command `\thebibliography`.

13. MORE HELP FOR THE HARDY

The acm_proc_article-sp document class file itself is chockfull of succinct and helpful comments. If you consider yourself a moderately experienced to expert user of LaTeX, you may find reading it useful but please remember not to change it.