RECOGNITION AND REPRESENTATION OF USER INTEREST

A Thesis

by

RAJIV RAVINDRANATH BADI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2005

Major Subject: Computer Science

RECOGNITION AND REPRESENTATION OF USER INTEREST

A Thesis

by

RAJIV RAVINDRANATH BADI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Frank Shipman
Committee Members,	Richard Furuta
	Takashi Yamauchi
Head of Department,	Valerie E. Taylor

December 2005

Major Subject: Computer Science

ABSTRACT

Recognition and Representation of User Interest. (December 2005)

Rajiv Ravindranath Badi, B.E., Bangalore University, Bangalore, India

Chair of Advisory Committee: Dr. Frank Shipman

With the growth of the internet and other media of communication, locating information on the topic of interest is less a problem of finding related documents than determining which particular documents are valuable. Often, the desired information is obscured within a long list of resources. Users become inundated with so much information that the task of sifting through it takes the majority of time on a given information task. Users look at multiple documents at once to find answers to their questions, and switch between documents to get the "complete" picture. New systems are needed that help users cull through related documents to gain the information they need.

As a part of the Document Triage Project, we have been looking at ways to help users in sifting through information. The Document Triage Project is developing tools to recognize, represent, communicate, and visualize user interest across applications. The topic of this thesis is recognizing user interest and providing an infrastructure to represent that interest so that it can be shared across the software applications involved in triage. Based on this inferred interest, applications can help users in their triage task by providing visualizations or other functionality. The applications could involve one or many reading interfaces (e.g., a browser, or an editor), an information organizing system (e.g., Visual Knowledge Builder) and search interfaces (the application providing the document collection; e.g., a search engine).

To recognize user interest, data is gathered from the user's reading, navigational and interpretive activities. Algorithms based on statistical models and qualitative analyses of user behavior in triage are used to infer interest. A light-weight infrastructure called Interest Profile Manager has been developed for the representation of interest values and the corresponding metadata. Interest Profile Manager also provides text processing capability, interest analysis functionality, sharing of data across applications and event propagation.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Frank Shipman for his guidance and support throughout my master's study. Working under him for the past one and a half years has been an exciting and fruitful experience. I would also like to thank Dr. Richard Furuta and Dr. Takashi Yamauchi for their support in my thesis.

I am grateful to all my friends in CSDL, especially Soonil Bae, for spending numerous hours in mentoring me and helping me in my work. Thanks goes to Hao-wei for helping me understand the inner workings of VKB, and to Micheal for the constructive comments, discussions and ideas all through my master's study. Thanks to Luis, Unmil and Neal for their patience, constructive criticism and knowledge sharing. Thanks to CSDL for the collaboration, camaraderie and caffeine.

I thank Ma and Papa for the undying love, unshakeable faith and unconditional support, and Rajani and Savji for their belief in me.

TABLE OF CONTENTS

CHAPTER Page

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

A. Problem

With the growth of the internet and other media of communication, locating information on the topic of interest is less a problem of finding related documents than determining which particular documents are valuable. Often, the desired information is obscured within a long list of resources. Users become inundated with so much information that the task of sifting through it takes the majority of time on a given information task. Users look at multiple documents at once to find answers to their questions, and switch between documents to get the "complete" picture. New systems are needed that help users cull through related documents to gain the information they need.

As a part of the Document Triage Project, we have been looking at ways to help users in sifting through information. The Document Triage Project is developing tools to recognize, represent, communicate, and visualize user interest across applications. The topic of this thesis is recognizing user interest and providing an infrastructure to represent that interest so that it can be shared across the software applications involved in triage. Based on this inferred interest, applications can help users in their triage task by providing visualizations or other functionality.

The journal model is *IEEE Transactions on Automatic Control.*

B. Background

Document Triage is the practice of quickly determining the usefulness and relevance of documents in a vast collection of documents (obtained from a search engine, for example). In triage, the attention of the user shifts from document to document to contextual overview (e.g. the list of search results, bookmarks, desktop or (another) visual overview of documents). Hence, document triage involves extensive reading (engagement with multiple documents at once) and hyper-extensive reading (engagement with subdocument components and fragmentary information), as opposed to intensive reading (engagement with a single document).

To support such activity, the Document Triage Project is investigating tools to support users in their triage tasks. In particular, the project is developing assistive software techniques based on inferred or expressed user interest. These techniques consist of four steps:

- **Recognizing User Interest and Document Value** - Systems can gain an understanding of user interest either explicitly or implicitly. Explicit interest indicators require the user to rate documents after reading/skimming them. It has been noted that with explicit ratings, users read a lot more articles than they rate [1]. Due to this and other shortcomings of explicit interest ratings which are discussed later, we chose to recognize user interest through a combination of inference based on implicit cues with occasional suggestions/questions designed to elicit explicit responses. Inferences of user interest in a document are based on user activity with the document, such as the time spent looking at a document, scrolling behavior, annotations, etc.

- **Representing User Interest** - Multiple software applications are involved while the user performs a triage task. These applications include the location/overview software, the reading software, and the software needed for the primary or motivating activity, such as writing a report, or making a decision based on some investigation. Each application can have its own methods for recognizing user interest but the representation of this interest must be shared across applications for the representation of interest to be helpful in providing support to the user's triage task as a whole.

- **Recognizing Documents of Interest** - Once software applications have a representation of user interest, they need to be able to identify documents that are related to this area of interest. Methods for determining when a document is related to an area of interest depend on the representation of interest. When an area of interest is represented as a set of documents or document components, or abstractions of these elements (e.g. term vectors, etc.), information retrieval algorithms can be used to recognize related documents.
- **Visualizing Interest Information** - Finally, after documents related to user interests are located, this must be conveyed to the user. The project will explore a variety of approaches to notifying users, including suggestion mechanisms and visualization techniques.

This thesis addresses recognition and representation of user interest (the first two steps) so as to help in the other two steps identified. Specifically, the focus is on recognizing user interest based on the user's behavior across multiple applications involved in the triage activity; all the applications collaborate to recognize the user's interest in a document. The applications could involve one or many reading interfaces (e.g., a browser, or an editor), an information organizing system (e.g., Visual Knowledge Builder (VKB) [2]) and search interfaces (the application providing the document collection; e.g., a search engine).

CHAPTER II

RELATED WORK

Much research has gone into gathering and recognizing user interest. The systems built to recognize user interest employ implicit indicators or explicit indicators, or a combination of both to recognize user interest. Explicit indicators (e.g. ratings), where users tell the system how interesting or uninteresting a given document is, is well-understood, easy to implement and fairly precise. However, stopping to enter explicit ratings can alter normal patterns of browsing and reading [3], and can lead to increased cognitive load on the user. Users may stop providing ratings when they perceive that there is no benefit in providing explicit ratings [4]. Research on the GroupLens system [1] found with explicit ratings that users rated much fewer documents than they read. Thus, even though explicit ratings are fairly precise in recognizing interest, their efficacy is limited due to the drawbacks listed above. Implicit indicators are a viable alternative. Nichols, in [5] identifies some implicit interest indicators and discusses the potential of implicit ratings. He states that "the limited evidence available suggests that implicit ratings have great potential, but their effectiveness remains unproven". The following subsections discuss approaches for recognizing user interest based on implicit interest indicators.

A. Reading time as the primary implicit interest indicator

Morita and Shinoda [6] study the issue of analyzing user interest on Usenet News Articles, and they conclude that only the time spent by the user in a document is an accurate indicator for interest. However, their analysis shows a very low correlation between the time taken to read the article and the length or the readability of the article. They mention that this probably indicates that not all articles are completely read, e.g. the decision of whether an article is interesting or not, especially in the not-interesting articles, is made by looking at the first couple of dozens of lines. This is an interesting observation especially in the context of document triage, where users rely on metadata or skimming to make judgments on the utility of documents.

Kim, Oard and Romanik [7] examine whether reading time is useful for predicting a user's preference for academic or professional journal articles, and whether printing behavior adds anything to what is already know about reading time. The paper concludes that users tend to spend a longer time reading relevant articles than non-relevant articles, although the threshold on reading time required to detect relevant documents would differ depending on the type of articles, e.g. news article, academic journal, professional journal, etc.

Kelly and Belkin [8] have analyzed the effectiveness of display time as a measure of user interest in a naturalistic study conducted over 14 weeks. They mention that the naturalistic study is more appropriate as it is generalized over contextual factors such as task, topic and collection. Their results demonstrate no general, direct relationship between display time and usefulness, and that display times differ significantly according to the specific task, and according to specific user.

B. Multiple implicit interest indicators

In Claypool et al. [3], the authors study the correlation between the user's explicit ratings with the time spent on a page, the time spent moving the mouse, the number of mouse clicks and the time spent scrolling. In the study, users provided explicit ratings for web pages as they browsed with Curious Browser. Curious Browser also collects data on implicit interest indicators as the users browse. The study shows that the time spent in reading and the time spent in scrolling are good indicators of interest, the number of mouse clicks is not a good indicator of interest and that there is a positive relationship between the time spent moving the mouse and the explicit rating. However, mouse movements alone appear useful only for determining which pages have the least amount of interest but are not accurate for distinguishing amongst higher levels of interest.

Goecks and Shavlik [9] design a neural network to learn user's interest unobtrusively from mouse clicks, scrolling behavior and the user's mouse activity. Although their study results indicate that the user's mouse activity may not correlate to the user's interest in the page (which is in contrast with the findings in [3]), it must be noted, as the authors mention, that their system's means of detecting mouse activity are not very complete.

Chan [10] introduces a metric for estimating interestingness of each visited page. A user profile is built of two components, a Web Access Graph and a Page Interest Estimator, is built. Interest is calculated as a function of (1) the frequency of page visits, (2) whether the page is bookmarked, (3) time spent on a page normalized by its length, (4) recency and (5) the number of links visited versus the number of links in the page. This work is geared towards recognizing the interest value of documents so as to build a “bag-of-words”, recommending documents deemed interesting and prefetching these documents based on the “bag-of-words”. The system evaluation for this approach predicts interest with an average accuracy of 70%. This shows that the weighted measure of interest is effective for calculating interest.

C. User annotations as interest indicators

Shipman et al. [11] discuss additions to the XLibris system to aid users while rereading documents by providing visualizations based on user annotations. The system using a mark parser identifies area highlights, area circling, underlining, use of margin bars, highlighting, comments, etc. The system assigns a higher emphasis value to interpretive marks (comments, symbols and callouts). Even though the recall and precision rates are low, the paper differentiates interpretive tasks from normal reading and navigational tasks.

All the related work mentioned above is based on a single application model, i.e. the research has focused on a single reading interface such as a web browser, document editor such as emacs, etc. As mentioned earlier, the goal of this thesis is to recognize and represent user interest in the context of document triage. The user's reading, navigation and interpretive activities combined with the document style attributes are considered for recognition of user interest. With multiple applications, a shared environment is required to represent interest and the related meta-information. The representation is to be such that all the applications can read/modify/delete the stored interest (or meta-information). The system infrastructure developed for representing/sharing interest across applications is the Interest Profile Manager. Our approach to developing the models of user interest is described next. This is followed by a description of the Interest Profile Manager.

CHAPTER III

INTEREST MODELS

The first step in recognizing user interest is to gather data from the applications involved in the triage activity. Various parameters are recorded from the document style information and the document reading and the document interpreting activities. The data could be based on the usage history or on the current activities of the user. Mathematical (including statistical) models are applied on the collected data to gather user's interest on specific web pages. Each one of the models considered makes use of only a part of the collected data. The next subsection details the document and interaction parameters gathered from Internet Explorer (IE) and Visual Knowledge Builder (VKB). Following this is a description of the mathematical models used to recognize user interest based on the collected data.

A. Data collection

User interest can be calculated after unobtrusively gathering some or all of the following data from a reading interface (e.g., IE) and organizing interface (e.g. VKB). Data gathered includes characteristics of the document, data coming from the document reading activity, and data coming from the document interpretation activity:

- Document Style Information
  - Number of pages
  - Number of characters
  - Number of words
- Document Reading Activity

  - Time spent in a document
  - Time spent moving the mouse
  - Mouse clicks (on links or otherwise)
  - Scrolling behavior
    - Direction of scrolling from the time the document is opened till the time it is closed
    - Speed of scrolling
    - Total time spent scrolling
    - Scroll offset (positive or negative)
    - Total number of scroll groups
  - Frequency of access of the document

- Document Interpretation Activity
  - Annotation in the document
  - Bookmarking
  - Printing
  - Highlighting

Document style information consists of a set of attributes of the documents themselves, rather than users' interactions with the document. This thesis primarily considers document length (e.g. number of pages, number of characters, number of words), although many other attributes of documents could be considered.

Document reading activity includes user actions during passive reading. Information included consists of time spent on a document, time spent moving the mouse around on the document, number of mouse clicks, characteristics of the user's scrolling behavior, and frequency of document access.

Document interpretation activity refers to user actions that leave a lasting record of user interest in the document, such as annotating, bookmarking, or printing the document. Annotation can be in the form of highlighting, underlining, use of margin bars, circling, use of comments, use of symbols, use of callouts, change of colors, etc. Document interpretation activities explicitly indicate that the user has encountered interesting sections of a document. Document interpretation data is collected not just for the reading interface, but also for the spatial hypertext tool - Visual Knowledge Builder (VKB). Even though annotation on paper documents could be considered as a good indication of interest, its implementation requires scanning annotated paper documents on computers. Although pen based computers such as Tablet PCs can mimic the annotation capabilities of pen and paper, pilot studies in the document triage project indicated that people are uncomfortable using a Tablet PC for reading and annotation due to frustration with the control of standard interfaces, e.g. scrollbars, using a pen. Hence, the focus of this thesis will be on interpretive tasks on a *conventional* computer such as a desktop or a laptop. User activity from Visual Knowledge Builder (VKB) gathered includes:

- Object edit (text, color, border thickness, size, etc.)
- Object move/scroll
- Object text analysis
- Clicking/Following links
- Workspace organization

B. Mathematical models

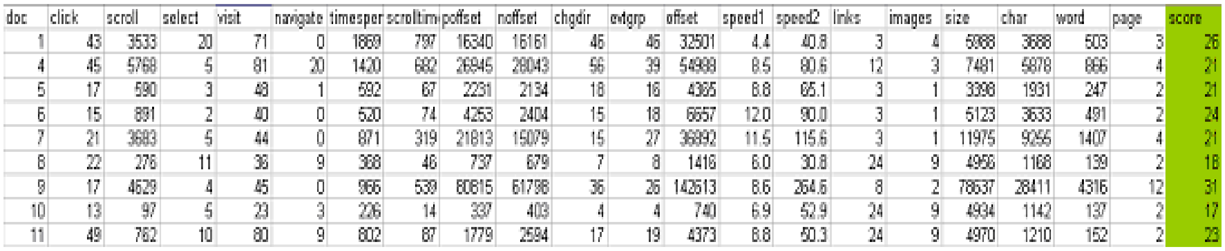
Scores are assigned to documents based on mathematical models based on the data described above. The following subsections discuss three models which can be applied to recognize user interest. These models have been arrived at from an analysis of the data gathered from a Document Triage study carried out in Fall 2004 [12].

The first model is based on a factor analysis of user activity in the reading interface (IE) and document attributes when compared to explicit user assessment of documents. The second model is the result of user activity in the reading interface and document attributes, but employing multiple regression to arrive at the model. The final model was developed through a combined quantitative and qualitative analysis of the data from prior triage studies.

1. Reading-activity based model

The reading-activity based model takes data available to the reading application (IE) from earlier studies on the use of the Visual Knowledge Builder (VKB) and Internet Explorer during a document triage activity. Following the activity, users were asked to identify the five most valuable and five least valuable documents. The factor-analysis based model was developed using correlations between perceived value and user activity and document style.

Figure 1 shows a snippet of the raw data from the study [12]. The raw data contains document style and user reading data from 24 subjects across 35 web pages. The 20 variables shown in the table are: number of clicks, number of scrolls events, number of text selections, number of visits to the web page, number of navigate events, total time spent in the web page, total time spent scrolling, total positive scrolling offset, total negative scrolling offset, total number of times the user changed



doc	click	scroll	select	visit	navigate	timesper	scrolltim	poffset	noffset	chgdir	evtgrp	offset	speed1	speed2	links	images	size	char	word	page	score
1	43	3533	20	71	0	1869	797	16340	16161	46	46	32501	4.4	40.8	3	4	5988	3688	503	3	26
4	45	5768	5	81	20	1420	682	26945	28043	56	39	54988	8.5	80.6	12	3	7481	5878	866	4	21
5	17	590	3	48	1	592	67	2231	2134	18	16	4365	6.8	65.1	3	1	3398	1931	247	2	21
6	15	891	2	40	0	520	74	4253	2404	15	18	6657	12.0	90.0	3	1	5123	3633	491	2	24
7	21	3683	5	44	0	871	319	21813	15079	15	27	36892	11.5	115.6	3	1	11975	9256	1407	4	21
8	22	276	11	36	9	368	46	737	679	7	8	1416	6.0	30.8	24	9	4956	1168	139	2	18
9	17	4629	4	45	0	966	539	80815	61798	36	26	142613	8.6	264.6	8	2	78637	28411	4316	12	31
10	13	97	5	23	3	226	14	337	403	4	4	740	6.9	52.9	24	9	4934	1142	137	2	17
11	49	762	10	80	9	802	87	1779	2594	17	19	4373	6.8	50.3	24	9	4970	1210	152	2	23

Fig. 1. Snippet of raw data

scrolling direction, total scrolling offset, two types of scrolling speed, number of links in the web page, number of images in the web page, file size of the web page, number of characters in the web page, number of words in the web page, and number of pages in the web page.

Principal factor analysis is used to simplify the dataset. Based on the correlation matrix, some variables are excluded for which the majority of significance values are greater than 0.05 and the correlation coefficients are greater than 0.9 (20 variables ⇒ 8 variables). Using principal axis factoring, the following table is obtained.

Two factors whose eigenvalues are greater than 1 are extracted. The eigenvalues associated with each factor show the variance explained by that factor. In Table I, two factors explain 82% of total variance.

The rotated factor matrix (Table II) is a matrix of the factor loadings for each variable onto each factor (factor loadings less than 0.4 have not been displayed). Three variables substantially affect both factors, which demonstrates that factors are not clear-cut in terms of variables. Factor 2 is significantly influenced by scrolling offset and document length in words, while Factor 1 is influenced by other user events and time spent.

From the factor score coefficient of Table III, the following factor scores are calculated:

Table I. Total variance explained (reading-activity based model)

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cum. %	Total	% of Variance	Cum. %	Total	% of Variance	Cum. %
1	4.693	58.668	58.668	4.508	56.345	56.345	3.569	44.613	44.613
2	1.859	23.239	81.908	1.574	19.681	76.026	2.513	31.413	76.026
3	.580	7.255	89.163						
4	.418	5.223	94.386						
5	.210	2.628	97.014						
6	.119	1.491	98.506						
7	.074	.923	99.428						
8	.046	.572	100.00						

Factor 1 = 0.289 * (# of clicks) + 0.063 * (# of scrolls) + 0.076 * (# of selections) + 0.110 * (# of visits) + 0.556 * (total time spent) - 0.331 * (scrolling offset) + 0.138 * (# of changes in scrolling direction) - 0.049 * (# of words)

Factor 2 = - 0.143 * (# of clicks) + 0.147 * (# of scrolls) - 0.043 * (# of selections) - 0.033 * (# of visits) - 0.053 * (total time spent) + 0.848 * (scrolling offset) + 0.045 * (# of changes in scrolling direction) + 0.077 * (# of words)

Instead of individual user events and document style attributes, the above given

Table II. Rotated factor matrix (reading-activity based model) a. Extraction method: Principal Axis factoring. b. Rotation method: Varimax with Kaiser normalization

	Factor	
	1	2
# of Clicks	.913	
Time spent	.878	.407
# of Visits	.792	
Change direction	.723	.554
# of Selections	.695	
Scrolling offset		.961
# of Scrolls	.545	.746
# of Words		.717

factors are used in regression. Statistical models are built around the factors. A multiple regression model is of the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon_i \tag{3.1}$$

Y: outcome variable

β_0 : constant

β_1 : coefficient of i^{th} predictor

X_i: i^{th} predictor

ϵ_i: difference between the predicted and observed value of Y

Applying the resulting model to the study, Y is the document score, and X_is are

Table III. Factor score coefficient matrix (reading-activity based model) a. Extraction method: Principal Axis factoring. b. Rotation method: Varimax with Kaiser normalization. c. Factor scores method: Anderson-Rubin

	Factor	
	1	2
# of Clicks	.289	−.143
# of Scrolls	.063	.147
# of Selections	.076	−.043
# of Visits	.110	−.033
Time Spent	.556	−.053
Scrolling offset	−.331	.848
Change Direction	.138	.045
# of Words	−.049	.077

user events, document style attributes and factors. The β values indicate to what degree each predictor affects the outcome if the effects of all other predictors are held constant. User events and document style attributes could be considered as predictors, but many of them are significantly correlated to each other, which violates the basic assumption of multiple regression: more predictors do not guarantee more accurate estimation. To build a model from the two factors, the Hierarchical (Blockwise entry) method is used. Table IV gives the summary of the model with the first row providing results when only the first factor is included and the second row includes both factors.

R in Table IV measures the values of the multiple correlation coefficients between the predictors (factors) and document score. R^2 is a measure of how much of the variability in the outcome is accounted for by the predictors. With a single factor,

Table IV. Factor analysis based model summary a. Predictors: (constant), A-R factor score 2 for analysis 1 b. Predictors: (constant), A-R factor score 2 for analysis 1 , A-R factor score 1 for analysis 1 c. Dependent variable: Document score

Factors in Model	R	R^2	Adj. R^2	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R^2 Change	F Change	df1	df2	Sig. F Change	
1	.588	.346	.326	5.049	.346	17.437	1	33	.000	
2	.691	.477	.445	4.582	.132	8.066	1	32	.008	1.853

the model accounts for 34.6% of the variation in document score, while including the second factor increases the variability accounted for to 44.5%. The adjusted R^2 shows how well this model generalizes. This value indicates that the cross-validity of this model is good. The Durbin-Watson statistic indicates whether the assumption of independent errors is tenable. The closer to 2 that the value is, the better: values less than 1 or greater than 3 are definitely problematic. 1.853 in this model is reasonably close to 2, and seems to be acceptable.

The sum of squares for the model (Regression in Table V) represents the improvement in prediction resulting from fitting a regression line to the data rather than using the mean as an estimate of the outcome. The residual sum of squares (Residual in Table V) represents the total difference between the model and the observed data. If the improvement due to fitting the regression model is much greater than the inaccuracy within the model then the value of F is greater than 1. When including both factors, the F-ratio is 14.618, which is highly significant ($p < 0.001$).

Table V. ANOVA (reading-activity based model) a. Predictors: (constant), A-R factor score 2 for analysis 1 b. Predictors: (constant), A-R factor score 2 for analysis 1 , A-R factor score 1 for analysis 1 c. Dependent variable: Document score

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	444.549	1	444.549	17.437	.000
	Residual	841.336	33	25.495		
	Total	1285.886	34			
2	Regression	613.929	2	306.965	14.618	.000
	Residual	671.956	32	20.999		
	Total	1285.886	34			

Each of the β values has an associated standard error showing to what extent these values would vary across different samples and whether or not the β values are significantly different from zero. The t-test associated with β values shows that the predictors are making a significant contribution to the model ($p < 0.0001$). From Table VI, the resulting model is as following:

Accumulated Document Score $= \beta_0 + \beta_1 * Factor1 + \beta_2 * Factor2$

Accumulated Document Score $= 21.057 + (2.232 * Factor1) + (3.616 * Factor2)$

In Table VI, the tolerance value is much greater than 0.2, and the average VIF is 1, which confirms that collinearity is not a problem for this model. The above given formula will be used to calculate user interest using this model.

Table VI. Coefficients (reading-activity based model) a. Dependent variable: Document score

Model		Unstandardized Coefficients		Std. Coefficients	t	Sig.	95% Conf. Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero Order	Partial	Part	Tolerance	VIF
1	(Constant)	21.057	.853		24.672	.000	19.321	22.794					
	A-R factor score 2 for Analysis 1	3.616	.866	.588	4.176	.000	1.854	5.378	.588	.588	.588	1.000	1.000
2	(Constant)	21.057	.775		27.186	.000	19.479	22.635					
	A-R factor score 2 for Analysis 1	3.616	.786	.588	4.601	.000	2.015	5.217	.588	.631	.588	1.000	1.000
	A-R factor score 1 for Analysis 1	2.232	.786	.363	2.840	.008	.631	3.833	.363	.449	.363	1.000	1.000

2. Limited collinearity reading-activity based model

Table VII. Limited collinearity reading-activity based model summary a. Predictors: (constant), scrolling offset b. Predictors: (constant), scrolling offset, # of visits c. Dependent variable: Document score

Model	R	R^2	Adj. R^2	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R^2 Change	F Change	df1	df2	Sig. F Change	
1	.643	.413	.395	4.783	.413	23.206	1	33	.000	
2	.727	.528	.499	4.355	.115	7.808	1	32	.009	1.963

Table VIII. ANOVA (limited collinearity reading-activity based model) a. Predictors: (constant), scrolling offset b. Predictors: (constant), scrolling offset, # of visits c. Dependent variable: Document score

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	530.914	1	530.914	23.206	.000
	Residual	754.972	33	22.878		
	Total	1285.886	34			
2	Regression	678.994	2	339.497	17.901	.000
	Residual	606.892	32	18.965		
	Total	1285.886	34			

The second model was developed using the same techniques as the first model but allowing for the resulting factors to result in some collinearity. As with the first reading-activity based model, the factor-analysis based model was developed using correlations between perceived value and user activity and document style. Tables VII, VIII, and IX show the relative performance and correlation of the one and two factor models allowing for some collinearity.

Accumulated Document Score $= \beta_0 + \beta_1 * (ScrollingOffset) + \beta_2 * (\#ofvisits)$

Accumulated Document Score $= 11.937 + 0.00006663 * (scrollingoffset) + 0.14 * (\#ofvisits)$

The resulting model is a little more powerful (the adjusted R^2 of 0.499) than the previous model (the adjusted R^2 of 0.445). However, the collinearity statistics and Durbin-Watson statistics indicate that this model is relatively more prone to the collinearity problem, although it is still within an acceptable range.

Table IX. Coefficients (limited collinearity reading-activity based model) a. Dependent variable: Document score

Model		Unstandardized Coefficients		Std. Coefficients	t	Sig.	95% Conf. Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero Order	Partial	Part	Tolerance	VIF
1	(Constant)	17.948	1.035		17.349	.000	15.843	20.053					
	Scrolling offset	8.065$E-$05	.00	.643	4.817	.000	.000	.000	.643	.643	.643	1.000	1.000
2	(Constant)	11.937	2.348		5.083	.000	7.154	16.720					
	Scrolling offset	6.663$E-$05	.000	.531	4.152	.000	.000	.000	.643	.592	.504	.902	1.108
	# of Visits	.140	.050	.357	2.794	.009	.038	.243	.523	.443	.339	.902	1.108

3. Qualitatively-defined model

The previous two models are based on the statistical analysis of prior study data. A third model was defined based on a combination of qualitative and quantitative assessment of previous study data. Further discussions in this subsection give the analysis behind the weight assignments in the model.

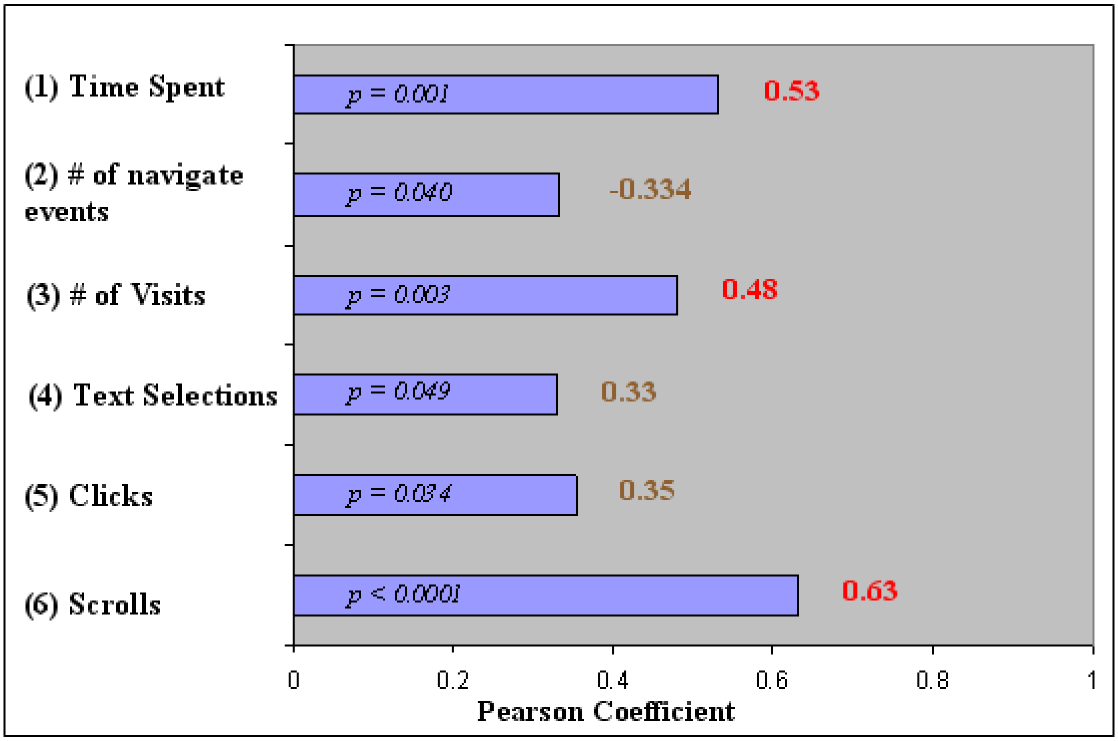


Fig. 2. Correlation between user behavior and user interest

Figure 2 shows a plot of Pearson Coefficient versus various user activities in the reading interface (i.e., Internet Explorer). The Pearson coefficients of variables with high values (and with p values < 0.005) are shown in red. From the figure we derive that time spent, # of visits and scroll events are positively correlated to user interest. The graph shows that users tend to spend more time on documents they find interesting. Similarly, their scrolling activity is higher in documents of interest. Hence, the 3 variables are assigned large weights of 12, 10 and 15 respectively.

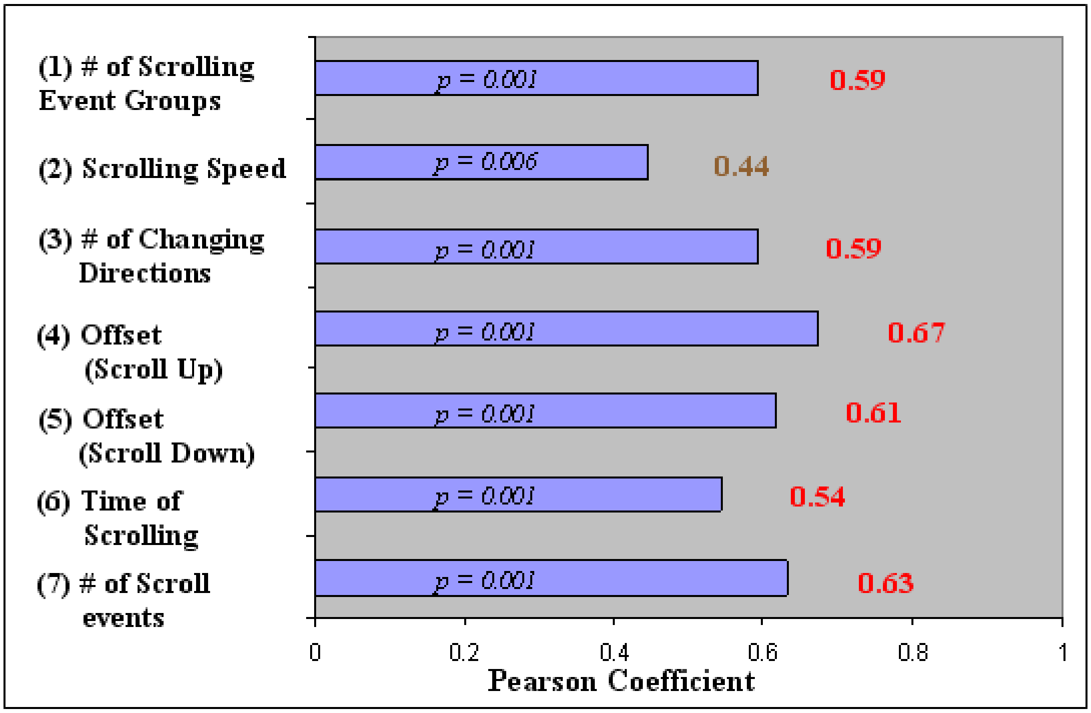


Fig. 3. Scroll parameters vs. user interest

Scrolling activity can be further broken down into fine-grained variables. These are shown in Figure 3. The figure shows that many scrolling parameters are positively correlated to user interest. The weights in the table for the scroll parameters have been arrived at from the graph. Figure 4 shows the correlation between various document style parameters and user interest. The values do not seem to be very significant. However, from the interviews carried out with subjects in the Oct 2004 Document Triage study, it has been found that users consider the amount of content in a document to be an important criterion while classifying documents. Some users prefer denser documents (more text) while other users prefer sparser documents (less text). Considering these factors, the variable corresponding to number of characters is assigned a weight of 4. Number of words and pages in a document are strongly related to the number of characters. Hence these variables have not been included in the model. The other interesting document style variable is the number of links per page. This is the only parameter significantly correlated to document score among per page parameters.

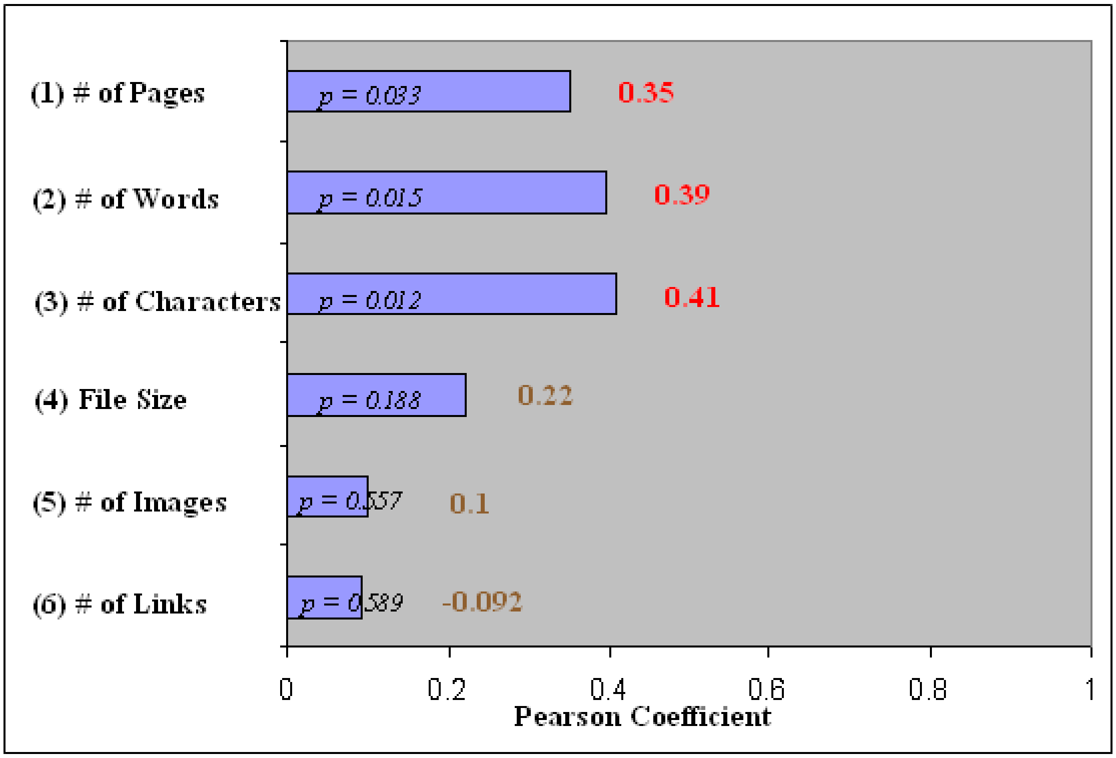


Fig. 4. Document style attributes vs. user interest

Other user activities such as printing and bookmarking are stronger indicators of user interest. These events cannot be caught without tweaking the internals of Internet Explorer. It seems natural to assume that documents that are bookmarked or printed are of strong interest to the user. Due to the non-availability of IE source code, and since users do not print and bookmark webpages frequently, these two parameters have been disregarded for analysis.

Table X. Weights for qualitatively-defined model

	Parameter	Weight
1	Time spent on page	12
2	Number of visits to document	10
3	Scroll: Scroll Events	15
4	Scroll: Scroll Up	12
5	Scroll: Scroll Down	10
6	Scroll: # of Scroll Event groups	8
7	Scroll: Scrolling speed	5
8	Scroll: # of instances of changing direction	8
9	Scroll: Time of scroll	6
10	VKB Edit event: Object Resize	2
11	VKB Edit event: Object Border Color Change	2
12	VKB Edit event: Object Move	2
13	VKB Edit event: Object Color Change	3
14	VKB Edit event: Object Text Edit	3
15	VKB Edit event: Object Border Thickness Change	3
16	Document Style: # of characters	4
17	Document Style: # of links per page	-5
	Total	100
18	VKB Edit event: Object Delete	Assign the object a score of 0

The table also presents various VKB edit events. These weights are lower and have been arrived at from qualitative assessments of user behavior on VKB. Deletion of VKB objects is a strong negative indicator of user interest. Hence, a score of 0 is assigned to the link object when it is deleted.

Table X summarizes this model. The weight field in the table is the degree to which the particular parameter contributes to the overall document score. Additionally, this model assigns the document a value of 0 (lowest possible) if the document is deleted in VKB.

4. Limitations of models

These three models were created based on prior study data from the same user activity with a variety of different hardware and software configurations. The first two models use statistical analysis of reading behavior to infer document value. The third model is a first pass at a model combining reading activity, organizing activity, and document characteristics, although it is still heavily weighted towards reading activity. Thus, these three models must be viewed as first steps towards the multi-application model intended. Their success, or lack thereof, will aid the design of the next generation of models.

CHAPTER IV

SYSTEM BUILDING

A. Interest Profile Manager

The Interest Profile Manager is a light-weight infrastructure which can be used to store data deemed *interesting* by applications. *Store* here refers to both in-memory storage and persistent storage. Applications can share data by means of the Interest Profile Manager. The design for the Interest Profile Manager is sufficiently flexible to be able to capture the kinds of interest that readers express through their actions (like scrolling behavior, annotation, etc.). It can be used to store state information as well. The Interest Profile Manager is designed to be independent of the application that uses it. Since the Interest Profile Manager is developed as a library, applications can use it as an "add-in", and make use of the interfaces provided by the Interest Profile Manager for run-time/persistent storage of data. The Interest Profile Manager provides storage/retrieval of data without involving the overhead of a Database. The Interest Profile Manager can store the run-time data structure as an XML file, and also create a run-time data structure from a specified XML file.

The design goals of Interest Profile Manager are listed hereunder:

- **Storage and retrieval of interest:** It allows storage and retrieval for reading data, navigational data and user annotation data.

- **Text processing capability:** Interfaces for manipulating term vectors, and performing operations on them such as getting the terms with highest frequency, removing stop words from the term vector, etc.

- **Interest Assessment functions:** A host of functions to help applications on interest assessment from the data (e.g. get interest value of the term “hypertext” for all “Green” objects in VKB.)

- **Storage of *other* data:** Allow applications to store more than just interest values in the Interest Profile Manager. This could mean metadata associated with interests, or other bookkeeping data.

- **Mechanism to assign belief values:** Applications can assign belief values (the level of *confidence* an application attaches to the interest value) for the inferred interest. e.g., user bookmarking a page can be inferred by the browser to be a sure sign of the user’s interest in the page, and hence the belief value associated with the interest is high.

- **Support for sharing data:** Applications involved in triage activity need to know the user’s *interest* data gathered from other applications. That is, if MS Word and Internet Explorer (IE) are being used in a triage task, IE should be able to read/write the attributes created by MS Word and vice versa. In other words, Interest Profile Manager provides a framework for sharing of interests across applications in a hub-and-spoke model.

- **Event notification:** Mechanism to notify the applications when changes have been made to attributes in its storage. When one application writes *interest* information into the Interest Profile Manager, other applications (stakeholders) are notified of the *write* event.

The design for Interest Profile Manager allows for sharing of data by applications across different machines. Figure 5 shows a scenario in which Interest Profile Manager can be used.

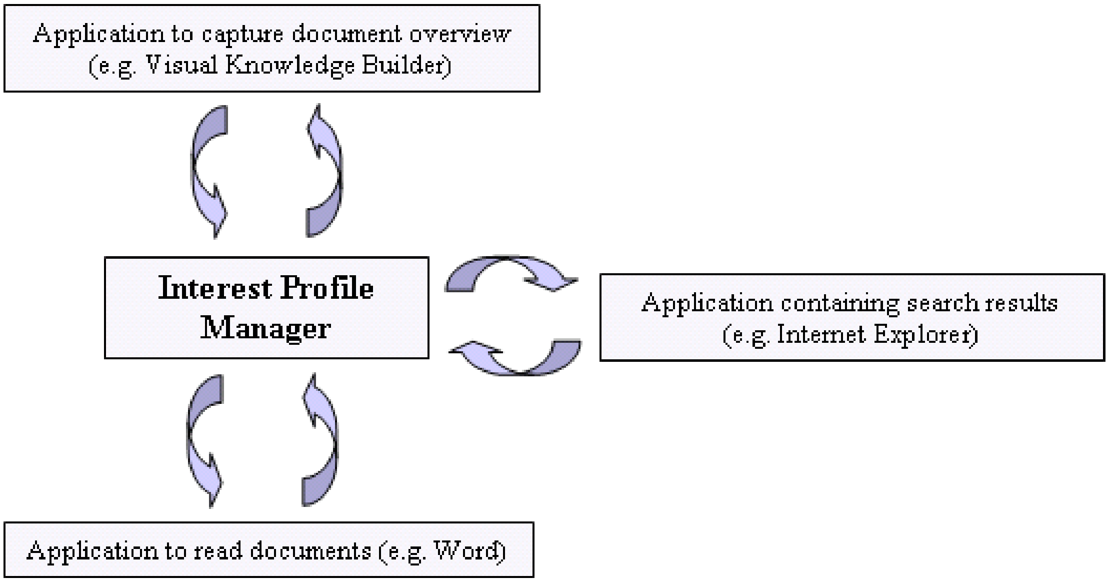

Application to capture document overview (e.g. Visual Knowledge Builder)
Interest Profile Manager
Application containing search results (e.g. Internet Explorer)
Application to read documents (e.g. Word)

Fig. 5. Anticipated usage scenario of Interest Profile Manager

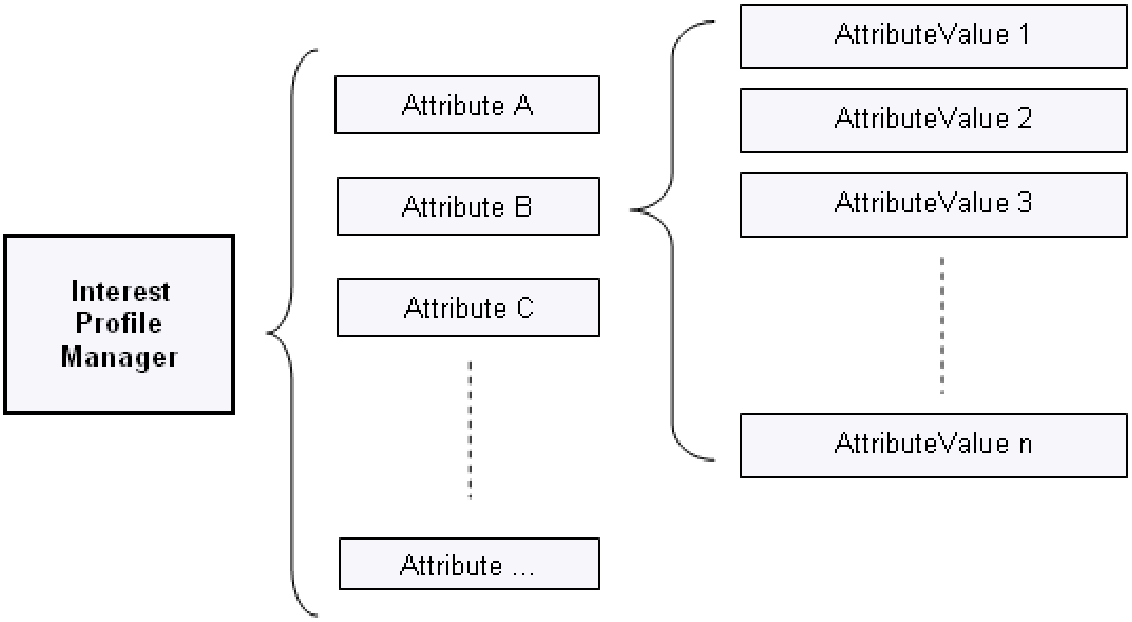


Fig. 6. High-level structure of Interest Profile Manager

To meet the design goals stated in this section, the Interest Profile Manager is designed to be a collection of Attributes, which, in turn, store AttributeValues (Figure 6). Apart from AttributeValues, an Attribute also stores the interest value, the belief value and timestamp (can help applications make observations based on the timeline of user interests). Applications can add, delete, update and retrieve Attributes and AttributeValues. AttributeValues have a provision wherein applications can specify the action to be taken on the AttributeValue. The Interest Profile Manager runs as a separate process and acts as the server. Applications connect to the server as Interest Profile clients. Clients can register to receive notifications of changes from the server. The Interest Profile client and server are connected through two TCP/IP data streams. One of the streams is used for data communication (using XML) and the other is used for asynchronous event notification. Many clients can connect to the Interest Profile Manager and effect changes simultaneously. Data integrity in the Interest Profile Manager is preserved even when multiple clients (threads) perform edit operations by using synchronized methods. Monitors can be automatically set on the server data in the Java Virtual Machine using these synchronized methods. All the client facing data editing APIs are declared as synchronized.

The Attribute class contains name, id, ControlString, belief, timestamp and interest as variables along with a Vector of AttributeValues. ControlString is a string which could be used by applications to store complex data which would otherwise be difficult to store as a collection of AttributeValues in an Attribute object. The ControlString is for future proofing the Attribute class. Figure 7 shows a sample XML file with Attributes and AttributeValues for a scenario containing a user's book interests.

AttributeValue class contains three variables namely Data, Action and Value. The variables data and value are used to store key and value pairs. The Action variable is for applications to set flags which would be required for processing the data stored in the AttributeValue. The flags could take values "TAKE_ACTION", "NO_ACTION", etc.

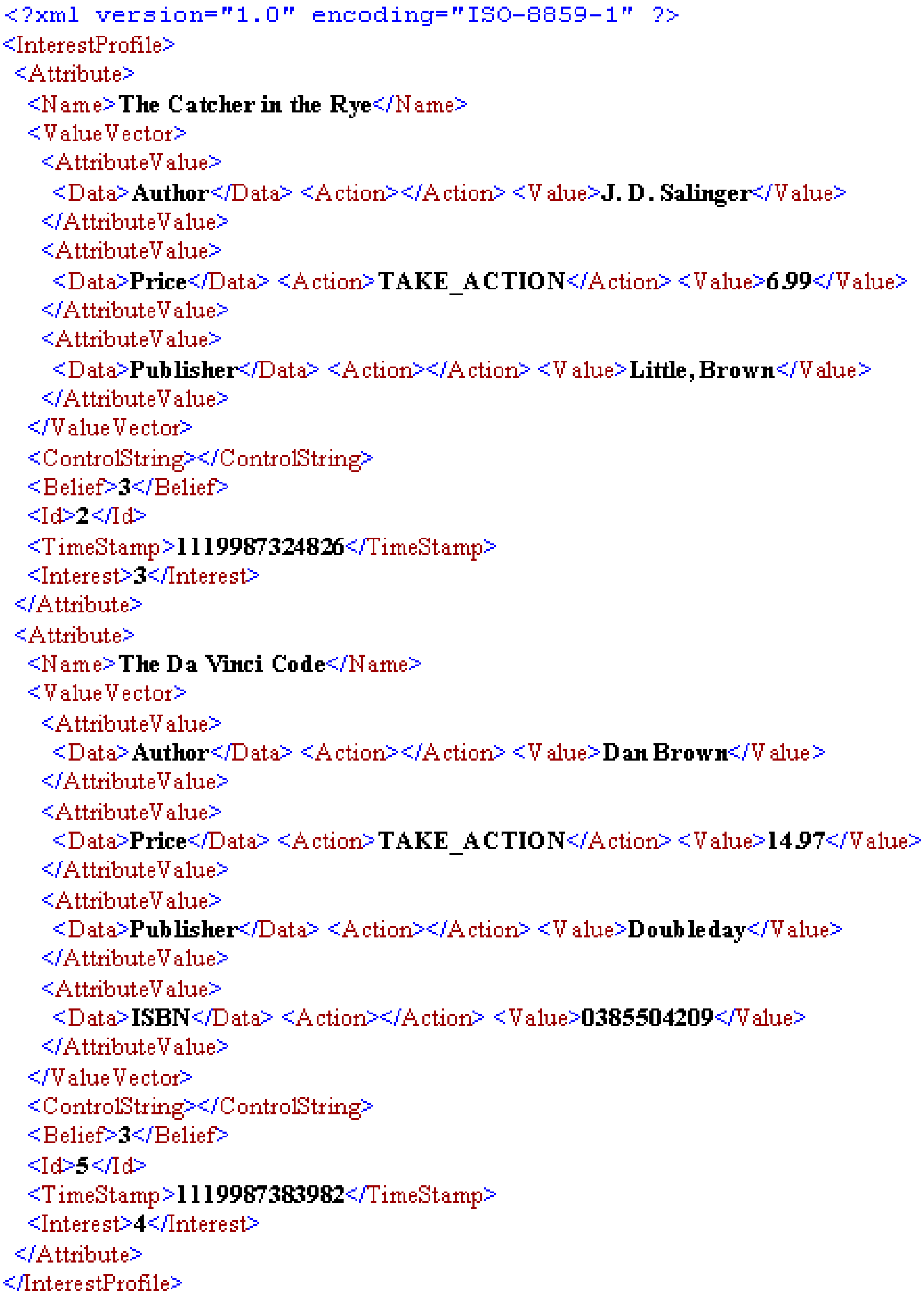
The Interest Profile Manager provides many methods for the operations mentioned in the design goals, namely:

- Overloaded methods to add Attributes
- Overloaded methods to remove single, multiple and all Attributes
- Methods for updating Attributes. Updating methods are provided for replacement, merging, updating AttributeValues, and updating and adding AttributeValues
- APIs to get high term frequency terms from Attributes, frequency calculation, term frequency, term vector
- Server APIs to add Event Handlers
- APIs to load and save XML file

Interest Profile Clients provide similar methods for effecting changes on the server. In addition, they provide certain other methods required to connect to the server, check connection to the server and stop the server.

B. Sample setup

The setup shown in Figure 8 shows a Document Triage scenario using the Interest Profile Manager. This setup is similar to Figure 5 in principle, but provides more detail. VKB and IE are the two applications communicating with Interest Profile Manager as clients. In the triage activity, VKB is the tool which allows the user



```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<InterestProfile>
 <Attribute>
  <Name>The Catcher in the Rye</Name>
  <ValueVector>
   <AttributeValue>
    <Data>Author</Data> <Action></Action> <Value>J. D. Salinger</Value>
   </AttributeValue>
   <AttributeValue>
    <Data>Price</Data> <Action>TAKE_ACTION</Action> <Value>6.99</Value>
   </AttributeValue>
   <AttributeValue>
    <Data>Publisher</Data> <Action></Action> <Value>Little, Brown</Value>
   </AttributeValue>
  </ValueVector>
  <ControlString></ControlString>
  <Belief>3</Belief>
  <Id>2</Id>
  <TimeStamp>1119987324826</TimeStamp>
  <Interest>3</Interest>
 </Attribute>
 <Attribute>
  <Name>The Da Vinci Code</Name>
  <ValueVector>
   <AttributeValue>
    <Data>Author</Data> <Action></Action> <Value>Dan Brown</Value>
   </AttributeValue>
   <AttributeValue>
    <Data>Price</Data> <Action>TAKE_ACTION</Action> <Value>14.97</Value>
   </AttributeValue>
   <AttributeValue>
    <Data>Publisher</Data> <Action></Action> <Value>Doubleday</Value>
   </AttributeValue>
   <AttributeValue>
    <Data>ISBN</Data> <Action></Action> <Value>0385504209</Value>
   </AttributeValue>
  </ValueVector>
  <ControlString></ControlString>
  <Belief>3</Belief>
  <Id>5</Id>
  <TimeStamp>1119987383982</TimeStamp>
  <Interest>4</Interest>
 </Attribute>
</InterestProfile>
```

Fig. 7. Sample XML file generated by Interest Profile Manager

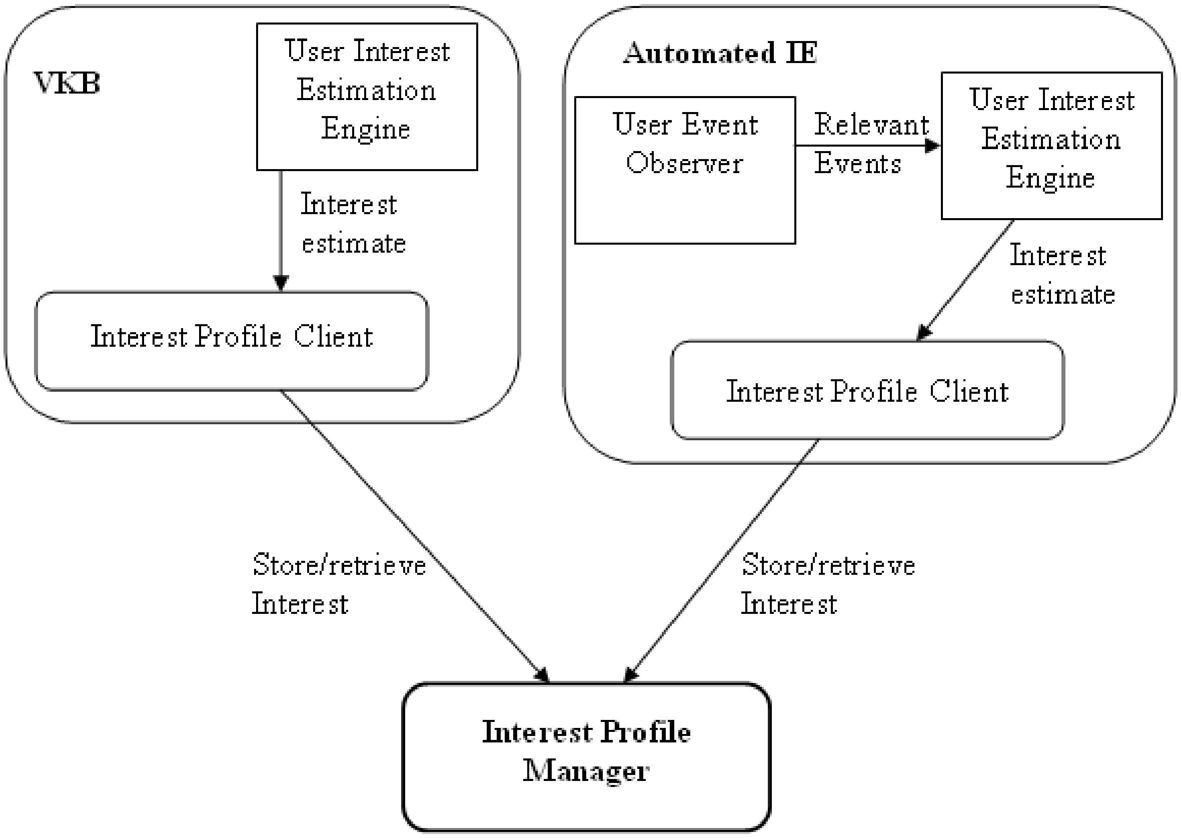


Fig. 8. Sample setup

to store, manage and organize links. In the VKB space, the user can annotate and associate metadata with the collections and objects containing links. Automated IE forms the reading interface in the triage task. Automated IE here refers to the enhanced version of Internet Explorer capable of gathering events on user activity. Both applications contain the Interest Profile Client library which allows them to interface with the Interest Profile Manager (Server). The User Interest Estimation Engine is the module in VKB and Automated IE to estimate user interest. This engine could use any of the models discussed in the previous section.

C. Interest estimation using data from VKB

Based on the criteria for evaluating user interest (as elaborated in the section "Interest Models" above), the User Interest Inference Engine identifies the text/objects of interest in the VKB work space. The User Interest Inference Engine also handles the task of gathering the attributes of objects (color, border thickness, etc.) along with the text contained in the objects. Belief values are assigned to these and then sent to the Interest Profile Manager for storage. When sufficient information about the workspace is stored in Interest Profile Manager, VKB can query for "analysis" on the data stored. VKB could query for the common text in all blue objects, for example. Interest Profile Manager, based on text analysis of the objects, returns the ordered list of common terms for the blue objects. The same information can be queried by Internet Explorer as well, as Interest Profile Manager is common to all the applications. Now, using the queried data, Internet Explorer could show a useful visualization to the user. For example, when the user visits a link contained in one of the blue objects, Internet Explorer could highlight the common terms in the document. Or, in an alternate scenario, Internet Explorer could scroll to the part of the webpage with the most occurrences of the terms, immediately after the page is loaded.

D. Interest estimation using data from Internet Explorer

Suppose the user visits a link in the VKB workspace and, based on reading activity, Internet Explorer infers that the user is interested in the contents on the page. Internet Explorer stores this information in the Interest Profile Manager. When the user navigates back to the VKB workspace, VKB now knows that the user is interested in the link she visited. Using this piece of useful information, VKB could change the display attribute of the object containing the link. For example, a halo around the object may be added to reflect the user interest value. This way, the user knows the pages of interest to her without taking explicit notes. When she opens the same VKB workspace after a gap of a few days, she can get started from where she left without having to revisit the previous links. She can selectively visit only the links which are *marked interesting* and ignore the ones which are *not marked as interesting*.

Note that the focus of this thesis is on recognizing and representing user interest. Visualizations have been mentioned only to provide context.

CHAPTER V

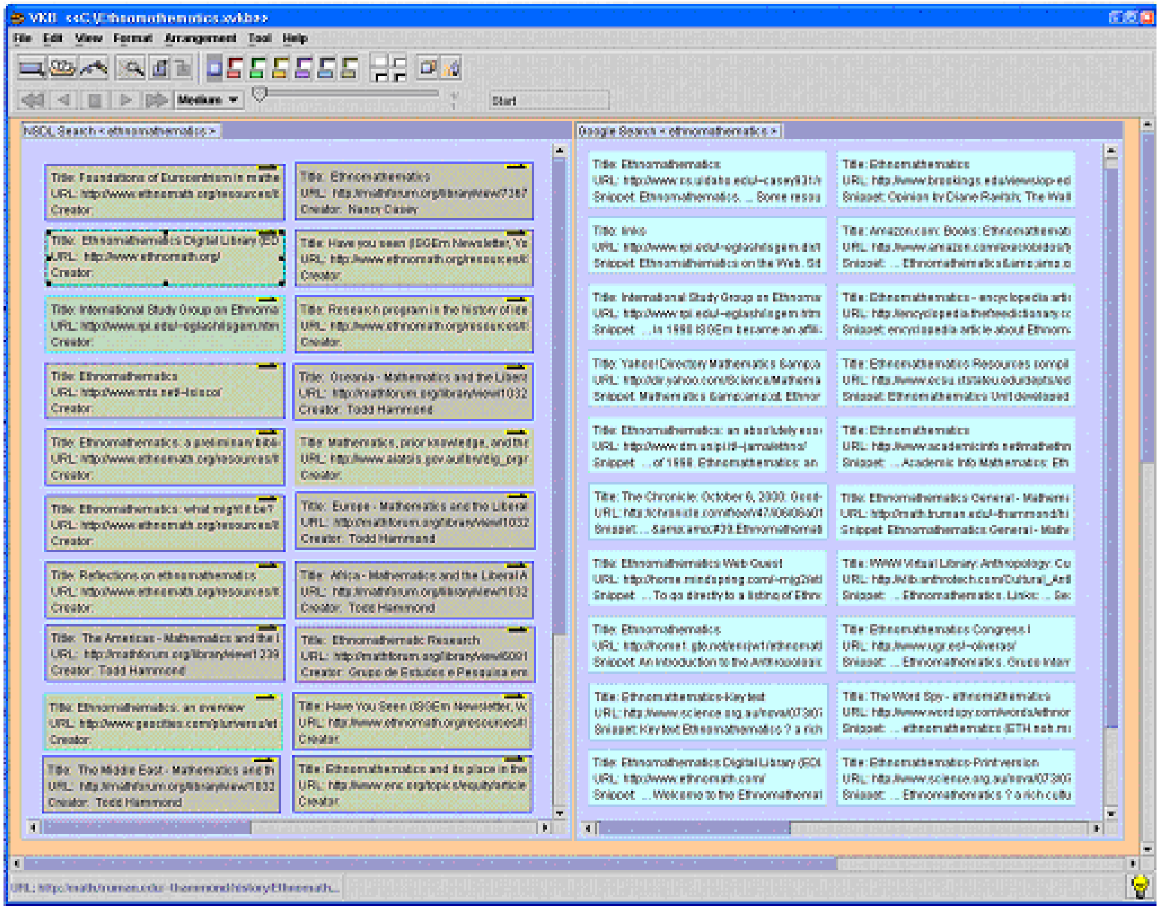
EVALUATION

A user study was conducted to evaluate the effectiveness of the interest models (presented in section - Interest Models) in recognizing user interest accurately. The study design was similar to the previous study [12] carried out as a part of the Document Triage project. The following subsections discuss the study design, the system setup and the results.

A. Study design

This study was conducted to evaluate if the identified implicit interest indicators are effective in recognizing user interest accurately. The study took place in the Center for the Study of Digital Libraries at Texas A&M University. 8 subjects were recruited via mass email. The subjects included graduate students and research associates from the university. The subjects were in the age range $22 - 32$. Subjects were required to have basic familiarity with using a computer and browsing the web, All 8 subjects were regular computer users for five or more years.

The subjects were placed in the role of a research librarian who had to select and organize documents for a high school teacher preparing a class on ethnomathematics (the study of a group's culturally-specific mathematical practices as they go about their everyday activities). Subjects started with 20 documents returned from the National Science Digital Library (NSDL) and 20 documents returned from Google placed in lists in VKB. VKB allows users to organize information objects (links to websites in this study) in a hierarchy of two-dimensional workspaces. All the subjects were given a brief training on VKB with emphasis on features considered relevant for the task. The same 40 links (from NSDL and Google) were provided to all the


VKB <<C:\Ethnomathematics.vkb>>
File
Edit
View
Format
Arrangement
Tool
Help
Medium
Start
NSDL Search <ethnomathematics>
Google Search <ethnomathematics>

Title: Foundations of Eurocentrism in mathe
URL: http://www.ethnomath.org/resources/l
Creator:

Title: Ethnomathematics
URL: http://mathforum.org/library/view/7367
Creator: Nancy Casey

Title: Ethnomathematics Digital Library (ED
URL: http://www.ethnomath.org/
Creator:

Title: Have you seen ISGEm Newsletter, Vo
URL: http://www.ethnomath.org/resources/l
Creator:

Title: International Study Group on Ethnoma
URL: http://www.rpi.edu/~eglash/isgem.htm
Creator:

Title: Research program in the history of ide
URL: http://www.ethnomath.org/resources/l
Creator:

Title: Ethnomathematics
URL: http://www.nts.net/~lisacor/
Creator:

Title: Oceania - Mathematics and the Libera
URL: http://mathforum.org/library/view/1032
Creator: Todd Hammond

Title: Ethnomathematics: a preliminary bibli
URL: http://www.ethnomath.org/resources/l
Creator:

Title: Mathematics, prior knowledge, and the
URL: http://www.aiatsis.gov.au/lbry/dig_prgr
Creator:

Title: Ethnomathematics: what might it be?
URL: http://www.ethnomath.org/resources/l
Creator:

Title: Europe - Mathematics and the Liberal
URL: http://mathforum.org/library/view/1032
Creator: Todd Hammond

Title: Reflections on ethnomathematics
URL: http://www.ethnomath.org/resources/l
Creator:

Title: Africa - Mathematics and the Liberal A
URL: http://mathforum.org/library/view/1032
Creator: Todd Hammond

Title: The Americas - Mathematics and the L
URL: http://mathforum.org/library/view/1239
Creator: Todd Hammond

Title: Ethnomathematic Research
URL: http://mathforum.org/library/view/6001
Creator: Grupo de Estudos e Pesquisa em

Title: Ethnomathematics: an overview
URL: http://www.geocities.com/pluriversalel
Creator:

Title: Have You Seen ISGEm Newsletter, V
URL: http://www.ethnomath.org/resources/l
Creator:

Title: The Middle East - Mathematics and th
URL: http://mathforum.org/library/view/1032
Creator: Todd Hammond

Title: Ethnomathematics and its place in the
URL: http://www.enc.org/topics/equity/article
Creator:

Title: Ethnomathematics
URL: http://www.cs.uidaho.edu/~casey931/
Snippet: Ethnomathematics. ... Some resou

Title: Ethnomathematics
URL: http://www.brookings.edu/views/op-ed
Snippet: Opinion by Diane Ravitch; The Wall

Title: links
URL: http://www.rpi.edu/~eglash/isgem.dir/
Snippet: Ethnomathematics on the Web. Sit

Title: Amazon.com: Books: Ethnomathemat
URL: http://www.amazon.com/exec/obidos/t
Snippet: ... Ethnomathematics&amp;a

Title: International Study Group on Ethnoma
URL: http://www.rpi.edu/~eglash/isgem.htm
Snippet: ... In 1998 ISGEm became an affili

Title: Ethnomathematics - encyclopedia arti
URL: http://encyclopedia.thefreedictionary.cc
Snippet: encyclopedia article about Ethnom

Title: Yahoo! Directory Mathematics &ampco
URL: http://dir.yahoo.com/Science/Mathema
Snippet: Mathematics &amp;ct; Ethnor

Title: Ethnomathematics Resources compil
URL: http://www.ecsu.ctstateu.edu/depts/ed
Snippet: Ethnomathematics Unit developed

Title: Ethnomathematics: an absolutely ess
URL: http://www.dm.unipi.it/~jama/ethno/
Snippet: ... of 1998. Ethnomathematics: an

Title: Ethnomathematics
URL: http://www.academicinfo.net/mathethn
Snippet: ... Academic Info Mathematics: Eth

Title: The Chronicle: October 6, 2000: Good-
URL: http://chronicle.com/free/v47/i06/06a01
Snippet: ... &amp;#39;Ethnomathemati

Title: Ethnomathematics General - Mathema
URL: http://math.truman.edu/~thammond/hi
Snippet: Ethnomathematics General - Mathe

Title: Ethnomathematics Web Quest
URL: http://home.mindspring.com/~mig2/eth
Snippet: ... To go directly to a listing of Ethn

Title: WWW Virtual Library: Anthropology: Cu
URL: http://vlib.anthrotech.com/Cultural_Ant
Snippet: ... Ethnomathematics. Links: ... Se

Title: Ethnomathematics
URL: http://horizon.gtd.net/en/wt/ethnomat
Snippet: An introduction to the Anthropologic

Title: Ethnomathematics Congress I
URL: http://www.ugr.es/~oliveras/
Snippet: ... Ethnomathematics. Grupo Inter

Title: Ethnomathematics-Key text
URL: http://www.science.org.au/nova/073/07
Snippet: Key text Ethnomathematics ? a rich

Title: The Word Spy - ethnomathematics
URL: http://www.wordspy.com/words/ethnor
Snippet: ... ethnomathematics (ETH.noh.m

Title: Ethnomathematics Digital Library (EDL
URL: http://www.ethnomath.com/
Snippet: ... Welcome to the Ethnomathemati

Title: Ethnomathematics-Print version
URL: http://www.science.org.au/nova/073/07
Snippet: ... Ethnomathematics ? a rich cultu

URL: http://math.truman.edu/~thammond/history/Ethnomath...

Fig. 9. VKB space provided to the subject

subjects. The documents varied in their level of difficulty, relatedness to the topic and volume of information. Although no time limit was set, all the subjects took less than one and a half hours for the task. Figure 9 shows the VKB document (with 40 links) as provided to the subjects. The subjects were to use only the given instance of VKB, and Internet Explorer as the web browser.

This is the same task and topic as in the studies reported in [12] and [13]. Before the task, the subject's were asked to fill out a demographical survey. This document is in Appendix A. Subjects were asked to fill out two questionnaires after the task, a questionnaire on document usefulness ratings and a questionnaire to gather the subject's input on the task and the system. These questionnaires are in Appendix B and

C respectively. They took part in a short interview after filling out the questionnaires. The interview was conducted for a deeper understanding of their task practices, document rating, usage of meta-data in evaluation of document relevance, etc. The main focus of the interview was on understanding the subject's triage practice and the criteria the subject used to rate the documents.

The subjects were provided with a computer with a wide-aspect ratio 20" flat screen panel for the task. The subjects had to organize the 40 links into visual structures for the high school teacher. They had to determine their own criteria for organizing the links, and were free to add, modify and delete the links. They were also free to modify other attributes in the VKB space such as background color, border thickness, font color, etc., as well as adding text or annotation to the VKB space as they deemed necessary.

Additionally, the subject's on-screen activity was captured using screen capturing video software. The subjects were also video taped from behind their back. This is to provide finer details such as the subject's focus of attention during the task. Finally, user actions in VKB and IE were logged and provide event times, URLs and Internet Explorer window identifiers. The Interest Profile Manager was used to store user behavior details from web pages. The next subsection describes the specific system components and configuration used to collect data during the study.

B. System setup for study

The system architecture is shown in Figure 10. The main components of the system are: VKB, Automated IE, a VB Control Module, the Interest Profile Manager, and the Interest Profile Shell Client.

A version of VKB containing two extra components, the Interest Profile Client

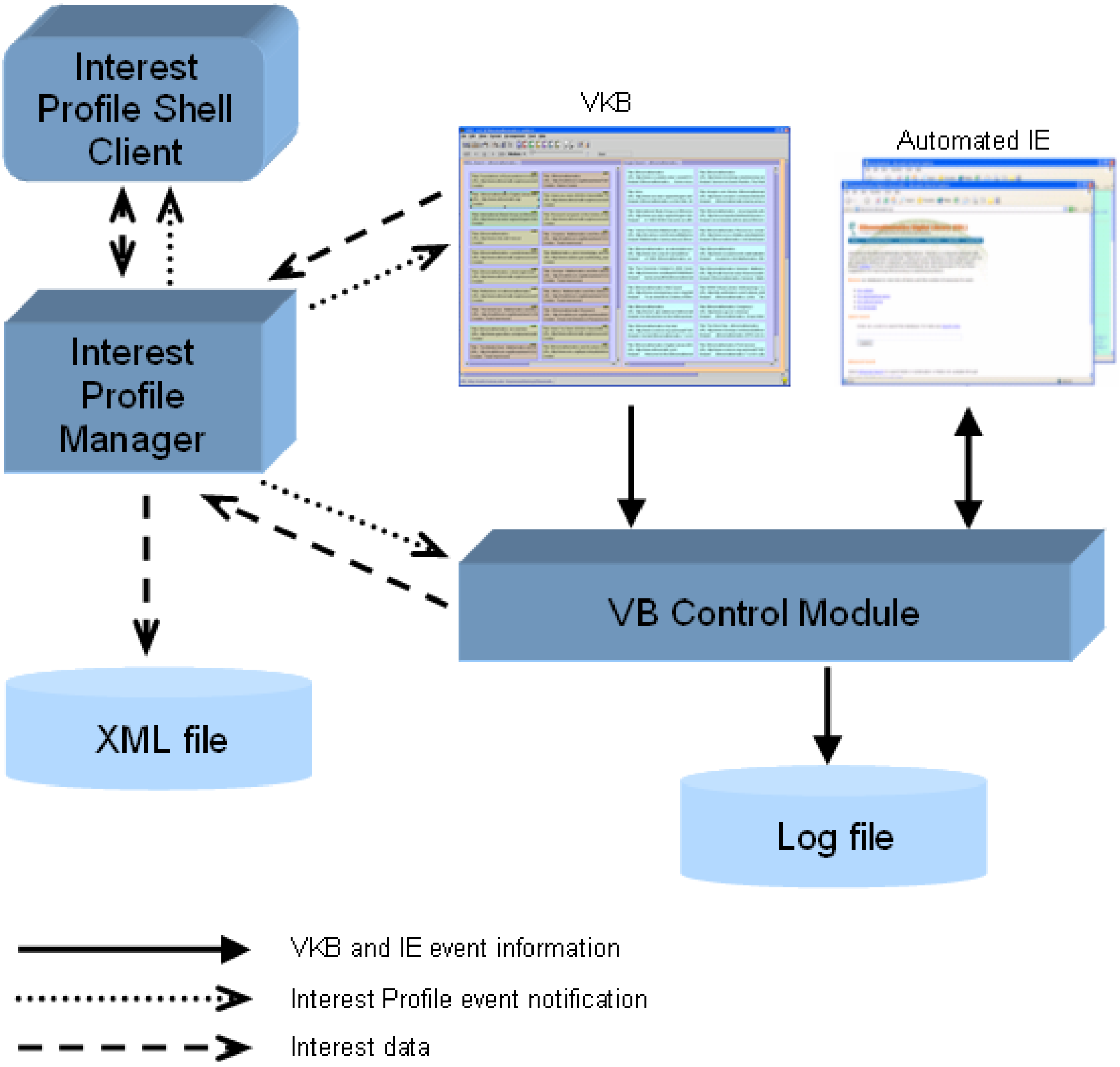
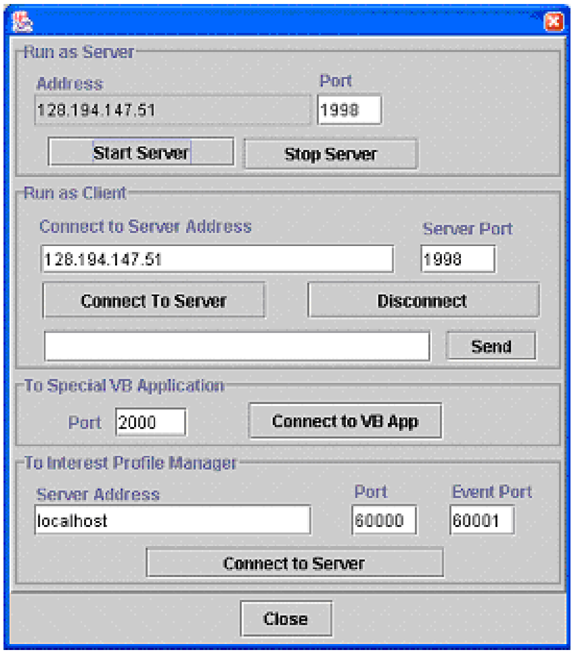

Interest Profile Shell Client
VKB
Automated IE
Interest Profile Manager
VB Control Module
XML file
Log file
VKB and IE event information
Interest Profile event notification
Interest data

Fig. 10. Overall system architecture

and a component to interact with the VB Control Module, was used for the study. The connection between VKB and the VB Control Module and the Interest Profile Manager and the VB Control Module are controlled via the interface shown in Figure 11. VKB sends various events to the VB Control Module once connected. When the user double-clicks on a link in VKB, the URL is sent to the VB Control Module which then invokes Automated IE to open the web page. VKB also



Address 128.194.147.51 Port 1998

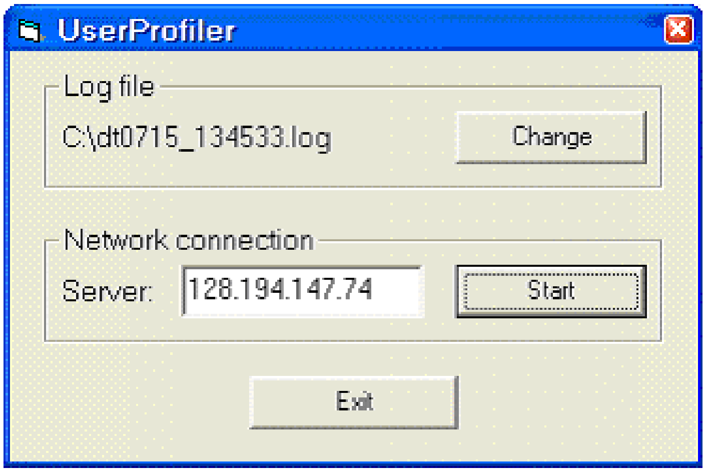
Connect to Server Address 128.194.147.51 Server Port 1998

Port 2000

Server Address localhost Port 60000 Event Port 60001

Fig. 11. Network connection dialog in VKB

sends click information, and focusing and defocusing events to the VB Control Module. The VB Control Module records the various events into a log file. VKB, when connected to the Interest Profile Manager, also sends edit events corresponding to the link objects. The edit events propagated are: AddSymbol, DeleteSymbol, MoveSymbol, ResizeSymbol, ChangeBackgroundColor, ChangeBackgroundPallet, ChangeBorderColor, ChangeBorderPallet, ChangeBorderWidth, ChangeTransparency, ChangeFont, ChangeFontColor, ChangeContent, ChangeAttribute, ChangeZOrder and ChangeCanvasColor. The Interest Profile manager API includes a function updateAndAddAttributeValues to send these edit events. This function updates the Attribute corresponding to the URL with the new data, adding the new data to the corresponding old data if they match. For example, assume an Attribute for http://ethnomath.org already exists in the Interest Profile Manager and it contains an AttributeValue for MoveSymbol with value 3. If the link object corresponding to http://ethnomath.org is moved again in VKB, using updateAndAddAttributeValues



Log file

C:\dt0715_134533.log

Network connection

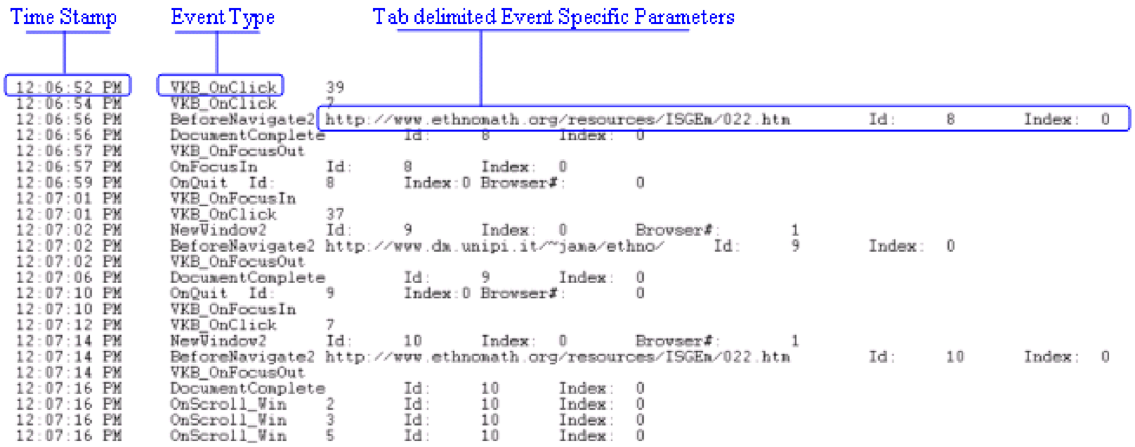
Server: 128.194.147.74

Fig. 12. User Interface of the VB Control Module

causes the value for MoveSymbol to change to 4.

The VB Control Module connects to both the Interest Profile Manager and VKB. Figure 12 shows the interface of the VB Control Module. It provides functionality for opening IE and gathering system and user events from the automated IE instance. Some of the events captured are: document load and navigate, document focus in and out, new window and close window, scroll, click, and select. The Control Module also gathers document data such as the number of characters, words, images and links in the web page. The event data is written into a log file along with time stamps and other event specific information. A snippet of a typical log file is shown in Figure 13.

The Interest Profile Manager stores the user events and web page parameters. An *Attribute* data structure is created for each URL the user visits. Multiple visits to the same URL result in the same Attribute being updated. The event data and the document information are contained in *AttributeValues* in the Attribute corresponding to the URL. A single Attribute from the XML file generated by the Interest Profile Manager is shown in Figure 14. The Data element of AttributeValue contains the name of the event or the document attribute. The Value element contains the value



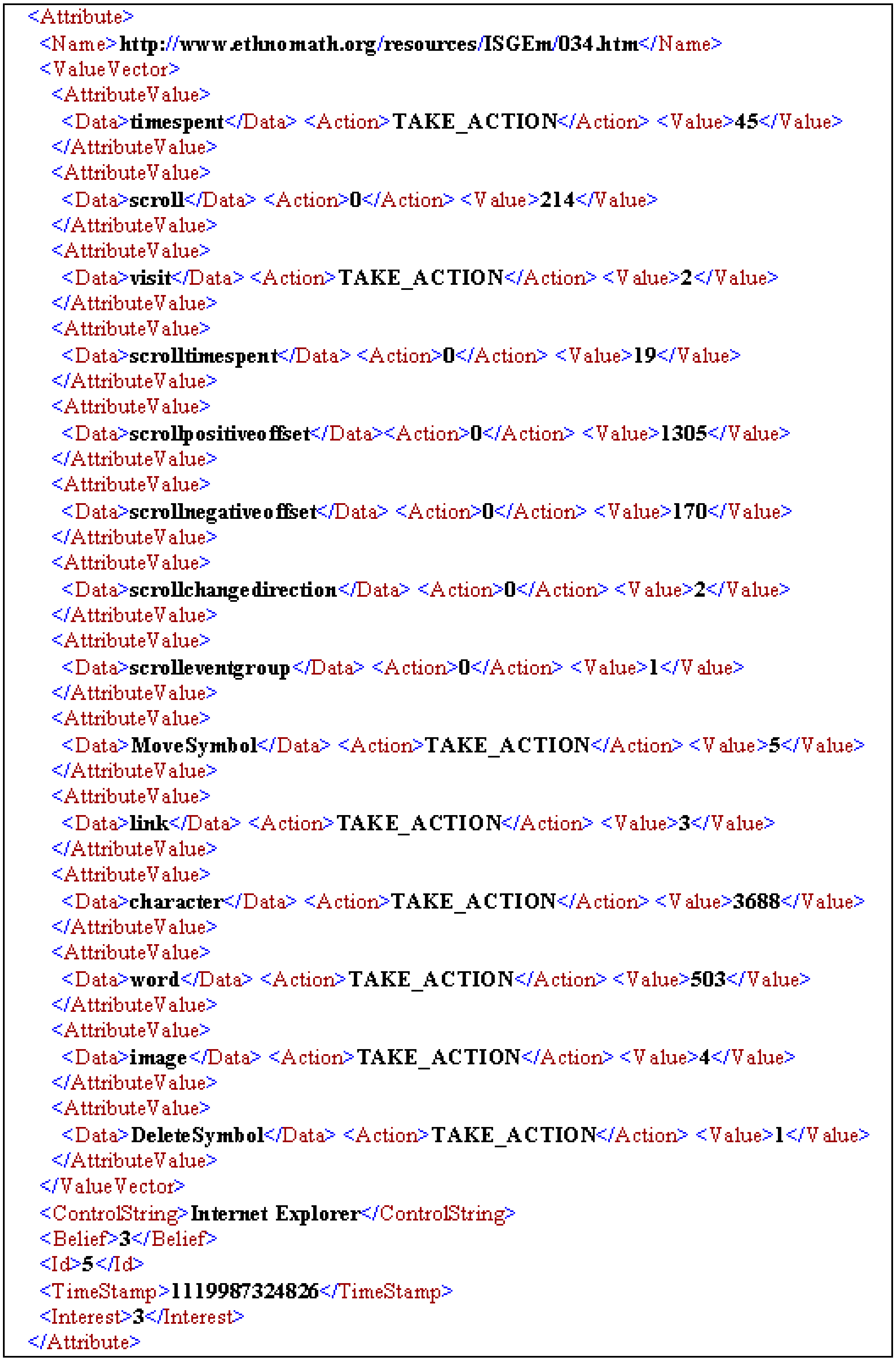
```
12:06:52 PM     VKB_OnClick     39
12:06:54 PM     VKB_OnClick     7
12:06:56 PM     BeforeNavigate2 http://www.ethnomath.org/resources/ISGEm/022.htm        Id:     8       Index:  0
12:06:56 PM     DocumentComplete        Id:     8       Index:  0
12:06:57 PM     VKB_OnFocusOut
12:06:57 PM     OnFocusIn       Id:     8       Index:  0
12:06:59 PM     OnQuit  Id:     8       Index:0 Browser#:       0
12:07:01 PM     VKB_OnFocusIn
12:07:01 PM     VKB_OnClick     37
12:07:02 PM     NewWindow2      Id:     9       Index:  0       Browser#:       1
12:07:02 PM     BeforeNavigate2 http://www.dm.unipi.it/~jama/ethno/     Id:     9       Index:  0
12:07:02 PM     VKB_OnFocusOut
12:07:06 PM     DocumentComplete        Id:     9       Index:  0
12:07:10 PM     OnQuit  Id:     9       Index:0 Browser#:       0
12:07:10 PM     VKB_OnFocusIn
12:07:12 PM     VKB_OnClick     7
12:07:14 PM     NewWindow2      Id:     10      Index:  0       Browser#:       1
12:07:14 PM     BeforeNavigate2 http://www.ethnomath.org/resources/ISGEm/022.htm        Id:     10      Index:  0
12:07:14 PM     VKB_OnFocusOut
12:07:16 PM     DocumentComplete        Id:     10      Index:  0
12:07:16 PM     OnScroll_Win    2       Id:     10      Index:  0
12:07:16 PM     OnScroll_Win    3       Id:     10      Index:  0
12:07:16 PM     OnScroll_Win    5       Id:     10      Index:  0
```

Fig. 13. Snippet of the VB Control Module log file

of the Data. Action field with TAKE_ACTION tells clients that the corresponding AttributeValue is recently updated and hence, action needs to be taken. The Action element is set to 0 once action is taken based on the change.

Consolidated event data corresponding to a user's visit to a web page is sent to the Interest Profile Manager using the *updateAndAddAttributeValues* function. The Attribute corresponding to the URL thus contains the cumulative sum of the various events across multiple visits during the triage activity. Document data meant to replace existing values, e.g. number of characters, is sent to the Interest Profile Manager using the *updateAttributeValues* function. It connects to the Interest Profile Manager as a client. Since it waits for VKB to connect to it as a client, no network connection parameters corresponding to this are provided in its user interface.

The Interest Profile Shell Client is used to monitor the working of the Interest Profile Manager and to save the results into an XML file. It consists of the Interest Profile Client wrapped in a shell interface. This shell command line interface provides commands such as connectToServer, save, load, setXmlFile, addAttribute, removeAttribute, removeAllAttributes, etc. to perform various operations on the In-



```
<Attribute>
 <Name>http://www.ethnomath.org/resources/ISGEm/034.htm</Name>
 <ValueVector>
  <AttributeValue>
   <Data>timespent</Data> <Action>TAKE_ACTION</Action> <Value>45</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>scroll</Data> <Action>0</Action> <Value>214</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>visit</Data> <Action>TAKE_ACTION</Action> <Value>2</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>scrolltimespent</Data> <Action>0</Action> <Value>19</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>scrollpositiveoffset</Data><Action>0</Action> <Value>1305</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>scrollnegativeoffset</Data> <Action>0</Action> <Value>170</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>scrollchangedirection</Data> <Action>0</Action> <Value>2</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>scrolleventgroup</Data> <Action>0</Action> <Value>1</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>MoveSymbol</Data> <Action>TAKE_ACTION</Action> <Value>5</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>link</Data> <Action>TAKE_ACTION</Action> <Value>3</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>character</Data> <Action>TAKE_ACTION</Action> <Value>3688</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>word</Data> <Action>TAKE_ACTION</Action> <Value>503</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>image</Data> <Action>TAKE_ACTION</Action> <Value>4</Value>
  </AttributeValue>
  <AttributeValue>
   <Data>DeleteSymbol</Data> <Action>TAKE_ACTION</Action> <Value>1</Value>
  </AttributeValue>
 </ValueVector>
 <ControlString>Internet Explorer</ControlString>
 <Belief>3</Belief>
 <Id>5</Id>
 <TimeStamp>1119987324826</TimeStamp>
 <Interest>3</Interest>
</Attribute>
```

Fig. 14. Attribute from the Interest Profile Manager XML file

terest Profile Manager. Since it is a client, it can connect to the Interest Profile Manager from a different machine. The Shell Client thus provides remote administration mechanism for the Interest Profile Manager. This application is also helpful in handling cases of unexpected error conditions.

C. Study results

After the subjects finished the task, they were asked to rate the 40 documents based on how useful they found the documents to be on a Likert scale ranging from 1 (not useful) through 5 (very useful). The Document Rating questionnaire is found in Appendix B. Five (of the 40) documents were disregarded from the analysis. Two documents were repeated, that is each of these documents had two objects in VKB. Both instances of each of these documents were removed from the analysis due to the inconsistency in analysis that would result from merging user activity across the two instances. Two other documents that were ignored had the same content even though their URLs were different. Additionally, one of the documents was a pdf file, and no data could be gathered from the pdf viewer.

For each subject, two files were generated, the Interest Profile Manager xml file and the VB Control Module log file. The Interest Profile Manager xml file contains details on all the URLs the subject visited, with each URL stored as a separate Attribute (shown in Figure 14). Each Attribute contains the user document reading and interpretation activity data, along with the document style information for the corresponding URL. The VB Control Module log file contains timing information for all user events in IE as well as the edit events in VKB. The analysis here refers to application of the three models (Reading-Activity Based model, Limited Collinearity Reading-Activity Based model and Qualitatively-Defined model) to the parsed data.

For each URL visited by each subject, each of the 3 models was applied to obtain the corresponding document score.

For each subject, these explicit ratings were compared with the document scores obtained from the 3 models. For the correlation analysis carried out between two models (Reading-Activity Based model and Limited collinearity reading-activity Based model) and the document ratings, the results showed no statistical significance. This is not surprising as the models have been built using data accumulated across all users from the previous study. The correlation analysis carried out for the Qualitatively-Defined model is shown in Table XI. The analysis shows that the current models cannot be used for effective evaluation of individual user interest.

Table XI. Correlation analysis for qualitatively-defined model using individual subject data

Subject ID	Pearson Correlation	Sig. (2-tailed)	N
1	−0.006	0.971	35
2	0.121	0.54	28
3	−0.144	0.416	34
4	−0.016	0.927	35
5	0.004	0.981	35
6	0.315	0.065	35
7	0.296	0.193	21
8	0.126	0.478	34

Further, the document ratings across all 8 subjects were summed up to give the composite document ratings. Also, the gathered user activity data for the individual URLs was summed up for all the users. All 3 models were applied on this accumulated data. Correlation analysis was carried out on this accumulated data between two models (Reading-Activity Based model and Limited Collinearity Reading-Activity Based model) and the composite document ratings. The correlation analysis results (Table XII) show that the current Limited Collinearity Reading-Activity based model performs better in predicting document usefulness than the other model. However, the Pearson correlation coefficient for the correlation between Limited Collinearity Reading-Activity based model and the composite document rating is 0.411, showing that the strength of the relationship is low.

Table XII. Correlation analysis for accumulated data

	Parameters	Reading-Activity Based Model	Limited Collinearity Reading-Activity Based Model
Document Model	Pearson Correlation	.028	.411
	Sig. (2-tailed)	.871	.014
	N	35	35

In another analysis, the document scores obtained from applying the models were compared with the explicit ratings, disregarding the subject ids. In other words, the document scores and the ratings were compared for all the URLs as if they were obtained from a single subject. The Limited Collinearity Reading-Activity based model applied on the gathered data for the individual data across all subjects performs better than the Reading-Activity based model. The correlation analysis between the document rating and the model is shown in Table XIII. However, the Pearson coefficient shows only a weak positive association between the two variables.

Table XIII. Correlation analysis for individual data across all users (limited collinearity reading-activity based model)

		Limited Collinearity Reading-Activity Based Model
Document Model	Pearson Correlation	.176
	Sig. (2-tailed)	.005
	N	258

D. Discussion

From the above subsection, we note that the Limited Collinearity Reading-Activity model when applied to accumulated data is significantly correlated to the document score. When the same model is applied to individual subjects, no correlation is seen. In the former case, idiosyncrasy of the subject does not affect the model performance. However, in the latter, the difference in style of users significantly impacts the value of the model.

In interviews that followed the task, the users were asked about why they rated documents lower and higher, about revisiting documents, and about their overall organizing strategy. The following discussion is based on both the quantitative results of the study as well as the qualitative assessment based on the task and the interviews.

Subjects rated documents mainly on the content. Some of the subjects considered web pages with more information as useful, rating academic journals and some comprehensive sources as more useful, while others rated them low as they felt that the teacher would prefer introductory material for her class. Most of the subjects relied on the domain names of the documents for authenticity, rating documents from .edu domains and digital libraries higher than other documents. It was interesting to note that a couple subjects relied on the update frequency of the web pages as a criterion in rating them, rating documents which were recently updated higher. They relied on the “Last updated” information in the web pages for this criterion. Some of the subjects rated a few documents even without visiting them, basing their decisions on metadata alone. For example, one of the subjects moved the Amazon link to a collection based on the URL and title information provided in the VKB symbol, and subsequently assigned a rating of 4 to the URL after the task.

Subjects revisited documents for different reasons. One of the subjects visited most links cursorily in the first pass “to see what was out there”, classified most documents in the second pass, and classified the remaining documents in the third pass. Another subject revisited some of the documents as he did not know if he had already visited them.

Another aspect for consideration is the deletion of links from the VKB space. The users had not been given any specific instructions regarding discarding links from the space. Some users chose to delete links which they felt were not useful. Some others chose to place such links in separate collections giving those names such as “Other”. Some subjects chose not to delete links and left them in their original collections without moving them to any new collections they created. One of the subjects deleted most of the links leaving only 10 links in the VKB space. The subject mentioned that the 10 links would suffice for the teacher, and that many of the other links could be reached from these 10 links. The subject provided explicit ratings for 33 documents. One of the hypotheses going into the study was that the subjects deleted links of documents that were not useful. However, this hypothesis is not substantiated by the study results.

The usage of color in VKB was also idiosyncratic based on the subjects' strategy for completing the task. One of the subjects used elaborate color coding using red for web pages with good examples, purple for web pages with introductory material, etc., and using thick red borders for web pages with good examples which were already assigned a different color. The subject mentioned that she would set the color and border thickness of a symbol after reading the corresponding web page and in the second pass, drop the matching symbols into collections. Another subject who came up with a chapter based classification, used color to show relationship between the chapter headings and their contents.

From the quantitative and qualitative assessments, it follows that user activity is highly idiosyncratic and it will be difficult for a single model to work well across a population. To improve performance, future models will need to increase the number of triage activity and document characteristics included in the model and better balance their contributions to the inferred value. That said, the results of the models on aggregated data indicate the potential for the models to be used in collaborative applications.

CHAPTER VI

CONCLUSION AND FUTURE WORK

As part of the Document Triage Project, we are looking at ways to help users in their document triage activity. The tools developed in this research are meant to help in recognizing, representing, communicating and visualizing user interest. Document triage involves multiple applications such as reading interfaces and document organizing interfaces. Hence, our focus is on developing tools to support document triage involving multiple applications. The objective of this thesis is two-fold, recognition of user interest and representing it. To recognize user interest, data is gathered from the user's reading, navigational and interpretive activities. The Interest Profile Manager was developed as an infrastructure to represent user interest across multiple applications.

The Interest Profile Manager performs the dual role of representing user interest and as a communication medium across applications. All applications involved in the triage task communicate with one another through the Interest Profile Manager. The Interest Profile Manager also provides an event infrastructure for applications to notify and take action based on changes made to the data stored in the Interest Profile Manager. The Interest Profile Manager provides capability to store and retrieve XML files. The shared data stored in the Interest Profile Manager is used for estimating user interest. Models for estimating user interest can be plugged into the infrastructure as modules built upon Interest Profile Clients.

A study was conducted to evaluate the efficacy of 3 models to estimate user interest. Eight subjects took part in the study. The subjects were to don the role of a research librarian who had to select and organize documents for a high school teacher preparing a class on ethnomathematics. For the gathered data, 3 models were applied to the data to evaluate their efficacy in recognizing user interest. None of the models were found to be effective in recognizing user interest for a given subject. The Limited Collinearity Reading-Activity Based model is a statistically significant model for recognizing documents of interest when it is built using accumulated data and compared against cumulative assessment. From the qualitative analysis, we note that some enhancements can be made to the current models for improved estimation of user interest.

This thesis presents a first step to supporting document triage using models of user interest. In particular, while the Interest Profile Manager enables communication of the results of interest models, the models developed and evaluated were single-application interest models. Multi-application interest models should perform better as they have more information about the user interaction across the applications involved in the triage activity. The Interest Profile Manager is a step in this direction. However, much research needs to go into development of multi-application interest models and their applicability.

REFERENCES

[1] B. Sarwar, J. Konstan, A. Borchers, J. Herlocker, B. Miller, and J. Riedl, "Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system," in *Proc. ACM Conference on Computer Supported Cooperative Work*, Seattle, Jan. 1998, pp. 345–354.

[2] F. M. Shipman III, H. Hsieh, P. Maloor, and J. M. Moore, "The visual knowledge builder: A second generation spatial hypertext," in *Proc. 12th ACM conference on Hypertext and Hypermedia*, Budapest, Hungary, 2001, pp. 113–122.

[3] M. Claypool, P. Le, M. Waseda, and D. Brown, "Implicit interest indicators," in *Proc. International Conference on Intelligent User Interfaces*, Santa Fe, 2001, pp. 33–40.

[4] J. Grudin, "Groupware and social dynamics: Eight challenges for developers," *Communications of the ACM*, vol. 35, pp. 92–105, Jan. 1994.

[5] D. M. Nichols, "Implicit rating and filtering," in *Proc. 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary, Nov. 1997, pp. 10–12.

[6] M. Morita and Y. Shinoda, "Information filtering based on user behaviour analysis and best match text retrieval," in *Proc. 17th ACM Annual International Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 1994, pp. 272–281.

[7] J. Kim, D. W. Oard, and K. Romanik, "Using implicit feedback for user modeling in internet and intranet searching," Tech. Rep., College of Library and Information Services, University of Maryland, College Park, 2000.

[8] D. Kelly and N. J. Belkin, "Display time as implicit feedback: Understanding task effects," in *Proc. 27th Annual ACM International Conference on Research and Development in Information Retrieval*, Sheffield, UK, 2004, pp. 377–384.

[9] J. Goecks and J. Shavlik, "Learning users interests by unobtrusively observing their normal behavior," in *Proc. 5th international conference on Intelligent User Interfaces*, New Orleans, 2000, pp. 129–132.

[10] P. Chan, "A non-invasive learning approach to building web user profiles," in *Proc. ACM SIGKDD International Conference*, San Diego, 1999, pp. 7–12.

[11] F. M. Shipman, M. N. Price, C. C. Marshall, G. Golovchinsky, and B. N. Schilit, "Identifying useful passages in documents based on annotation patterns," in *Proc. 7th European Conference on Research and Advanced Technology for Digital Libraries*, Trondheim, Norway, Aug. 2003, pp. 101–112.

[12] S. Bae, R. Badi, K. Meintanis, J. M. Moore, A. Zacchi, H. Hsieh, C. Marshall, and F. Shipman, "Effects of display configurations on document triage," in *Proc. IFIP Interact Conference*, Rome, Italy, Sep. 2005.

[13] F. Shipman, H. Hsieh, J. M. Moore, and A. Zacchi, "Supporting personal collections across digital libraries in spatial hypertext," in *Proc. ACM and IEEE Joint Conference on Digital Libraries*, Tucson, AZ, June 2004, pp. 358–367.

APPENDIX A

DEMOGRAPHICS & DOMAIN SURVEY

Subject ID: ________________

Please fill in value(s) or circle your response.

1. Age: _______

2. Gender: Male / Female (circle appropriate)

3. Classification:

   ○ Student

   Seeking Degree: ______ (specify if Non-degree seeking)

   Major: ___________

   ○ Non-Student

   Job Title: ____________________

4. Amount of computer use

   ○ Heavy (more than 20 hours/week)

   ○ Moderate (between 10-20 hours/week)

   ○ Light (less than 10 hours/week)

5. Location of computer use (check all that apply)

   □ Home

   □ School/Work

- □ Other specify: ____________________

6. Type of computers used (check all that apply)

   - □ Desktop
   - □ Laptop
   - □ Other (PDAs, etc.)

7. I typically use a computer to: (check all that apply)

   - □ Send and receive email (if checked, about how many emails do you receive per day? ______)
   - □ Chat and Instant Messaging
   - □ Browse the Web for things I find interesting
   - □ Do Web-based research on specific topics
   - □ Read Newspapers online
   - □ Shop online
   - □ Create documents
   - □ Play games
   - □ Other (please specify) ________________

8. I usually print the following items on the screen: (check all that apply)

   - □ Long emails
   - □ Short emails
   - □ Long documents
   - □ Short documents

□ Newspaper or magazine articles

□ Informational web pages (other than receipts)

9. I usually read the following items on the screen: (check all that apply)

   □ Long emails

   □ Short emails

   □ Long documents

   □ Short documents

   □ Newspaper or magazine articles

   □ Informational web pages

10. I have used computers regularly for _________ months / years. (circle appropriate)

APPENDIX B

DOCUMENT RATING

Subject ID: _______________

		Not useful			Very Useful	
1	http://www.ethnomath.org/resources/ISGEm/034.htm	1	2	3	4	5
2	http://www.ethnomath.org/	1	2	3	4	5
3	http://www.rpi.edu/~eglash/isgem.htm	1	2	3	4	5
4	http://www.mts.net/~lsisco/	1	2	3	4	5
5	http://www.ethnomath.org/resources/ISGEm/023.htm	1	2	3	4	5
6	http://www.ethnomath.org/resources/ISGEm/022.htm	1	2	3	4	5
7	http://www.ethnomath.org/resources/ISGEm/030.htm	1	2	3	4	5
8	http://mathforum.org/library/view/12390.html	1	2	3	4	5
9	http://www.geocities.com/pluriversu/ethno.html	1	2	3	4	5
10	http://mathforum.org/library/view/10328.html	1	2	3	4	5
11	http://mathforum.org/library/view/7367.html	1	2	3	4	5
12	http://www.ethnomath.org/resources/ISGEm/031.htm	1	2	3	4	5
13	http://www.ethnomath.org/resources/ISGEm/035.htm	1	2	3	4	5

14	http://mathforum.org/library/view/10326.html	1	2	3	4	5
15	http://www.aiatsis.gov.au/lbry/dig_prgm/e_access/pmphlet/m0032661_v_ṗdf	1	2	3	4	5
16	http://mathforum.org/library/view/10327.html	1	2	3	4	5
17	http://mathforum.org/library/view/10322.html	1	2	3	4	5
18	http://mathforum.org/library/view/60813.html	1	2	3	4	5
19	http://www.ethnomath.org/resources/ISGEm/032.htm	1	2	3	4	5
20	http://www.enc.org/topics/equity/articles/document.shtm?input=ACQ-111552-1552	1	2	3	4	5
21	http://www.cs.uidaho.edu/˜casey931/seminar/ethno.html	1	2	3	4	5
22	http://www.rpi.edu/˜eglash/isgem.dir/links.htm	1	2	3	4	5
23	http://www.rpi.edu/˜eglash/isgem.htm	1	2	3	4	5
24	http://dir.yahoo.com/Science/Mathematics/Ethnomathematics/	1	2	3	4	5
25	http://www.dm.unipi.it/˜jama/ethno/	1	2	3	4	5
26	http://chronicle.com/free/v47/i06/06a01601.htm	1	2	3	4	5
27	http://home.mindspring.com/˜mjg2/ethalt.html	1	2	3	4	5
28	http://home1.gte.net/ericjw1/ethnomathematics.html	1	2	3	4	5
29	http://www.science.org.au/nova/073/073key.htm	1	2	3	4	5
30	http://www.ethnomath.com/	1	2	3	4	5
31	http://www.brookings.edu/views/op-ed/ravitch/20050620.htm	1	2	3	4	5
32	http://www.amazon.com/exec/obidos/tg/detail/-/0412989417?v=glance	1	2	3	4	5
33	http://encyclopedia.thefreedictionary.com/Ethnomathematics	1	2	3	4	5

34	http://www.ecsu.ctstateu.edu/depts/edu/projects/ethnomath.html	1	2	3	4	5
35	http://www.academicinfo.net/mathethno.html	1	2	3	4	5
36	http://math.truman.edu/~thammond/history/Ethnomathematics.html	1	2	3	4	5
37	http://vlib.anthrotech.com/Cultural_Anthropology/Ethnomathematics/	1	2	3	4	5
38	http://www.ugr.es/~oliveras/	1	2	3	4	5
39	http://www.wordspy.com/words/ethnomathematics.asp	1	2	3	4	5
40	http://www.science.org.au/nova/073/073print.htm	1	2	3	4	5

APPENDIX C

QUESTIONNAIRE

Subject ID: _______________

	Questions	Strongly Disagree		Neutral	Strongly Agree	
1	I feel comfortable reading documents on a computer.	1	2	3	4	5
2	It will be easy to go back later and understand the rationale behind my organization.	1	2	3	4	5
3	I enjoyed doing this task.	1	2	3	4	5
4	Did you have problems using the system? If yes, please explain.			YES		NO
5	Do you think it will be easy for someone else to understand the way you organized the documents? Please explain why.			YES		NO

6 What would help you organize the documents better (any software, hardware enhancement/modification)?

7 What features would you like to add to the reading interface (IE) to make it better?

8 What features in the organizing interface (VKB) were helpful?

9 What new features could be helpful additions to the organizing interface (VKB)?

VITA

Name: Rajiv Ravindranath Badi

Address: 301 Harvey R. Bright Building, Computer Science Department,
Texas A&M University, College Station, TX 77843-3112

Education:

- Master of Science, Computer Science, Texas A&M University, December 2005
- Bachelor of Engineering, Computer Science and Engineering, Bangalore University, 2001