A foray into Understanding the Next Billion Search Users

Naman K. Gupta
Language Technologies Institute
Carnegie Mellon University, Pittsburgh
nkgupta@cs.cmu.edu

Carolyn P. Rose
Language Technologies Institute
Carnegie Mellon University, Pittsburgh
cprose@cs.cmu.edu

ABSTRACT
This paper presents an investigation into the search behavior of low literacy users in the developing world, like the participants from rural areas in India for our study. The field of personalization research has been hindered by a paucity of appropriate data for inducing effective user models that target the real problems in information access for needy populations, such as low literacy users. Our goal is to address these limitations using a data-driven, user centered methodology. We present results from an experimental study that demonstrates that some important assumptions underlying current probabilistic models of information seeking on the web that govern the behavior of popular search services such as Google and Yahoo are not valid for our target user population. We present an analysis that offers specific suggestions for better supporting the information seeking practices of such users.

Author Keywords
Information Retrieval, Personalization, Low literacy.

ACM Classification Keywords
H3.3. Information Storage and Retrieval.

INTRODUCTION
Currently, internet penetration [8] in the developed world has reached a point that much of the growth in internet users for the next billion users must almost necessarily be from the developing world, where a large majority of users, particularly from rural regions, are low literacy users. What that means is that many of those new users will have very different needs from the great majority of internet users today, who are largely capable of using current search technology to meet their information seeking needs.

Imagine a student from a rural area in India or Africa with limited web experience and limited education level, or a foreign student with low English comprehension - just entering a school or university environment in the United States. For these users, the experience of using search technology is quite different than the experience many of us have effortlessly every day. For such users, their low comprehension of the language may act as a hindrance in formulating an effective query phrase. If relevant information is provided in response to their query, they may or may not recognize it as such. Long lists of search results may be overwhelming to them.

Many of the current models in the area of probabilistic retrieval [2, 5, 7, 9], which are embedded in popular search services such as Google and Yahoo, build in the assumption that users are able to distinguish effectively between relevant and non-relevant documents by examining the text around the links that have been provided in response to their query, that they click on those links that meet their needs best, and that the search ends when they have found what they are looking for. We challenge all of these assumptions when dealing with inexperienced and low literacy populations, and thus a targeted effort is necessary to assist such populations to search efficiently and effectively. Understanding the search strategies and needs for support of such populations is unchartered territory, and arguably essential at this time as internet penetration continues to expand into developing regions. The analysis of data presented in this paper contributes towards understanding the needs of this emerging market.

The emerging area of personalization of information access technologies has made some progress towards adapting the behavior of these technologies to the specific needs of particular user groups [1, 6]. However, the field of personalization research has been hindered by a paucity of appropriate data for inducing effective user models that target the real problems in information access for needy populations, such as low literacy users. Our goal is to address these limitations using a data-driven, user centered methodology. Fortunately, we have access to a large population of users who fit into our target user population, who have recently become part of a community where they have access to technology and support for their English communication skills, but are still relatively early in the process.

In the remainder of the paper we explain the experimental study we ran as part of this effort, along with results that confirm our suspicion that the assumptions underlying current probabilistic models of search behavior are not valid for our target user population. We present an analysis that

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

concludes with specific suggestions for better supporting the information seeking practices of such users.

METHOD

User Participants

We conducted this user study as training exercise with 300 Home Room Tutors (HRTs) currently undergoing a 1-Year IT Diploma training at the IIIT, Hyderabad campus. They are college undergraduates with diverse majors, having similar cultural and educational background as our target user population.

Experimental Procedure

The study was conducted in 6 sessions with 50 participants each over a period of 2 days. Each session extended for 2 hours. Initially the experimenter, giving a short self-introduction, explained the purpose and motivation behind the study. Then a brief walkthrough of the study was given to the participants. The experiment survey extended for 1 hour and 10 minutes duration: 10 minutes for completing a background information questionnaire, 10 minutes for installing a Search Activity Logging Toolbar and other browser configurations, 20 minutes for understanding the information seeking task and completing the Pre-search write-up. They were then given another 30 minutes for the Search activity and subsequent Task write-up. Once finishing the survey, the participants uploaded the log files recorded by the toolbar using the toolbar itself, and subsequently uninstalled it.

Experimental Task and Manipulation

The Experimental task itself was an exploratory information-seeking task [3] based on the following template:

Imagine that you are a new professor assigned to teach the course <*Course Name*> for the first time to 11th grade students, and you want to make sure the content is up-to-date with the latest <technology/literature>. The specific topics you will be focusing on are < *Broad Topic/Less Broad Topic/ Specific Topic*>. Write a brief content summary for the course curriculum with reference books to be followed during the course.

The slots in the template were filled in differently for each condition based on the experimental manipulation described below. Participants were asked to mention the characteristics of the students, which seem relevant to them for their assigned search task: Age, Gender, Educational Background, Medium of Instruction in School, Experience with Computers, Experience with Internet/Searching, Personal Interests, Others factors.

Before accessing any information online, they were asked to prepare a Pre-search write-up based on prior knowledge. Then using any search engine – Google, Bing, Yahoo etc, they were told to prepare a Post-search write-up having all the information required for the given Search Task.

They were asked to evaluate their familiarity with Information-seeking task topics, and also their perceived search task difficulty on a scale of 1-5, 5 being most difficult.

Experimental Manipulation

The difficulty of an information-seeking task is expected to have an effect on search strategy and task success. We operationalized the task difficulty as a combination of how familiar the topic is – Topic Familiarity, and what the level of specificity is with which the information need is formulated – Specificity. The experiment was a 3X2 factorial design, where the Specificity is a 3 level between subject factor – High, Medium, Low and the Topic Familiarity is 2 level between subject factors – High, Low. This design allows us to avoid order effects and confounds from interaction between Topic and Specificity. These 6 variations were defined in the 6 Experiment sessions with 50 participants each. The instructions across all the 6 sessions were same just the necessary variations in the

3x2		Topic Familiarity	
		Low	**High**
Specificity	**Low**	Any 2-3 topics on Foundational Computer Science (1A)	Any 2-3 Topics on World History in the 20th Century (2A)
	Medium	Broad Topics - Computer Hardware and Operating System (1B)	Broad Topics – World Wars and US-Russian Cold War (2B)
	High	Blue Ray discs and Unix Operating System (1C)	Watergate Scandal and Collapse of Soviet Union (2C)

Table 1: 3x2 Factorial Design with Specificity and Topic Familiarity as variables.

Information-seeking task statement according to the above factors.

Logistical Issues during Experiment

As the study was conducted in a real classroom in a resource poor environment, there were some logistical issues that resulted in some unfortunate data loss:

- Intermittent and slow Internet connectivity. This led to some incomplete surveys.
- Some participants neglected to upload the Search activity logs

TOOLS AND MATERIALS

The following Tools and Materials were used for the experiment:

- A 4 page Web-based survey[1] designed using www.surveymonkey.com. The survey included following question types- Background Information, Instructions for Installing Logging Toolbar, Search Task statement, Pre-Search and Post-Search Write-ups and instructions for uploading Search activity logs.
- Firefox browser compatible with both Windows and Ubuntu systems was used for the experiment.
- Lemur Query Log Toolbar[2] was used to log all Search based activities performed during the Experiment.

DATA COLLECTION AND PROCESSING

The data was collected in the following formats:

Survey Data. We collected a total 360 survey responses over the 6 study sessions. This included spurious responses filtered out during Pre-Processing described below. These surveys contained the following details:

- Background Information – Unique ID, Type of High School, Medium of Instruction in School and University, Experience and Frequency with Computers, Frequency of using Search Engines.
- Student Characteristics deemed relevant for the Search Task by the Participants
- Pre-Search and Post-Search Write-ups
- Self-reported Topic Familiarity and Search Task Difficulty.

Activity and Search Log Data. We collected 280 Activity and search logs using the Lemur Toolbar. These logs contained the following event details:

- Search Related – Details (Query string, timestamp) of all queries issued. Details (Result rank, URL, timestamp) of results clicked from results
- Viewed Pages – Details (URLs, content, Time on Page, timestamp) of all the pages viewed.
- Browser Events – Details (RClick, Add/Close New Tab/Window, Copy, Scroll events) of any browser activity during the experiment. This allows us to build a sequence of events during the Search session.

Gold Standard Data. We collected 6 Survey and Search Logs, one for each of the 6 conditions from 6 high literacy graduate students at a top-tier US university.

Data Pre-Processing

The incomplete responses in the Survey data were removed giving a total of 305 responses. This further reduced to 296 responses after removing double submissions from some participants.

Out of these logs, only 200 logs had Search Related information. This might have happened in cases where people did not use any search engine in performing the task, used other search engines than the specified (Google, Bing, Yahoo). Some Participants used the default Firefox Wecome Google search page which was not logged by the Toolbar.

Data Processing

For each Participant response including the Gold Standard responses, we build 5 different Language models [5] with commonly used Laplace Smoothing [2, 7]. Language models capture the distribution of words used by a user or population. Language models can be compared using metrics that measure how different their associated word distributions are, and thus can be used to rank users according to how different or similar they are to the Gold Standard Users. The five models computed for each user as well as the Gold Standard users are the following:

- AllSearchResultsModel – includes the content from all the top 10 search results returned in response to each of the queries issued by the participant. This is to evaluate the relevance of the queries compared to the ones issued by Gold Standard Users.
- ClickedResultsModel – includes the content from all the results from queries that were clicked by the participant. This is to evaluate the participant's ability to choose a relevant result from the results page.
- AllViewedPagesModel – includes the content from all the pages viewed by the participant, directly or by navigating across pages.
- Pre-Search Write-up Model – to evaluate their knowledge and understanding of the task prior to Searching.
- Post-Search Write-up Model – to evaluate their ability to pick content relevant to the task.

A variable referring to the names of these 5 models is referred to in the remainder of the paper as Model-Label. Language models for each participant in a study session were compared using KL divergence [3] with corresponding Gold standard Language model for that session. KL divergence measures the difference between two distributions. In this context, it is used as a way of evaluating how similar the behavior of the user is to that of the corresponding Gold Standard User for the condition.

ANALYSIS AND DISCUSSION

Our experimental manipulation consisted of 2 topics crossed with 3 levels of specificity. We first tested for evidence that the experimental manipulation resulted in differences in task difficulty. In order to answer this question, we tested whether the two independent variables (i.e., Topic Familiarity and Specificity) predicted differences in On-Topic ratings of the pre-search answers, since we expected that if users found a question more difficult to answer, and had less prior knowledge, they

[1] www.cs.cmu.edu/~nkgupta/SearchStudy/

[2] http://www.lemurproject.org/querylogtoolbar

would have a more difficult time producing an answer that seemed within the range of relevant answers. Using a binary logistic regression, we determined that Specificity did not have an effect on proportion of On-Topic pre-search answers but Topic Familiarity did such that the less familiar topic was associated with a higher proportion of off-topic answers (i.e., 42% for the less familiar topic as opposed to 30% for the more familiar topic).

In order to validate the use of KL divergence as a performance measure for search behavior and the write ups, a human judge read through each pre-search response entered by the users and marked whether the response was on-topic for the search query. Due to the relatively low level of English literacy of the users, some users did not fully understand what they were reading when they were reading in English. Presumably because of this, 35% of pre-search responses were deemed off-topic. In order to validate the use of KL Divergence as a performance metric, we computed an ANOVA with Specificity, Topic Familiarity, and Model Label as independent variables, On-topic as a random variable, and KL-Divergence as the dependent variable, in order to evaluate whether the on-topic/of-topic judgment on the pre-search answer predicted a difference in KL Divergence between the language models for the Gold Standard users and those of the participants. Language models for users whose pre-search answer was deemed on-topic had significantly lower KL-Divergence than those of the other users, which indicates that the sub-population of users who were not able to produce on-topic pre-search answers behaved in a way that deviated more from the Gold Standard users than that of the other users ($F(1,868)=16.05$, $p<.0001$, effect size .21 s.d.).

Both Topic Familiarity and Specificity had a significant effect on KL Divergence. But the most interesting effect was that of Model-Label on KL Divergence ($F(4,868)=281.4$, $p<.0001$), which did not interact with these task related variables. Since that is a main effect with no interaction with task related variables, we focus on that main effect analysis for the remainder of this analysis leading up to design recommendations. A post-hoc analysis demonstrated that not only did users with off-topic pre-search answers behave in a way that deviated more from the behavior of Gold Standard users, but their behavior became more deviant over the course of the activity. In line with this, their post-search answer deviated significantly more from that of the Gold Standard users than their pre-search answers did. Furthermore, their click behavior deviated more from that of the Gold Standard users than their query behavior. And their viewing behavior deviated more than their click behavior.

DESIGN RECOMENDATIONS AND CURRENT DIRECTIONS

In this paper we have presented an experimental study in which we have explored the specific needs of low literacy users in the developing world conducting a search task.

Our analysis suggests that support at the later stages of information seeking, such as when they are deciding which links to click on or when they are navigating to find specific information from these links is even more necessary than support at the query formation stage, where their has been a large focus. Furthermore, typical approaches to personalization where models used to fine-tune rankings are based on those links the user has clicked on in the past are likely to make the problem worse for these users rather than better. Thus we propose that instead, models of expert user behavior are used for the ranking rather than the personal models of these users. Other support for distinguishing relevant information from irrelevant information may also be necessary.

This study is a pilot effort contributing some new insights towards modeling low literacy information seeking behavior on the web. In our current work we are preparing to conduct a much larger study with 6,000 users with even less computer experience and lower literacy than the users from this study.

ACKNOWLEDGMENTS

We thank the administrative team at our collaborating university for managing the logistics of the Study.

REFERENCES

1. Downey, D. Understanding the relationship between searcher's queries and information goals. *In Proc. CIKM, 2008.* pp 449-458.
2. Chen, S.F., Goodman, J. An empirical study of smoothing techniques for language modeling. In *Proc. ACL, 1996.* pp 310-318
3. Kulback, S. The Kulback-Leibler distance. The American Statistician. 1987. 41: 340-341.
4. Kules, B., Capra, R. Designing Exploratory Search Tasks for User Studies of Information seeking supoort systems. *9th ACM/IEEE Joint Conference on Digital Libraries. JDCL 2009.*
5. Song, F., Croft, B. A general language model for information retrieval. In *Proc. ACM SIGIR*, *1999*. pp. 279-280.
6. Teevan, J., Dumais, S.T., Horvitz, E. Characterizing the Value of Personalizing Search. *In Proc. ACM SIGIR, 2007.*
7. Zhai, C., Lafferty, J, A study of smoothing methods for language models applied to ad hoc information retrieval. *In Proc. ACM SIGIR 2001*
8. Google User Statistics. http://www.comscore.com/Press_Events/Press_Releases/2009/1/Global_Internet_Audience_1_Billion
9. Agichtein, E., Brill, E, Dumais, S.T. Improving web search ranking incorporating user behavior. *In Proc. SIGIR, 2006.* pp 19-26.