

# Support for Web-based Information Seeking for Lifelong Learning in the Developing World

Naman Gupta and Carolyn Penstein Rosé, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, {nkgupta, cp3a}@andrew.cmu.edu

**Abstract:** This paper presents an analysis of a small user study investigating difficulties rural Indian users have with search technology in the context of an Information Technology course.

We present a qualitative analysis of the result write-up of the information-seeking task along with a log analysis of their corresponding search behavior. The goal of the analysis is to identify difficulties such users have with information seeking tasks so that appropriate support can be designed for them. Specific difficulties identified in the qualitative analysis indicate that students tend to simply ignore material they have trouble understanding, which leads to varying negative consequences. Our analysis demonstrates that we can extract predictors from logs of student search behavior that helps us identify where these students are having difficulty during the search task.

**Keywords:** Personalization, Low-literacy Users, Web-log Analysis, Information Seeking Task, Search strategies.

## Introduction

With the rise of the internet, both in terms of opportunities for on-line social interactions and learning oriented discussions as well as web-based information seeking, more and more learning, including both formal and informal learning, is taking place on-line. So far much of this has taken place primarily within the developed world. However, as internet penetration begins to make its way into less developed regions, different support needs will become salient. In order for the ideal of lifelong learning to be realized in the developing world, support for information seeking that targets the unique needs of students within these communities is needed. In this paper we present the results from a user study investigating the support needs of a low-literacy target student population.

Imagine a student from a rural area in India 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.

The users in our study are unique among studies of search behavior previously published (Agichtein et al., 2006; Duggan et al., 2008; White et al., 2009). Thus, while the study we present is based on a relatively small number of users, the analysis we provide is still a contribution in that it raises questions related to information literacy instruction for a student population with needs that have not yet been addressed with existing technology. The users in our study are 11<sup>th</sup> grade students from the Indian state of Andhra Pradesh who have come to study at a university developed as an outreach to the rural youth of that state. Students at this university were selected because they were the top ranking students in the village schools they came from. Many of these students had never seen or used a computer before coming to the campus. Furthermore, although they had studied English for 10 years prior to coming to the campus, more than half of them had done their schooling primarily in their mother tongue, Telugu. Each student who comes to study at this university is given a laptop, and most of the instruction is delivered in a computer supported fashion. All of its courses are conducted purely with English as the medium of instruction. Thus, the students who come to the campus are faced with two major challenges. First, they must adapt to the computer-based infrastructure, and second, they must adapt to English-based instruction. At the time of the study, the students were about to complete their first year at the university. During this year, they were provided Internet access for a short period of 1 month. So most students have a minimal experience of searching on the web.

As universities in the developing world, such as the one that provided the context for our research, begin to reach further into rural communities and provide opportunities for students from those communities to obtain a quality education and move to a more developed part of the country, the specific needs of these types of students will need to be addressed.

In the remainder of the paper we present the details of the study we conducted along with both a qualitative and quantitative analysis of the data. From this analysis we identify important questions for subsequent research and formulate some initial design desiderata for supporting web-based information seeking practices of our target student population.

## Method

### Experimental Procedure

The study was conducted with 11 participants chosen randomly from the students present on campus. 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: 10 minutes for completing a background information questionnaire, 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 Lemur Query Log toolbar.

The Experimental task itself was an exploratory information-seeking task based on the characteristics defined in (Kelly et al., 2009; Kules et al., 2009).

*Imagine that you have uncle in Pittsburgh who recently went to a dentist and was diagnosed with an abscess in his tooth. He had to undergo a painful treatment for the infection. You have to search for the necessary information on the Internet, in order to prevent your uncle from having a recurrence of the abscess or any other tooth disease in general.*

Before accessing any information online, they were asked to prepare a Pre-search write-up based on prior knowledge. Then using any search engine, they were told to prepare a Post-search write-up having all the information required for the given search task. No distinction was made in what search engine was used. This task was designed as an exploratory task such that in order to prepare a comprehensive and complete write-up, it elicited an extended search session. The task was also designed such that the students could relate to the given situation and has some educational content from a learning perspective.

### Tools and Materials

The following Tools and Materials were used for the experiment:

- A 3 page Web-based survey designed using [www.surveymonkey.com](http://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 was used to log all Search based activities performed during the Experiment.

Completely anonymized logs generated by the above toolbar were used along with the Survey Data. The details of the Log Data and Processing methods are as follows:

*Survey Data.* We collected a total 11 survey responses. These surveys contained the following details:

- Background Information – Unique ID, Type of High School, Medium of Instruction in School, Experience and Frequency with Computers, Frequency of using Search Engines.
- Subject 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 11 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 Survey and Search Logs, for the same Task Statement, from 4 high literacy graduate students at a top-tier US University.

## Qualitative Analysis

### Overview of Qualitative Analysis Approach

The task statement, presented during this information-seeking study, represents the level of English students at the campus are expected to be able to respond to in assignments they are given every day. Although the task itself might have been slightly difficult for them as compared to what they are used to. In order to respond to this information need, students must understand that what they need to find out is how to prevent tooth abscesses from occurring. While this would be obvious to any native speaker of English, this wasn't obvious to all of our participants because of their level of English competency. Some students did not comprehend the word "abscess", and thus did not focus their search or their answers on this tooth condition specifically. More frequently, however, they were able to understand that "abscess" was the tooth condition they were concerned about, but they missed that what they were supposed to search for was information related to prevention specifically. Because comprehension difficulties of one sort or another were identified in the answers provided by over 80% of our participants, we focused our assessment on this issue specifically, first analyzing their responses, and then analyzing their behavior in search of patterns that are indicative of specific difficulties.

Typically support for information seeking focuses on problems such as problem identification, query formulation, or information overload (Sutcliffe, 1998; Iivonen et al., 1998; Limberg, 1999). One might argue, then, that the research problem we are addressing is not information seeking per se. However, we argue that the comprehension difficulties we identified in our participants, beyond the obvious issues related to basic literacy, are also information seeking issues specifically. Students in this educational context must be able to receive instructions in English and respond in English using information technology such as search. We are not arguing that the basic literacy problems are not interesting and important research issues in their own right. We are simply arguing that learning to cope with the reality that students in this type of educational context must face before those literacy issues are addressed is a separate research problem deserving of a targeted effort.

Students' Pre-search and Post-search responses were first examined for evidence that they were searching for information related to tooth abscesses. They received one point if there was evidence in their Pre-search answer, and one point if there was evidence in their Post-search answer. Including information about prevention was more rare than information about abscesses, and almost always only occurred within the Post-search answer. Thus, students were given one additional point if either their pre or Post-search answer gave evidence that they were searching for information related to prevention. Altogether students could receive three points based on these criteria. In addition to these opportunities to receive positive points, we also assigned negative points for specific difficulties we identified. For example, students received one negative point if their answers were not primarily focused on prevention, even if they mentioned prevention. Students received a negative point if they included information that wasn't related to tooth problems in either their Pre-search or Post-search answers. Finally, if students gave evidence of not understanding some specific word needed to complete the task, they received another negative point. Criteria for assigning these positive and negative points are explicated in greater detail in the sections that follow, along with examples. All answers were rated first by a researcher who developed the criteria. That rater's analysis is presented in this paper. Instructions related to each of these points were written up and given to an independent rater not involved in this qualitative analysis. That rater applied the instructions to the same answers. The analysis was validated by computing a correlation between the total score assigned by each rater to each student. The correlation was .81.

Table 1: Coded Pre and Post Search Answers for All Participants.

Participant Number	Pre-search Mention Abscess	Post-search Mention Abscess	Mention Prevention	Not Primarily Prevention	Post-search Includes OffTopic	Demonstrates Specific Misunderstanding	Total Score
1	0	0	0	1	1	1	-1
2	0	0	0	1	1	1	-1
3	0	0	0	1	1	0	-.67
4	0	1	1	1	1	0	0
5	0	1	0	1	0	0	0
6	1	1	0	1	0	0	.33
7	0	1	1	1	0	0	.33
8	0	0	1	0	0	0	.33
9	1	1	1	1	0	0	.67
10	0	1	1	0	0	0	.67
11	1	1	1	1	0	0	.67

Table 1 presents a coded version of the pre and post search answers contributed by all 11 participants. As described above, students received one positive point for each of the first three columns and one negative point for each of the second three columns. The total score was then divided by 3 so that total score ranged between -

1 and 1. Participants are listed in Table 1 based on total score, starting with the least scoring students and ending with the best scoring students. The worst students received none of the positive points and all or most of the negative points, whereas the best students received most or all of the positive points and none or almost none of the negative points. Thus, based on the criteria we investigated, our participant population represents almost the full range of possible performance levels.

### Pre-search and Post-search Mention Abscess

The most basic criteria for demonstrating that a student was carrying out the search task correctly was evidence in the Pre-search or Post-search answers that information specifically about tooth abscesses was being searched for. 4 out of 11 answers did not provide any such evidence. We see here that in order to understand this data, we must triangulate between pre and Post-search answers as well as the search behavior itself.

Most interesting was student 8's response, which was about prevention of tooth problems, but never mentioned anything about tooth abscesses specifically.

[Pre-search answer] I would better advice him to brush twice daily. I will ask him to stop eating very sweet eatables like toffees, candies and harmful substances like tobacco and so on. From any of my friends, who is a doctor, I will be collecting some more information [remainder of answer omitted...]

This is actually a good answer since the main advice available on the web for preventing tooth abscesses is to observe good dental hygiene in general. However, it is not possible to verify from this answer that the student was specifically looking for information related to tooth abscesses and not tooth issues in general. His Post-search answer did not mention anything about abscesses specifically either, and neither did any of his queries, which included only general tooth issues, specifically "tooth pastes", "toothpastes", "tooth dentist", "doctors", "preventions for tooth decay". It appears from the progression of queries that this student started out thinking about general tooth care in order to identify types of problems that may occur if one does not properly care for teeth, and then discovered that an English word for such difficulties was "tooth decay", and thus searched for prevention of tooth decay rather than prevention of tooth abscesses specifically. What this suggests as a possible student strategy is to ignore all but the parts of instructions that are comprehensible, try to make sense of the comprehensible pieces, and then proceed from there. This is a strategy that has been documented in the literature on language contact between speakers of different languages when they must cope with their lack of shared language in order to work together (Hatch, 1983). The simplistic progression from "tooth pastes" to "toothpastes" was also a pattern we observed frequently, and which has been documented in the literature on search for novice users (Holscher et al., 2000). Such users, when they are not satisfied with the response they get from search technology do not necessarily have effective strategies for revising their queries in order to obtain a more satisfying result.

Student 2's Pre-search answer was the furthest off from the expected focus, though it does offer some evidence that the student was thinking about how to approach the task:

[Pre-search answer] I can get answers from [www.answer.com](http://www.answer.com). I can prepare for the exminations from [www.extramarks.com](http://www.extramarks.com). I can use online dictionary. I can read news. I can study many books in onlline. I can get health tips, study tips, language tips.

Other aspects of the student's answer and behavior cast doubt that the student understood that he was supposed to be searching for information to help his uncle and not to prepare for an exam. For example, his queries reveal that he was far off from performing the task as intended, specifically "income", "study skills", "chemistry", "study for engineer", "learning", "learn more", "my study", "physics controls", "boilogy books", etc. This could be viewed as an extreme case of students responding to only the portions of the task that they were able to comprehend and piece together.

Student 10's Pre-search answer did not mention abscesses and furthermore focused on treatment rather than prevention.

[Pre-search answer] I would like to take him to a doctor. I give him medicine. I will going to be take care of his food. I pass this to my family members and I take suggestion from them and I search for more information in internet if I have a chance.

However, this student's Post-search answer demonstrated that in fact he did understand that he should be looking for information related to abscesses and perhaps just did not know anything specific about tooth abscesses prior to the search task. This seems to be the most probable reason for the students (8 out of 11) not mentioning abscesses in the Pre-search answers since many more students (i.e., all but 4 out of 11) mentioned

abscesses in their Post-search answers. This is an encouraging sign, then, since it suggests that this population may be capable of using search technology to address knowledge gaps, although perhaps in a degraded manner due to their literacy problems.

### Mention Prevention and Not Primarily Prevention

While as mentioned above, student 8 did not mention abscesses but did mention prevention, a more frequent error, occurring in almost half of the cases, was that prevention was never mentioned either in the Pre-search answer or the Post-search answer. In these cases, students did not receive a positive point for mentioning prevention. These students tended to focus on treatment. We suspect a probable reason for this is that the overwhelming majority of information that is found when one queries “tooth abscess” on the web is related to treatment rather than prevention. Nevertheless, the fact that five students mentioned both abscesses and prevention within their answers shows that it is possible for this student population to find information about prevention of tooth abscesses on the web, and therefore, the task itself was within a reasonable scope for this population. The pattern of discussing treatment of tooth abscesses without addressing the prevention issue is consistent with the hypothesis that students responded to the portion of the task description that they understood and ignored the rest, which was introduced above. Of the five students who never mentioned prevention either in their Pre-search or Post-search answers, none of them ever literally mentioned prevention in any of their queries either.

On the other hand, all of these students made multiple query attempts. Thus, there is also another possible explanation. The fact that they made multiple query attempts offers evidence that they were not quite satisfied with the results they obtained with their initial queries. It is possible that they realized their search results were not addressing the search task fully but we not sure how to get it to give them the information they wanted, even though the word “prevention” literally shows up in the task description, which they could have simply copied from the task description itself.

Here is one query progression, from student 6, that suggests a severe lack of ability to formulate a query appropriately, specifically “He had to undergo an painful treatment for the infection”, “MY UNCLE HAS A TOOTH PROBLEM WT CAN I DO FOR HIM”, “MY UNCLE HAS A TOOTH PROBLEM WT CAN I DO FOR HIM”, “dentist and was diagnosed with an abscess in his tooth”. It’s clear from this progression that the student understood what the problem was with his uncle’s teeth. However, the student did not effectively articulate the information need, in relation to this problem. The closest thing mentioned was, “what can I do for him”. However, the queries that suggest prevention were phrased in a much more indirect way than those a more experienced user would attempt using a search interface. More importantly, however, the queries that suggested prevention were earlier rather than later in the progression, and thus, while the student may have recognized that information about prevention was not forthcoming in response to “what can I do for him”, a further attempt to get information related to prevention was not given.

All of the other query progressions appeared much more typical in style as queries. For example, consider the queries from student 1, which demonstrate a typical, very terse query style, even though they demonstrate some comprehension difficulties, which will be addressed in the next section: specifically, “uncle”, “allergies”, “uncle”, “tooth paste”, “tooth disease”, “about google search engine”, “about yahoo”, “about MSNB”, “about MSN”, “oral health”, “oral health care”, “oral health diseases”, “www.yahoo.com”, “tooth disease”. The fact this student used several queries related to search engines rather than the task in the middle of the progression suggests that this student realized he was having trouble with the task. Nevertheless, the fact that he mentioned several different search engines suggests he had some broad awareness of search technology coming in to the task.

Even when students mentioned prevention either in their Pre-search or Post-search answers, it was frequently not the main focus. In cases where it was not the case that either the complete Pre-search answer or the complete post search answer could be construed, even using very liberal criteria, as focusing mainly on prevention, students received a negative point. This occurred in all but two cases. If the problem of never mentioning prevention was due to students ignoring that part of the task description, then this problem must be indicative of a different cause, unless the information included about prevention was included by chance. Student 7, for example, did not receive the negative point because altogether the answers and behavior suggest a more skillful attempt than we see in most other students in our participant population. His behavior and answers offer supporting evidence for the conjecture that a problem other than ignoring part of the task description is going on in the case where students mention prevention but don’t focus on prevention. In the case of student 7, the Pre-search answer was almost a perfect answer even without search:

[Pre-search answer] In order to prevent the recurrence of tooth disease one should brush the teeth twice a day. One should change the brush for every two months. one should visit dentist twice a year.

Furthermore, the one query given focused on prevention, specifically “How to eradicate the tooth diseases”. Neither the Pre-search answer nor the query specifically mentioned abscesses, but the post search answer focused specifically on abscesses, and thus although that mainly focused on treatment, altogether this student’s answer and behavior demonstrates a good understanding of the task and facility with search technology, although the Post-search answer is not perfect because of its focus on treatment:

[Post-search answer] One tooth abscess treatment option, is the use of antibiotics, when the tooth is already dead and sealed by a dentist. The antibiotics kill the germ responsible for the tooth abscess and the body then repairs the bone and tooth, normally pretty good. The second tooth abscess treatment, is root canal treatment, which is the drilling out ... [remainder of answer omitted...]

Thus, in this case we see that even if a student understands the task completely, the Post-search answer may still fail to meet the requirements of the task. In this case, the problem may be distinguishing relevant information from irrelevant information. This student only issued a single query. If multiple queries give evidence of dissatisfaction with search results, we do not see any such dissatisfaction here.

### Post-search Includes OffTopic and Demonstrates Specific Misunderstanding

So far we have focused on missing information in student answers. Another issue was where inappropriate material was included or where the behavior or answer gave evidence that a word from the task description was explicitly misunderstood. In response to both of these situations, a negative point was assigned to students, although the second case was rare, only being assigned twice. An example of the first case was student 1’s Post-search answer, which was vaguely on the task topic, and was clearly focused differently than desired:

[Post-search answer] There is now a substantial literature concerning the concept of health and its application in dentistry in which various theoretical approaches and conceptual frameworks are discussed. 1, 3-9 Consequently, many of the basic conceptual issues involved in this field will be familiar. For example, the limitations of using clinical disease-based measures ... [remainder of answer omitted...]

Student 1, whose query progression was mentioned in the previous section, was one of the students who demonstrated a misunderstanding of the vocabulary used in the task description. Notice that his first and third queries were both “uncle”. His answer after search makes it clear that his reason for this query was that he did not know what “uncle” means (note that this answer was inserted in the Pre-search answer slot in the form, but it’s clear that the student used information from the web to answer the question):

[Pre-search answer] Good oral care starts from the beginning of your uncle's life. Even before his first teeth emerge, certain factors can affect their future appearance and health. For instance, tetracycline, a common antibiotic, can cause tooth discoloration. For this reason, they should not be used by nursing mothers or by expectant mothers in the last half of pregnancy. Since uncle teeth usually emerge around six months ago, standard oral health procedures like brushing and flossing aren't required for infants. However, infants have special oral health needs that every new parent should know about. These include guarding against uncle and making sure your uncle is receiving enough fluoride. <http://www.cochrane.org/>

This student recognized that he did not understand a term. However, his behavior makes it clear that he was not aware of how to deal with this problem using web technology. He made two attempts using the search interface. However, he was not able to clear up his difficulty. A more detailed analysis of his click behavior revealed why. In typing in the query “uncle”, the student ended up on some pages related to “crying uncle”, which is an idiom in American English that is used when kids are planning a game and one child wants another child to stop doing something that hurts him. Problems like this, that indicate a lack of ability to address lack of understanding even when it is detected and not simply ignored should be noted when formulating plans for support.

### Discussion of Qualitative Analysis

From this qualitative analysis, we identify several potential problems that we plan to investigate further in future work. Two issues we identified above have already been reported on the literature related to search for novice users, specifically problems with query formation and problems with query reformulation. However, some issues appear to be specific to low literacy or low comprehension users, in some cases showing reminiscent

patterns to those reported in the literature on Creole formation and communication between populations of people who must communicate for business purposes but do not speak the same language (Hatch, 1983), in particular, ignoring part or all of the task description and inability to distinguish relevant and irrelevant information. In the next section, we will present a quantitative analysis confirming some of these conjectures.

## Quantitative Analysis

In this section we present a quantitative analysis of our data, to follow up on some of the conjectures we formulated in the course of our qualitative analysis.

Recall that we collected data from 11 students and 4 gold standard users, which we used to evaluate the behavior of the student users. The first preprocessing step was to construct a set of models representing the behavior of each user as well as combined models representing the behavior of the set of gold standard users. One type of information captured by language models is the distribution of words in the Web pages looked at by that single user or group of users.

We defined 3 basic types of User actions – Query (Q), ClickResults (R) and ClickResults+Navigation (R+RN). The Search behavior for a participant was represented as the percentage of time spent on each above defined User actions. In our quantitative analysis we first explored whether these behavior distributions predicted the pattern of scores we observed in the qualitative coding of the task results discussed in the qualitative analysis above. One stable finding was that the percentage of time spent on ClickResults correlated with the number of positive points assigned ( $R^2=.3$ ,  $p<.05$ ). Students who had more difficulty with the task showed a more even distribution over the types of tasks rather than investing the majority of the time on the search results. This might be indicative of “random” behavior of confused students, or students who are struggling with the task and feeling unable to make good use of the results of their queries, either because they weren’t able to formulate effective queries or because they weren’t able to comprehend the results properly, or without considerable effort. For these students, the standard deviation across probabilities assigned to specific types of actions was low because there was no clear preference for one behavior over another. Because of this, standard deviation across percentage of time spent on the different user behaviors had a marginal negative correlation with number of negative points assigned in the qualitative analysis of search products, specifically the pre and post search write-ups, such that students who had low standard deviations across these percentages tended to be assigned larger numbers of negative points in the analysis of their search products. This analysis shows that the behavior shows promise that regular patterns can be identified in search behavior that can enable automatic identification of users who are struggling with their search task.

We also explored evidence in the text of the documents retrieved itself for evidence of specific difficulties with the task. For each user as well as the set of gold standard users, we built 3 different language models that represented distributions of words found on the pages retrieved in response to the 3 types of user behaviors: Query (Q), ClickResults (R) and ClickResults+Navigation (R+RN).

Once we have the language models, we can evaluate the behavior of students by comparing their language models to those of the group of gold standard users. A pair of 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. Language models for each user in a study were compared using KL divergence (Kulback, 1987) with combined Gold standard Language model of the 4 Gold Standard Users. 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 those of the Gold Standard Users. Since the task defined in this user study is an elaborate Information seeking task, the usual methods of IR evaluation are not applicable here (Kelly et al., 2009). So we make an assumption about the ‘perfect’ search behavior for the highly educated and experienced Gold standard Users in our study. This is reasonable assumption in this context and it allows us to compare the Search behavior patterns in the two User populations and identify relevant variations during such an information-seeking task.

The models were built were based on distributions of words found on the pages users spent time looking at after different user actions during the search session. The description of the 3 models computed for each user as well as the Gold Standard users based on the three based types of user actions is as follows:

- AllSearchResultsModel – includes the content from all the top 10 search results returned in response to each of the queries issued by the user. 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 that were clicked by the user. This is to evaluate the user’s ability to choose a relevant result from the results page.
- ClickedResults+NavModel – includes the content from the above clicked results pages along with the subsequent navigated pages. This is to evaluate the user’s ability to effectively navigate through the Clicked results to find the relevant information.

The most predictive of the three models was the ClickedResultsModel, which had a moderate negative correlation with the number of positive points assigned to each student ( $R^2=.41$ ,  $p<.05$ ), such that students with

higher divergence scores earned fewer positive points. The correlation with negative points and total points was weaker, but still significant. What this demonstrates is that we can make predictions about students, when we know their task, about whether they are heading for a high quality result based on the distribution of words found on the pages they click on. Taken in combination with the results of the probability distributions associated with student actions, we have some evidence that we can distinguish users who are struggling from users who are achieving some success in their task.

## Conclusion and Current Directions

In this paper we have presented a qualitative and quantitative analysis of data collected in a small user study investigating web-based information seeking behavior of low-literacy users in the developing world. Our results show promise that we can identify users who are struggling with their search task. We also identified some specific difficulties that users have and possibly counter-productive coping strategies, such as ignoring portions of task descriptions that are not clear to them. We have two specific directions for future research.

First, we plan to conduct a much more extensive study, with a much wider variety of tasks and a much larger population of users. It is a limitation of this study that it is based on only 11 students. Nevertheless, even within this small sampling of students from our target student population, we see a wide variety of levels of success, and we identify some predictors that are easy to extract from the logs of their search behavior that predict how much success they are having in their task.

Our second main direction for future research is to develop specific support technology that can be triggered for students who are identified as having trouble. For example, we can trigger dictionary pop-ups so that students who are struggling with some words will be more likely to find useful definitions rather than misleading web pages related to synonymous idioms. We could provide students with an interface to type in an extended information need description, such as their task description, which could be used to identify from their search behavior which portions they might be ignoring.

## References

- Agichtein, E., Brill, E., Dumais, S.T. (2006). Improving web search ranking by incorporating user behavior. In Proceedings of SIGIR, 19-26.
- Chen, S.F., Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In Proc. ACL. pp 310-318.
- Duggan, G.B., Payne, S.J. (2008). Knowledge in the head and on the web: Using topic expertise to aid search. In Proc. SIGCHI, pp. 39-48.
- Hatch, E. Simplified input and second language acquisition. (1983). In R. Anderson (Ed) *Pidginization and Creolization as Language Acquisition*. Newbury House Publishers.
- Hölscher, C., Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 33 no. 1-6, pp. 337-346.
- Iivonen, M. and Sonnenwald, D.H. (1998). From translation to navigation of different discourses: A model of search term selection during the pre-online stage of search process. *Journal of the American Society for Information Science*, Vol. 49, No. 4, pp. 312-326.
- Kelly, D., Dumais, S., Pedersen, J. (2009). Evaluation challenges and directions for information seeking support systems. In *IEEE Computer*, pp. 60-66.
- Kules, B., Capra, R. (2009). Designing Exploratory Search Tasks for User Studies of Information seeking support systems. 9th ACM/IEEE Joint Conference on Digital Libraries. JDCL.
- Limberg, L. (1999). Experiencing information seeking and learning: A study of the interaction between the two phenomena. In *Proc. of Information Research*, Vol. 5 No. 1.
- Song, F., Croft, B. (1999). A general language model for information retrieval. In *Proc. ACM SIGIR*, pp. 279-280.
- Sutcliffe A. and Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with computers*, Vol. 10, pp. 321-351.
- White, R.W., Dumais, S.T., Teevan, J. (2009) Characterizing the influence of Domain Expertise on Web Search Behavior. In *Proc. WSDM*.
- Zhai, C., Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. ACM SIGIR*.

## Acknowledgments

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