discuss how we have utilized such an approach and then used this feature based representation to cluster users into sets that represent different overall approaches to the search process. Another way of modeling the result of search over time is to build language models from collections of documents encountered during the process. This can be the sum total of documents referenced in the top n documents returned from each query, all the links clicked on, all the links users have spent a substantial amount of time on, etc. Using this approach, it is also possible to compare search behavior between users or between sets of users, but combining language models within sets and comparing those models across sets. Later in the paper we will discuss how we have used such a technique to compare the search success of our target student population with that of more skilled users as well as focusing on our target users, and comparing across different tasks and different support conditions.

In order to make sense of sequences of click behavior, a first task is frequently segmenting the logs into sessions devoted to individual information needs. Most such work [Agichtein et al. 2006; Metzler et al. 2007] involves analysis of large-scale query logs. However, identifying the boundaries of a query session for a particular information need is hard task in such logs. Metzler et al. [2007] describe a methodology for segregating queries for a specific information need, by computing the similarity between two queries using lexical, morphological, and probabilistic modeling methods. Downey et al. [2008] highlight the importance of applying more complex techniques for the task of identifying session boundaries in query logs. In order to enable understanding of the various qualitative aspects of user search behavior for large scale query log data, as Kellar et al. [2008] describe, it is critical to augment the log data with contextual information. We have avoided needing techniques like this in the work presented in this paper since each student participated in the study for exactly one session.

Needs of Novice Users and Low Literacy Users

The work we report on in this paper fills a gap in the literature on search. Prior work has focused either on use of information technology of illiterate or extremely low literacy users on the one hand, or relatively high literacy users who may simply be novices with respect to search technology. Our focus, in contrast, is on users with moderately low literacy and English competency and moderately low computer literacy. Thus, our target users fall in between two extremes represented by prior work. We refer to the first body of prior work as analyzing search behavior of novice users and the second for low

literacy users. In contrast, we refer to our work as analyzing search behavior of emerging Internet users.

First we review the literature on search for novice users. This work typically contrasts the behavior of novice users with that of more typical "expert" users. For example, Zhang et al. [2005] evaluated the relationship between domain knowledge, search behavior and search success. The performance of search specialists, domain specialists and novices was also compared in Marchionini et al. [1989]. Most such research has shown differences in search behavior and success as a function of domain expertise and search experience. Holscher and Strube [2000] showed the heterogeneous needs and capabilities of search-engine users, which have to be catered to differently. They showed the ineffectiveness in query reformulation and navigation strategies of novice users during information seeking activities. In a search study with adolescents, Guinee et al. [2003] highlight their tendency to fall back on familiar cognitive paradigms: starting with their previous knowledge and adhering to time-tested practices learnt earlier. This demonstrates the potential for building search support that begins by using this prior knowledge as a foundation for scaffolding information seeking.

Work on search for low literacy users has focused on users who have much more rudimentary English and technical skills than our target users. In a study in rural school setting, Henry [2005] identified the importance for higher order thinking skills for using Internet successfully, apart from the fundamental literacy skills such as reading and writing.

It is important to keep in mind the contrast between our user population and that of users in typical evaluations of IR systems when we consider the impact these evaluations have had and are continuing to have on the direction the IR community takes in further development of the technology. Most of the current models in information retrieval [Ponte and Croft 1998], which are embedded in popular search services, build in the assumption that users are able to effectively distinguish 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 non-native English speaking populations, and thus a targeted effort is necessary to assist such populations to search efficiently and effectively [Birru et al., 2004]. 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.

Cross-Linguistic Information Retrieval

It is clear that one of the major issues facing our target user population is low English competence. The main support for Web search for non-native English speakers is Cross-Linguistic Information Retrieval (CLIR). In a study with 10 non-native English speakers, Aula et al. [2009] observed preference of querying in the native language by the users. When searching in the non-native language (ie. English), they found it difficult to do query reformulation and refinement when the original query was not successful. One might propose, then, that a potential solution is to make use of technology developed for the purpose of CLIR, where a search query is issued in the native language of the user, translated into the language of the target materials, and the results are presented in the user's language or the language of the materials [Kraaij et al. 2003; Lavrenko et al. 2002]. When effective, this technology would indeed at least address some problems related to query formation in a language in which the user has low competence. However, evaluations of CLIR have demonstrated that, while it is partly successful, there are major issues with ambiguities in search terms that degrade the search performance when queries are translated [e.g., Gao et al., 2002]. Furthermore, as we will demonstrate, our target users have other difficulties besides this. And thus, it is not a complete solution.

Later in the paper we will propose and provide some preliminary evaluation of an alternative approach. We aim to explore the potential of translating the natural language description of the information need itself in the desired language with the aim of providing necessary cue words allowing the user to formulate a query in the desired language themselves - English in our case. In this way, the process is more transparent to the user, and if the translation fails, the user is better able to recover by rephrasing the information need and trying again. The assumption is that the user is more likely to recognize a bad translation, which only requires recognition of whether the English and Telugu (in this case) match, than to produce a translation independently, which either requires the ability to understand the English unaided or generate English unaided.

3. STUDY 1: QUALITATIVE ANALYSIS OF INFORMATION SEEKING DIFFICULTIES

Because our population of emerging Internet users is distinct from populations that have been observed in connection with information seeking technology in the past, we first

present a qualitative analysis of a small number of users in order to illustrate the types of difficulties that we aim to further specify through a large scale quantitative analysis presented later.

Experimental Procedure

The study was conducted with 11 participants chosen randomly from the students present on campus at Rajiv Gandhi University for Knowledge Technologies (RGUKT), Andhra Pradesh, India. 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 experimental session extended for 1 hour: first, 10 minutes for completing a background information questionnaire followed by 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 recorded log files using the Lemur Query Log toolbar[2].

The experimental task itself was an exploratory information-seeking task based on the characteristics defined by Kelly et al., [2009]; Kules [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, after the search, they were told to prepare a post-search write-up having all the information they found for the given search task. They were asked to use the popular Google search engine to conduct the search. This task was designed as an exploratory task such that in order to prepare a comprehensive and complete write-up, an extended search session would be required. The task was also designed such that the students could relate to the given scenario based on their prior experience and that it would have some educational content so that it would be worth spending their instructional time.

[2] www.lemurproject.org/querylogtoolbar. This software is used to record all click behavior in a browser so that search behavior can be carefully monitored during the study and detailed logfiles can be kept for later analysis.

Tools and Materials

The following Tools and Materials were used for the experiment:

- A 3 page Web-based survey 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.
- The Firefox browser compatible with both Windows and Ubuntu systems was used for the experiment.
- The 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 produced during the search task. The details of the log data and processing methods are as follows:

Survey Data. We collected a total of 11 survey responses. These surveys elicited the following information:

- Background Information – Unique ID, Type of High School, Medium of Instruction in School, Experience and Frequency with Computers, Frequency of using Search Engines.
- 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.

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 seeking task adequately, students would need to be able to 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 was not 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 search behavior to look for patterns that indicate specific difficulties.

Typically support for information seeking focuses on issues such as problem identification, query formulation, or information overload [Sutcliffe and Ennis 1998; Iivonen and Sonnenwald 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. 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 rarer 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 was not 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 0.81.

Table I 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 1 appearing in the first three columns and one negative point for each 1 appearing in the last three columns. The total score was then divided by 3 so that total score ranged between -1 and 1. Participants are listed in Table I 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.

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

St. #	Pre-search Mention Abscess	Post-search Mention Abscess	Mention Prevention	Not Primarily Prevention	Post-search Includes OffTopic	Specific Error	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

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. Four out of eleven 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 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 and Strube 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. Two other students showed a similar pattern to student 8.

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 is 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.

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.

Post-search Includes Off-Topic 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 is 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 investigate further in the quantitative analysis presented in the next section. 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 our target 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 and offering further insights on how me might support this user population with technology.

4. STUDY 2: QUANTITATIVE EVALUATION OF POTENTIAL SUPPORT

Recall that our emerging Internet users are not native speakers of English, and yet they are required to do their school work in English. Furthermore, about 56% of the available content on the web is in English [Netz.tipp.de 2002], though this percentage is decreasing over time. Nevertheless, in order to make use of information available in English, they are required to interact with it in English. We are considering two scenarios with our target users:

- Case 1: They are presented with well-formed information needs in English, for which they have to prepare a report using web search. This can be difficult for them if they do not understand the information need perfectly.
- Case 2: They have a personal information need, where they might have trouble articulating that need in English in order to search on the web.

Both of these scenarios present challenges for our target users, and our goal is to gain insight into how we can support these users in their information seeking. We thus explore the potential of translating the information need from English to Telugu in case 1 and Telugu to English in Case 2. The purpose of the translation would be to foster a better understanding of the information need and provide necessary cue words to the user allowing him to formulate his own query (in English) using the translated information need. Note that in this study we do not address any of the technical issues related to producing the translation, but only evaluate the effect of providing the translation.

The dataset we explore was conducted in the context of an experiment designed in response to the findings from the qualitative analysis presented in the previous section. There we noticed that our target students had a tendency to skip over portions of a task statement that they didn't understand or others for which they were not able to reliably evaluate the relevance of the information found in response to their attempted queries. Our hypothesis was that providing users with a translation of the information need would allow them to triangulate the results of their search and thus overcome some of this difficulty. The students were asked to search for information on the web when they were presented with search task statements in different language combinations – only Telugu (condition A); only English (condition B); both English & Telugu (condition C).

In condition A, when the search task is given in just Telugu, the hypothesis was that the students (with Telugu as their native language) have a complete understanding of the information need, but due to their low English proficiency they might face some

difficulty in issuing queries in English that are representative of the information need and in evaluating the results returned by the search engines, since they are also in English.

In condition B, the search task was given just in English, just as in the qualitative study. In this case, we expect the students to have only a 'reasonable' understanding of the information need and to have a tendency to either skip what they do not fully understand or possibly use the cue words in the task description for picking query words for the search task. That means that even if they do not have complete understanding of the information need, it is possible that they issue 'reasonable' queries, which could possibly return documents relevant to the information need. However, they would likely have trouble fully evaluating the relevance of the information returned.

In condition C, the search task is presented in both English and Telugu. This will not only allow complete understanding of the search task, but also allow them to map important keywords across the two task descriptions. We initially expected students to be most successful in this condition. For each variation, we estimated the students' search effectiveness using a range of language models based on their post query search behavior and search task response. One motivation for doing such a translation would be to improve their active vocabulary and obviate the need for such a translation over time, so that they can search more effectively and efficiently. However, we do not formally investigate this vocabulary learning issue in this study.

4.1 Motivation

As highlighted earlier, most previous research has shown differences in search behavior and success as a function of domain expertise and search experience. In this paper, we aim to understand the behavioral differences between non-native users and more experienced educated users on a more detailed level. We also explore the search effectiveness of non-native speakers, when they are given the information need in different language combinations, including their native language, the target language, or both. Most of the prior research, done on large-scale query logs, either work without reference to a specific information need or make unreliable estimates of the information need, and also work with unreliable estimates of query session boundaries, all of which add noise to the model of user behavior they are able to obtain from that work. In our research, we conduct a study with well-defined information-seeking tasks to be completed within a controlled time duration, which allows us to make very controlled comparisons across conditions. At the same time, our study has a large user base with approximately 2000 participants, which is much higher than similar user studies done in

related research in personalization [Holscher and Strube 2000; Bhavnani 2005; Kelly and Cool 2002], which gives us sufficient statistical power to draw reliable conclusions from these controlled comparisons.

4.2 Method

We conducted a large-scale user study with approximately 2000 participants, who were given an elaborate information-seeking task, ensuring that we could have precise knowledge of both the information need and the boundaries of the search session.

Participants. The participants in our study are unique among previously published studies on search behavior [Agichtein et al. 2006; Duggan and Payne 2008; White et al. 2009]. The major highlight of this user study is the absolute size of the participants group. As in the earlier study, the participants in the study were 11^{th} grade students from the Indian state of Andhra Pradesh who have come to study at a university, initiated as an outreach to the rural youth of that state. Most of these students had never seen or used a computer before coming to the campus. Though most of them (83%) had done their schooling primarily in their native language, Telugu, they had studied English as a second language for almost 10 years prior to coming to the campus. Remaining students had English as the medium of instruction in prior schooling. A more comprehensive background characterization of the students is shown in Table II. Each student at this university is provided with 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. These students perform all educational activities like homework, exams and reports in English. 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 our study, the students were about to complete their first semester at the university. During this time, they were provided daily computer access with no Internet access. So most students had minimal prior experience with searching on the web.

Experimental Procedure. The study was conducted during regularly scheduled Information Technology classes at the university during one-week of time. The students were divided into 40 classes with 50 students each. Due to some absentees, a total of 1910 students participated in the experiment.

To ensure the experimental integrity, an instruction video was shown to all the participants giving a short introduction of the study – explaining the purpose and

motivation behind the study. The instruction video was in Telugu, to ensure that the participants understood all the instructions without any misunderstanding. Since many of the students did not have prior search experience, a brief walkthrough of a sample search task was given, illustrating the basic application of search engines – issuing a query, glancing through the results list, clicking on results, navigating further on the clicked results.

The experimental survey extended for 1 hour duration altogether: 10 minutes for installing a search activity logging toolbar[3], 10 minutes for completing a background information questionnaire, 10 minutes for understanding the information seeking task and completing the pre-search write-up. They were then given another 25 minutes for the search activity and subsequent task write-up. Once finishing the survey, the participants uploaded the log files recorded by the toolbar. We required students to install the toolbar themselves since they were working on their own laptop computers. This was a requirement of the university. As part of the instructions for installing the toolbar, we made sure the students were aware of how to uninstall and/or disable the toolbar so that their search behavior would not continue to be monitored beyond the study session.

Table II. Background Characterization of the participants

Background	**Values**	**Number of participants**	**Percentage of participants**
Language of Instruction in School	English	302	15.8%
	Telugu	1594	83.5%
	Others	14	0.7%
Computer Experience	0-3 Months	1548	83%
	3-12 Months	124	6.5%
	1-2 Years	88	4.6%
	2-5 Years	56	2.9%
	Above 5 Years	57	2.9%
Frequency of Computer Use on Campus	Never	1027	53.4%
	Rarely	487	25.3%
	Weekly	331	17.2%
	Daily	78	4.1%
Web Search Frequency	Never	1490	77.5%
	Rarely	289	15.0%
	Weekly	115	6.0%
	Daily	28	1.5%

Experimental Task and Manipulation. The experimental search tasks themselves were exploratory information-seeking tasks based on the characteristics defined in [Kules

[3] www.lemurproject.org/querylogtoolbar

2009]. 5 different topics of interest were chosen with the information need presented at 2 levels of specificity – low & high. The search topics were estimated to be of varying interests to different sub-population of the participants – Cricket (Male-oriented); Food (Female-oriented); Fluid dynamics (curriculum-oriented); Movies (non-curriculum oriented); Eye diseases (general). Search task specificity has been characterized in a variety of ways – 'open' and 'closed' tasks [Marchionini 1989], 'general' and 'specific' tasks [Qiu 1993]. We define the following terminology:

- Low specificity (LS) – pertaining to a broader and more abstract formalization of the search task, abstracting over more than one more specific instantiation of the same task, such as pertaining to the customary foods of a country rather than a specific region.
- High specificity (HS) – pertaining to a more focused and concrete formulation of the task.

Based on prior work [Kim and Allen 2002], we expected HS search tasks to require fewer viewed pages and less time devoted than LS search tasks. This would be consistent with the idea that HS tasks might require less effort to complete than LS tasks and hence might be easier. The tasks were designed to be similar to regular course assignments the students were accustomed to working on in terms of the level of English and approximate difficulty. The search task statements were presented to the participants according to the following 3 variations - only English (condition A); only Telugu (condition B); both English & Telugu (condition C). This gave a total of 30 search task variations altogether, a 3 (Language) X 2 (Specificity) X 5 (Topic) factorial design. The Telugu translation of the search tasks was done by a human annotator and verified with translation back to English by a second human annotator. Each student was assigned 1 out of the 30 search task variations, while ensuring equal number of participants for each variation.

4.3 Data

Completely anonymized logs generated by the Lemur query log toolbar were created as students worked through the survey. The details of the Log Data and Processing methods are as follows:

Survey Data. A total of 1910 survey responses were collected over the 40 study sessions after factoring out absentees & spurious and duplicate responses during pre-processing. 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.
- Pre-Search and Post-Search Write-ups.
- Self-reported Topic Familiarity and Search Task Difficulty.

Activity and Search Log Data. From the 1910 students, we collected only 1422 activity and search logs using the lemur toolbar. This number of was significantly lower than the actual number of participants due to a variety of logistical issues. These logs contained the following types of information about search behavior:

- **Search Related** – Details (query string, timestamp) of all queries issued. Details (result rank, URL, timestamp) of results clicked from query results page.
- **Viewed Pages** – Details (URL, content, time on page, timestamp) of all the pages viewed.
- **Browser Events** – Details (RClick, add/close new tab/window, scroll events) of any browser activity during the experiment. This allowed us to build a sequence of events during the search session.

Gold Standard Data. Because of the large size of the dataset, it was not possible to obtain human relevance judgments of the search queries, the clicked results or the survey post-search write-up for the obtained responses. We estimate each of the above search behavior characteristics, with appropriate language models described in the next section. For this purpose, we conducted another study with 30 high literacy students in a top-tier US University. These students, due to their high English proficiency and high computer & web search experience, can be expected to find the most appropriate information for the given search tasks. They were given the task statements in Condition B, English only. For each of the 10 search tasks, we collected data from 3 such 'gold standard' users. We understand that because of the exploratory nature of the search tasks, there is no perfect sequence of behavior or answer to the task. Hence we use the combined search behavior of the 3 users as the estimated gold standard data for the task. This builds in some tolerance to acceptable differences in behavior to the search effectiveness measurements we make using this gold standard.

4.4 Log Analysis Methodology

In order to estimate the effectiveness of the search behavior of the participants, we compared the various actions performed by the participants during their search session with those of the gold standard users described above. The possible actions for the users are issuing a search query (Q), clicking on a search result (R), and subsequent navigation on the clicked results (RN). We can estimate the quality of these actions in terms of how relevant the information returned was in comparison with that obtained by the gold standard user actions. For example - the relevance of the queries issued can be estimated by an evaluation of the concatenated text of the top 10 results. This was done in two ways: (i) using just the title and snippets concatenated for the top 10 results; (ii) using the complete text of the top 10 results. This was similar to the behavior based query similarity measure described in [Balfe and Smyth 2005]. Intuitively, the snippets & title text seems more representative of the queries because of the query-biased snippets generated by the search engines today. Similarly, the ability to identify relevant results for the given information need can be estimated by either the snippet & title of the clicked results; or by the whole text viewed by the user. Hence, for each user response including the gold standard responses, we built 6 different unigram language models [Ponte and Croft 1998; Chen 1996] with commonly used Laplace smoothing [Zhai and Lafferty 2001]. 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 description of the 6 models computed for each user as for well as the Gold Standard users is as follows:

- **QueryResultsSnippetModel** – includes the content from the snippets returned for 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.
- **QueryResultsModel (Q)** – includes the content from all the top 10 search results returned in response to each of the queries issued by the user. As in the previous model, this is meant to evaluate the relevance of the queries compared to the ones issued by gold standard users, except we use the whole pages associated with the links rather than the snippets.

- **ClickedResultsSnippetModel** – includes the content from the snippets returned for 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.
- **ClickedResultsModel (R)** – includes the content from all the results that were clicked by the user rather than just the snippets, which were used in the previous model. This is to evaluate the user's ability to choose a relevant result from the results page.
- **ClickedResults+NavModel (RN)** – 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.
- **Post-SearchWrite-upModel** – built from the search task write-up prepared by the users. This can be used to evaluate the final search effectiveness of the user for the given task.

The different language models for each user were compared using Kulback Liebler (KL) divergence [Kulback 1987] with corresponding concatenated language models build from the 3 gold standard users for that search task condition. 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.

Due to the relatively small size of the language models for the 2 snippets models and the write-up model, we used Rouge recall scores [Lin and Hovy, 2003] instead of the KL-Divergence scores. Rouge is the most common metric used evaluation of extractive summarization and sentence compression systems. It measures the overlap of n-grams between candidate summaries and the reference summaries. Rouge scores are particularly relevant for short texts such as sentence level comparison in sentence compression systems, however it should be noted that it is a very stringent metric. Since it is a n-gram overlap metric, and thus only gives credit for overlap when the same words are used, it can give low scores if the comparative texts are taken from different sources even when the meaning is the same.

Since the task defined in this user study is an elaborate information-seeking task, the usual system-oriented methods of IR evaluation are not applicable in our work, which requires a user-oriented perspective for evaluation [Kelly 2009]. So we make the simplifying assumption that the behavior of expert users can be used to obtain a "gold

standard" set of documents that can then be used as a reference point for evaluating the behavior of our target users. Although no two searches will yield the same set of documents, we can compute word distributions from the sets of documents found during each person's search session, and these can be compared using statistical language modeling techniques. Thus, we don't need to require that the same set of documents is found in order for behavior to be measured as good. Instead, we simply need to see that the distribution of words within the obtained collection, and thus the range of content included within those documents, is similar. We validated the reasonableness of this assumption using statistical tests that verify that divergences between the gold standard user models were significantly lower on average that divergences between the experiment users and the gold standard users.

In order to do this analysis, for each experiment user, and for each model, we made 3 comparisons, one for each pair of gold standard users. First, for each model and each pair of gold standard users, we computed the divergence between that pair model and that of a 3rd gold standard user. Normally this divergence was computed using KL-Divergence. However, as mentioned, for the snippet models and the final write up, we used Rouge scores instead because of sparsity. For comparison, we computed the divergence between that pair model and that of the experiment user. Thus, we were always making a controlled comparison of divergence within the set of gold standard users with that of the divergence of the experiment user and the same set of gold standard users. Because the comparison was totally controlled in this way, it was valid to test the significance of the difference using a 2-tailed paired-t test, and the differences were significant in all cases, indicating that the gold standard users were more similar to one another than to the students.

We also analyzed their query formulation strategies. We compared the queries (Q) issued with the English task statements (TS) for 10 search tasks. This comparison was done for each of the 30 task variations, to evaluate the impact of the presence of the English task statement on their query formulation strategy. Similar to the measures presented in [Metzler et al. 2007], we used the following lexical measures to evaluate the similarity between the query words and the task statement words:

- **Exact phrase match** – For each query, we identified the longest substring from the task statement that appears as a substring in the Query and then computed the ratio of its length over the total length of the query. We then averaged across queries. This was meant to be a proxy for the extent to which students borrowed from the task statement in their query formulation process.

- **Term overlay** – For each query, we counted the number of terms appearing anywhere in the task statement and then computed the ratio of its length over the total length of the query. We then averaged across queries. Here the intent was similar to that of Exact Phrase Match, except it was less restrictive.
- **Cosine Similarity** – We constructed a vector space consisting of all words appearing either in the task statement or any query issued by the user. We then computed a vector representation of the task statement and a separate vector representation of all of the queries concatenated together. We then computed the cosine distance between these two vectors. Again, this is meant as a similarity measure between the queries and the task statement. What this provides that the other two metrics do not is a measure of how much of the task statement was represented in the set of queries, rather than just a measure of how much of the query terms were borrowed from the task statement.

4.5 Results from Log file Analysis

In this section, we present the results of our experimental manipulation on the performance of the student users in comparison with gold standard users in terms of their Query, Click, and Navigation behavior as well as their write-up. The goal of this analysis is to explore which factors affect the success of our target student population. Primarily we are interested in whether the language manipulation shows significant effects, which would then suggest potential supports that might improve the search success of our target user population. Looking at the effects over these multiple models allows us to understand where the impact of the experimental manipulation is experienced in the search process. It allows us to determine whether the effect is primarily occurring at one specific stage, or whether the effect is compounded over time as the search process progresses from query to click to navigation, and finally to a write-up. Beyond this understanding, however, we are also interested in the extent to which these are general effects, and thus, we explore how the language manipulation potentially interacts with Topic or Specificity.

Tables III and IV present an overview of the results we discuss in this section. The letter indicates which equivalence class the cluster belongs to. For example, if a row contains cells with letters A and B, then cells with A are not statistically different from one another, and cells with B are not statistically different from one another, but cells with A are statistically different from cells with B. In the case of an interaction, pairwise

comparisons were only indicated between Language conditions within Specificity conditions.

Table III: KLD comparisons, note that Lower is Better

	High/ Both	High/ English	High/ Telugu	Low/ Both	Low/ English	Low/ Telugu
Query (Q)	.71(.28)^{A}	.73(.28)^{A}	.67(.3)^{A}	.61(.2)^{B}	.61(.2)^{B}	.61(.2)^{B}
Click (R)	.56(.17)^{AB}	.62(.19)^{A}	.59(.2)^{B}	.54(.2)^{AB}	.54(.2)^{A}	.48(.2)^{B}
Click+ Navigation (RN)	.59(.17)^{A}	.66(.2)^{B}	.64(.2)^{AB}	.60(.2)^{A}	.60(.2)^{B}	.53(.2)^{AB}

Table IV: Rouge scores, note that Higher is Better

	High/ Both	High/ English	High/ Telugu	Low/ Both	Low/ English	Low/ Telugu
Query Snippet	.06(.03)^{A}	.06(.03)^{A}	.06(.03)^{A}	.07(.03)^{B}	.07(.04)^{B}	.07(.04)^{B}
Click Snippet	.05(.04)^{A}	.05(.04)^{A}	.07(.05)^{B}	.06(.04)^{C}	.06(.05)^{D}	.05(.04)^{D}
WriteUp	.04(.04)^{A}	.04(.03)^{A}	.05(.04)^{B}	.06(.05)^{C}	.05(.05)^{CD}	.05(.04)^{D}

In the case of Query and Click behavior, we evaluate them both with KL-Divergence on language models computed from the full documents using the QueryResultsModel (Q) and ClickResultsModel (R) mentioned above as well as in terms of Rouge scores computed over the QueryResultsSnippet model and ClickResultsSnippet model listed above. Navigation results are only computed using KL-Divergence over the ClickedResults+Nav (RN) model mentioned above. In the case of navigation behavior, snippets are not relevant since they are not used beyond the main results page from each query. Since write-ups are short, we evaluate the quality using Rouge scores. Readers should note that it is the relative value of the scores and not the absolute value of the scores that is important. In particular, the Rouge scores presented in Table IV are very low, indicating very little overlap in behavior between experts and novices based on this metric. Similarly, the relatively high KLD scores presented in Table III indicate a large deviation on average between the novice and expert users. Thus, this assessment does not support the conclusion that the support was sufficient to close the gap between the performance of the novice and expert users. However, the differences in scores across conditions demonstrate where novice students performed better or worse in terms of comparison with expert behavior.

In order to get a sense of the search as a process, we would logically first investigate success at the query stage. We would then move on to clicks from the result page. Next we would explore success at navigation from clicked results. And finally we would look at the results. However, it only makes sense to deconstruct the full process to identify where issues have come up if we can verify that the experimental manipulation had some effect that persisted until the write up stage. Otherwise, where we may be able to find places where students were struggling and could be supported with a language oriented intervention, it would be the case that it is probably not a worthy investment if students are either able to compensate at a later stage for a deficit experienced at an earlier stage, or if deficits that occur over time, and are possibly compounded, eventually wash out the effect. Thus, we begin our analysis by first investigating the effect of the experimental manipulation on the write-up, which is a way of measuring ultimate task success. We do not report on the separate effect of the Topic factor since although it did frequently show a significant effect, it did not have a consistent effect across analyses, and it never had a significant interaction with the language factor or the specificity factor.

4.5.1 Results at the Write Up Stage

Using an ANOVA model with Language, Topic, and Specificity as independent factors, we found a significant effect of task language $F(2,1377) = 3.14$, $p < .05$. The condition where students had access to both English and Telugu was the best condition. It was significantly better than the English only condition (effect size .25 s.d.), and the Telugu only condition was in the middle, not being statistically different from either of the other conditions. There was no significant effect of task specificity on write-up success. However, there was a significant interaction between specificity and language $F(2,1377) = 10.42$, $p < .0001$, such that for High specificity tasks, Telugu > Both and English, (effect size Telugu-Both is .43 s.d., Telugu-English is .5 s.d.), and for Low specificity tasks, Both is significantly better than Telugu (effect size .25 s.d.), and English is not different from either. From this analysis we conclude that we have evidence that the language manipulation does have an effect on ultimate task success, but which configuration is best depends on the type of task. For high specificity tasks, representation of the task in Telugu only is best. For low specificity tasks, representation of the task in both English and Telugu is best. In both cases, having access to the Telugu is beneficial. The question is whether having access to English also is beneficial. In both cases, the appropriateness of the current situation where our target users are expected to conduct all of their schooling purely in English medium is called into question. In the

following sections, we explore the search process step-by-step to determine where the struggles are emerging for the students.

4.5.2 Results at the Query Stage

At the query stage, we do not see a significant effect of language manipulation either in the QueryResultsModel or the QueryResultsSnippetModel. In the QueryResultsModel, we see a statistical trend related to language $F(2,1305) = 2.0$, $p = .14$, such that the condition with Both English and Telugu is best, which is consistent with the overall effect of the language manipulation we observed on the write up. For both models, low specificity is best than high specificity, for example, with the QueryResults model $F(1,1305) = 55.22$, $p < .0001$, effect size .38 s.d.. However, there was no significant interaction with the language factor. So we see here weak evidence that the effect of the language manipulation occurs from the beginning of the search process. However, since the effect is not significant at this stage, we must conclude that either the effect compounds over the course of the search process, or additional problems emerge at later stages.

We expected that we would see the biggest difference in behavior at the query stage where users might be tempted to use the English task statement in a shallow way, simply borrowing terms from the task statement even if they did not understand them. We expected that in the case where users had access to both the English and the Telugu, they might use the terms from the English task statement more insightfully. We tested these hypotheses using Exact Phrase Match, Term Overlay, and Cosine Similarity mentioned above. For all three of these measures, there was both a significant effect of Language and a significant interaction between Language and Specificity. In all three cases, Telugu had the lowest similarity with the task statement. For Exact Phrase Match, English and Both were not statistically different, and Telugu was only different for High specificity tasks. For the other two, Telugu was always significantly more different than the other two, but English also showed significantly more similarity than Both. For Cosine Similarity, the difference between English and Both was only significant for Low specificity tasks. These results are consistent with our hypotheses about behavioral effects of the language manipulation, however we see with respect to effects on the Query performance that the changes in behavior did not always lead to consistent positive or negative effects on success.

4.5.3 Results at the Click Stage

At the click stage, what we evaluate is the user's ability to identify a subset of returned results that are likely to be relevant. Not all students have the same starting point for this, however, since students who issued poor queries will have fewer relevant links to choose from. On the other hand, they are not required to click on some minimum number of links. So if their judgment of what is relevant based on snippets is equally good, they can still do equally well by clicking on fewer results. We evaluate the goodness of clicks in two ways. The ClickResultsSnippet model is perhaps a better measure of how good the students' judgment is since it measures how close the snippets that are associated with links clicked on match snippets of links clicked on by experts. The other model measures how well the content in the links clicked by students matches that of the links clicked by experts.

At this level we see significant effects of the language manipulation in the ClickedResults model, $F(2,1041) = 4.0$, $p < .05$ such that Telugu is significantly better than English (Telugu-English effect size .2 s.d) and Both is not statistically distinguishable from either. In the ClickedResultsSnippet model there was no significant main effect of the language manipulation, but there was a significant interaction between Language and Specificity, $F(2,1041) = 11.8$, $p < .0001$, which was just a trend in the ClickedResults model ($p = .14$). For high specificity tasks, Telugu is significantly better than the other two, effect size .39 s.d. in the most extreme case. For low specificity tasks, Both is significantly better than the other two, effect size .27 s.d. in the most extreme case. This interaction effect is consistent with what we saw in connection with the write-up. The significant interaction supersedes the main effect, and thus it is not a concern that the main effect is not consistent with the main effect we saw on the write-up.

4.5.4 Results at the Navigation Stage

The results at the navigation stage demonstrate how good students were at identifying promising links on pages they clicked on from the results page. This is a different kind of task than selecting results, where there is a snippet to tell you what was relevant. Instead, the choice must be made based on the text that is on the link as well as the text that is around the link. Despite these differences, the results were consistent with the pattern we have been seeing. There was a significant effect of language $F(2,1021) = 3.9$, $p < .05$ such that English is significantly worse than Both (effect size .15 s.d.) and Telugu only (effect size .2 s.d.). There was a marginal interaction between language and specificity $F(2,1021) = 3.33$, $p < .06$. However, this time the result was different than before. For

Low specificity tasks, Telugu was significantly better than the other two, with effect size .3 in comparison with English and .25 in comparison with Both. For high specificity tasks, Both is better than English only (effect size .2) and Telugu is in the middle.

4.5.5 Discussion

Overall we see that the language manipulation had a significant effect on behavior at the query stage that was consistent with our hypotheses, however, the effects related to success were not apparent until later in the process. Because the language manipulation did show an effect on success at the write-up stage, we must conclude either that the effect of the change in behavior at that stage compounded over time, or that the effect of the manipulation was not limited to the change in behavior that resulted at the query stage. In other words, students may have continued to refer to the task statement differently within the different conditions at later stages as well. Ultimately, we see that the language in which the task statement is presented to students does matter for their success, however, which form is best depends on the task statement. Thus, it may not be best for students to always give them access to both English and their native language. Instead, it might be better to monitor their behavior and offer translation support when we detect that students are struggling. In the next section we explore a cluster based approach to detecting when students are struggling based on patterns in their behavior.

4.6 Results from Cluster Analysis

The results of the experimental manipulation give some evidence that some struggles with using search technology are language related and that they may be able to be partly addressed through some strategic use of machine translation technology. In that analysis, data mining technology was mainly used as a way of doing assessment of search activities for a large population of users. In this section we use unsupervised clustering for a different form of assessment that may be useful in supporting our target user population, or other populations where users may struggle with search. We then describe typical profiles of students who were classified in each cluster. It should be noted that while this analysis reveals what patterns of behavior were associated with more or less success, this analysis is not meant to make any causal claims about the connection between those behaviors and success. This analysis is only meant to illustrate what problematic behavior looks like within this user population. And the goal of classifying students into clusters associated with more or less success mainly fits into the overall vision for a total solution as a means for identifying the students who are in need of

support during search and language practice later. Our approach would be to provide support only as needed and to encourage as much autonomy as possible, especially as students increase their competence and need less help over time.

4.6.1 Computing Clusters

We computed our clusters based on features that can be extracted from user behavior during windows of time, regardless of whether we have accurate session boundaries or not. The features we used were as follows: QueryRate, Words per Query, Clicks per Query, Prob_R_Time, Prob_Q_Time, Prob_RN_Time, and Prob_QN_Time.

Table V: Definition of the Clustering Features

Feature	Definition
QueryRate	Average number of queries per minute
Words per Query	Average number of words per query
Clicks per Query	Average number of results clicked per query
Prob_R_Time	Percentage of total time in a search session spent browsing clicked results
Prob_Q_Time	Percentage of total time in a search session spent looking at query results page.
Prob_RN_Time	Percentage of total time in a search session spent browsing the pages navigated from the result pages
Prob_QN_Time	Percentage of total time in a search session spent looking at query results pages navigated from other query pages. For ex - Viewing subsequent results beyond the top 10 results; query reformulation without clicking on a result.

We used K-Means clustering. We set the number of clusters to 4. Cluster 1 had 170 instances. Cluster 2 had 439 instances. Cluster 3 had 390 instances. And Cluster 4 had 402 instances. We experimented with other numbers of clusters, but 4 seemed preferable because of the relatively even distribution of instances to clusters.

4.6.2 Analyzing Clusters

We found that clusters were associated with significant differences in write-up success even though the only features we looked at for creating the clusters were behavioral, as discussed above, $F(3,1383) = 11.25$, $p < .0001$. This result suggests that we can make predictions about how successfully users are moving towards a satisfying result by examining these behavioral features, which are easy to compute from their click stream data. Cluster 2 instances were significantly more successful than the others. Clusters 3

and 4 were significantly more successful than Cluster 1. The effect size of the difference between Cluster 1 and Cluster 2 was .5 s.d. The ratio of instances assigned to clusters did not vary significantly between conditions for Language, Specificity, or the interaction between the two. There was however a significant difference in proportion of cluster assignments by Topic. Nevertheless, when we controlled for Topic, we still found a significant effect of Cluster assignment on task success $F(3,1367) = 8.25$, $p < .0001$. In future research we will test the extent to which the assignment of instances to clusters generalizes to different tasks.

Table VI: This table displays the behavioral patterns associated with the 4 clusters.

	Cluster 1 **Queriers**	Cluster 2 **Readers**	Cluster 3 **Perusers**	Cluster 4 **Navigators**
QueryRate	.17(.14)^{D}	.08(.06)^{B}	.14(.09)^{C}	.07(.05)^{A}
Words per Query	4.2(3.6)^{A}	4.0(2.3)^{A}	4.0(2.3)^{A}	3.3(2.1)^{B}
Clicks per Query	.48(.61)^{C}	1.6(1.1)^{A}	.99(.77)^{B}	1.5(1.4)^{A}
Prob_R_Time	.09(.11)^{D}	.74(.11)^{A}	.43(.14)^{B}	.26(.14)^{C}
Prob_Q_Time	.85(.14)^{A}	.15(.09)^{C}	.33(.15)^{B}	.12(.09)^{D}
Prob_RN_Time	.03(.08)^{D}	.11(.11)^{C}	.19(.14)^{B}	.61(.15)^{A}
Prob_QN_Time	.02(.05)^{A}	0.0(.01)^{B}	.01(.05)^{A}	0.0(0.0)^{B}

Table VI displays the behavioral patterns associated with each cluster. Each cell shows the average feature value, with the standard deviation in parentheses. As with Tables III and IV, the letter indicates which equivalence class the cluster belongs to. For example, if a row contains cells with letters A and B, then cells with A are not statistically different from one another, and cells with B are not statistically different from one another, but cells with A are statistically different from cells with B. For example, we see that for QueryRate, instances assigned to Cluster 1 have a significantly higher value than those in all other cells. Instances assigned to Cluster 3 have significantly bigger values than those in Clusters 2 and 4. And those in Cluster 2 have significantly bigger values than those in Cluster 4.

Based on the patterns we see, we have assigned names to the clusters. For example, Cluster 1 has a distinctively high QueryRate as well as a distinctively high proportion of time spent looking at query results pages, thus we have named them Queriers. As a result, instances within this cluster are associated with a distinctively low amount of time on results pages or even clicks to results pages. These users appear to be having trouble

getting promising looking results back based on their queries. It is not surprising then that instances within Cluster 1 were the least successful in their associated write-up.

At the opposite end of the spectrum, the most successful cluster was Cluster 2. We have named users whose instances ended up in Cluster 2 Readers because they spent a distinctively high proportion of their time on the results pages from queries. They had the second lowest QueryRate, but among the highest clicks per query. Thus, they appear to have been very successful with their queries in terms of brining up promising looking links.

Users in Clusters 3 and 4 were in the middle in terms of success, being significantly worse than users in Cluster 2 but significantly better than users in Cluster 1. We named users in Cluster 3 Perusers. They had a moderately high QueryRate as well as Clicks per Query. Thus, they appear to have been more successful in their query behavior than users in Cluster 1, but less so than users in Cluster 2. They spent the majority of their time roughly evenly split between examining the set of links returned from their queries and on the pages linked to those lists. The distribution of time spent on the different types of actions was flatter with users in this cluster than in the other clusters. The pattern might be indicative of not having a clear strategy. This is consistent with other work showing less systematic behavior in search when users are not familiar with a domain and are thus more likely to get confused [Marchionini 1989; Zhang et al. 2005]. Users in Cluster 4 spent much more time on pages linked to the results pages. These links that they navigated to from results pages were frequently advertisements. These users did not appear to be able to distinguish relevant pages from irrelevant pages. Based on their QueryRate and Clicks per Query, they look a lot like users in the most successful cluster, however, the major distinguishing factor is this high tendency to navigate away from results pages rather than spending time reading the results pages. Thus, we have named these users Navigators.

4.6.3 Replicating Clusters

The cluster analysis would not be very useful if it were not possible to reliably assign a cluster to a pattern of behavior without knowing personal information about the user or that user's need. Thus, we used a Decision Tree learning algorithm in the Weka toolkit [Witten & Frank 2005] to test through a cross-validation experiment where on each fold we trained a model on 90% of the data and then tested on the remaining 10%. We did this across all 10 folds and then averaged performance. We found that the features that were used to create the clusters in the first place were also informative enough so that

instances could reliably be assigned to clusters based on those features. We achieved a 96.9% accuracy (.96 Kappa), which indicates excellent reliability. A pruned version of this model is displayed in Figure 2.

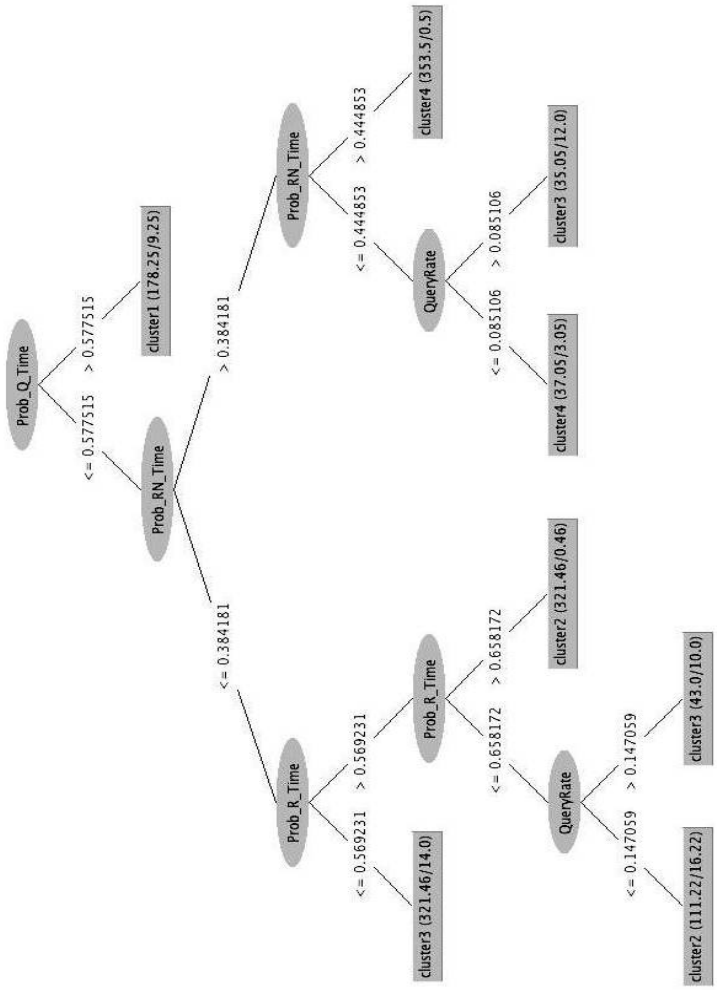

Fig. 2 A Simplified version of the trained decision tree for classifying students within clusters. The first feature that the model splits on is Prob_Q_Time (shown at the left), indicating the importance of the percentage of time spent on Queries. The left child, Prob_RN_Time, of the root splits into two subtrees Prob_R_Time and Prob_RN_Time respectively.

Thus we see that we can achieve a high level of accuracy at replicating the clustering of the data based only on features of the user's behavior. With this in mind, it would be possible for a search engine to monitor behavior of users and assign them to clusters based on the behavior that is identified, without knowing what the task is. We consider

these results somewhat preliminary since we have tested them only on a very particular user population, and it is not clear whether the values of the cluster feature that would be predicted for other types of users would mean the same thing in terms of whether they are struggling or not. However, these promising initial results give an indication that this is a potentially fruitful direction for continued research.

5. CONCLUSION AND CURRENT DIRECTIONS

In this article we have presented an exploration of search behavior, search success, and recommendations for support that are based on a data-driven methodology with data mining as a key component. We have argued that when it comes to information seeking, low English literacy and low computer literacy populations have very different needs from the great majority of Internet users today, who are more capable of using current search technology to meet their information seeking needs.

For such users, their low comprehension of the language may act as a hindrance in formulating effective queries. If relevant information is provided in response to their query, they may or may not recognize it as such. Irrelevant information that is also returned, which more literate users would quickly identify and ignore, confuses such users, and may lead them in the wrong direction.

Results of our investigation suggest that our target user population could be supported through an integration of data mining and machine translation technology. We have demonstrated how unsupervised clustering is effective in identifying subsets of users whose pattern of behavior is indicative of varying levels of success across topics and levels of abstraction in problem formulation. Our experimental study demonstrates that for certain types of tasks our target users would benefit from having their information need accessible both in English and their native language. Ultimately, similar technology might help us understand further what aspects of language encountered during search pose problems for our target users. These issues will be addressed in our ongoing work.

The results we present in this article are promising and support the idea that data mining technology has much to offer the investigation of search for emerging Internet users.

The work as it stands has some notable limitations. First, before the results can really achieve educational impact, the other components of the total solution would need to be developed. The results presented here are meant to be a proof of concept of sorts, demonstrating the potential value of such a solution in order to justify the development effort. We believe that potential has indeed been demonstrated. Another limitation is

that the dataset used in the research consisted mainly of data from one very specific user population. Data from more advanced users was only used as a comparison for purposes of evaluating the success of our experimental manipulation. A key component of the total vision is that users who are struggling would be identified by means of their search behavior. Our cluster analysis demonstrates that we can reliably identify what cluster a user belongs in, assuming that user is from the same distribution that our dataset represents. However, what we did not evaluate is the ability of the trained model to distinguish users within that distribution from users who are outside of that distribution. If a user comes to the search task with substantially better developed skills, then the features used for classification within that cluster model, such as amount of time spent per page or proportion of time spent on queries rather than browsing would likely take on a different significance, and thus the prediction of the model for those users would likely be incorrect. As we pointed out, the results of the cluster analysis are mainly suggestive, and a larger scale effort based on data from multiple user populations is needed. Also, while the data set covers a variety of search tasks, we have only begun to scratch the surface of possible types of search tasks. We must acknowledge our inability to make claims beyond the range of search tasks that we have experimented with in this large scale study. Thus, a very important next step will be to address these generality issues both with respect to search tasks and user populations.

ACKNOWLEDGMENTS

This research was supported in part by NSF SBE 0836012 and NSF HCC 0803482.

We gratefully acknowledge all the help of our colleagues Raj Reddy, Praveen Garimella, Srimanth Kadati and all of the mentors and students at Rajiv Gandhi University for Knowledge Technologies, without whose collaboration this work would not have been possible.

APPENDIX

Here we detail the list of actual texts of the search tasks presented to the students for Study 2. Note that for each task statement below, we have indicated the topic and level of specificity in bold square brackets.

1. **[Food, Low]** Imagine that you are food critic at a major Newspaper. You have to write an article highlighting Differences between local (AP) Cuisine with other cuisines in India.

2. **[Food, High]** Imagine that you are food critic at a major Newspaper. You have to write an article highlighting Differences between local (AP) Cuisine with the Karnataka Cuisine.
3. **[Eye diseases, Low]** Imagine that you are a Engineering student at a university. You have to write a report listing 2-3 common eye diseases/ailments and their treatments.
4. **[Eye diseases, High]** Imagine that you are a Engineering student at a university. You have to write a report listing 2-3 common eye diseases in old people and their treatment.
5. **[Cricket, Low]** Imagine that you are a Newspaper editor. You have to write an article discussing the impact of increasing popularity of IPL Cricket on the future of International Cricket, ODI and Test Cricket. You can take into consideration the financial interest of the players and the broadcasters.
6. **[Cricket, High]** Imagine that you are a Newspaper editor. You have to write an article discussing the impact of increasing popularity of T20 Cricket on the future of International Cricket, ODI and Test Cricket. You can take into consideration the financial interest of the players and the broadcasters.
7. **[Fluid dynamics, Low]** Imagine that you are a Engineering student at a university. You have to write a report describing important concepts in Fluid Dynamics in Physics. Also mention a few applications of Fluid Dynamics.
8. **[Fluid dynamics, High]** Imagine that you are a Engineering student at a university. You have to write a report describing Bernoulli's Principle in Physics. Also mention a few applications of the principle.
9. **[Movies, Low]** Imagine that you are a Newspaper editor. You have to write an article discussing the negative impact of movies on Young Children.
10. **[Movies, High]** Imagine that you are a Newspaper editor. You have to write an article discussing the negative impact of smoking in Movies on Young Children.

REFERENCES

AGICHTEIN, E., BRILL, E., and DUMAIS, S.T. 2006. Improving web search ranking incorporating user behavior. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 19-26.

AULA, A. 2003. Query Formulation in Web Information Search. In *Proceedings of IADIS International Conference WWW/Internet, 2003*, P. ISAIAS and N. KARMAKAR, (Eds.) IADIS Press, 403-410.

AULA, A. and KELLAR, M. 2009. Multilingual search strategies. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, 3854-3870.

BACHMANN, M., GOBERT, J., and BECK, J. 2010. Tracking Students' Paths through Student Transition Analysis, In *Proceedings of the 3rd International Conference of Educational Data Mining*, 269-270

BALFE, E. and SMYTH, B. 2005. An Analysis of Query Similarity in Collaborative Web Search. In *Proceedings of the 27th European Conference on Information Retrieval*, 330-344.

BHAVNANI, S.K. 2005. Strategy Hubs: Domain portals to help find comprehensive information. *Journal of the American Society for Information Science and Technology 57*(1), 4-24.

BIRRU, M., MONACO, V., CHALRLES, L., DREW, H., NJIE, V., BIERRIA, T., DETLEFSEN, E., and STEINMAN, R. 2004. Internet Usage by Low-Literacy Adults Seeking Health Information: An Observational Analysis, *Journal of*

Medical Internet Research 6(3).

BRANDT, S. and UDEN, L. 2003. Insight into mental models of novice Internet searchers , *Communications of the ACM 46*(7), 133-13.

CHEN, S.F. and GOODMAN, J. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 310-318.

DOWNEY, D., DUMAIS, S., LIEBLING, D., and HORVITZ, E. 2008. Understanding the relationship between searchers' queries and information goals. In *Proceedings of 17th ACM Conference on Information and Knowledge Management*, 449-458.

DUGGAN, G.B. and PAYNE, S.J. 2008. Knowledge in the head and on the web: Using topic expertise to aid search. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 39-48.

GAO, J., ZHOU, M., NIE, J., HE, H., and CHEN, W. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations, *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 183-190.

GRASSIAN, E. and KAPLOWITZ, J. 2001. *Information Literacy Instruction: Theory and Practice*, New York: Neal-Schuman Publishers.

GRIMES, C., TANG, D., and RUSSEL, D.M. 2007. Query Logs Alone are not Enough. In *Workshop on Query Log Analysis at WWW 2007: International World Wide Web Conference*, Banff, Alberta, Canada.

GUINEE, K., EAGLETON, M.B., and HALL, T.E. 2003. Adolescents' Internet search strategies: Drawing upon familiar cognitive paradigms when accessing electronic information sources. *Journal of Educational Computing Research 29*, 363 – 374.

HATCH, E.M. 1983. *Psycholinguistics: A second language perspective*. Newbury House Publishers, Inc., Rowley, MA.

HENRY, L. A. 2005. Information search strategies on the Internet: A critical component of new literacies. *Webology 2*(1), Article 9.

HOLSCHER, C. and STRUBE, G. 2000. Web search behavior of Internet experts and newbies. *Computer Networks: The International Journal of Computer and Telecommunications Networking 33*, 337-346.

HOWARD, L., JOHNSON, J., and NEITZEL, C. 2010. Examining Learner Control in a Structured Inquiry Cycle Using Process Mining. In *Proceedings of the 3rd International Conference on Educational Data Mining*, 71-80.

IVONEN, 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 49*, 312-326.

INGWERSEN, P. and JARVELIN, K. 2005. *The turn: Integration of information seeking and retrieval in context.* Dordrecht, The Netherlands: Springer.

JENKINS, C., CORRITORE, C.L., and WIEDENBECK. S. 2003. Patterns of Information Seeking on the Web: A Qualitative Study of Domain Expertise and Web Expertise. *Information Technology and Society 1*(3), 64-89.

JEONG, H., BISWAS, G., JOHNSON, J., and HOWARD, L. 2010. Analysis of Productive Learning Behaviors in a Structured Inquiry Cycle Using Hidden Markov Models. In *Proceedings of the 3rd International Conference on Educational Data Mining*. 81-90

KELLAR, M., HAWKEY, K., INKPEN, K.M., and WATTERS, C. 2008. Challenges of Capturing Natural Web-Based User Behaviors. *International Journal of Human-Computer Interaction 24*, 385 – 409.

KELLY, D., DUMAIS, S., and PEDERSEN, J. 2009. Evaluation challenges and directions for information seeking support systems. *IEEE Computer 42*, 60-66.

NEELY, T. 2006. *Information Literacy Assessment: Standards-Based Tools and Assignments*, Chicago: American Library Association.

KELLY, D. and COOL, C. 2002. The effects of topic familiarity on information search behavior. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, 74-75.

KIM, K. and ALLEN, B. 2002. Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology 53*(2), 109-119.

KRAAIJ, W., NIE, J.Y., and SIMARD, M. 2003. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics 29*, 381–419.

KULLBACK, S. 1987. The Kulback-Leibler distance. *The American Statistician 41*, 340-341.

KULES, B. 2008. Speaking the same language about exploratory information seeking. In *Information Seeking Support Systems Workshop*, Chapel Hill, NC.

LAVRENKO, V., CHOQUETTE, M., and CROFT, W.B. 2002. Cross-lingual relevance models. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 175–182.

LIMBERG, L. 1999. Experiencing information seeking and learning: A study of the interaction between the two phenomena. *Information Research 5*(1). 68.

LIN, C. Y. and HOVY, E. H. 2003. Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 71-78.

MARCHIONI, G. 1989. Information seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science 40*(1), 54-66.

MARCHIONI, G. 1995. *Information seeking in electronic environments*. Cambridge, UK: Cambridge University Press.

MARCHIONI, M. 2006. Exploratory Search: From Finding to Understanding. *Communications of the ACM 49*, 41-46.

METZLER, D., DUMAIS, S., and MEEK, C. 2007. Similarity measures for short segments of text. In *Proceedings of the 29th European Conference on Information Retrieval*, 16-27.

MONTALVO, O., BAKER, R. S., SAO PEDRO, A., and GOBERT, J. 2010. Identifying Students' Inquiry Planning Using Machine Learning. In *Proceedings of the 3rd International Conference on Educational Data Mining.* 141-150.

NEELY, T. 2006. *Information Literacy Assessment: Standards-Based Tools and Assignments*. Chicago: American Library Association.

NETZ-TIPP.DE 2002. Distribution of languages on the Internet. http://www.netz-tipp.de/languages.html.

PONTE, J.M. and CROFT, W.B. 1998. A Language Modeling Approach to Information Retrieval. *Research and Development in Information Retrieval*, 275–281.

QIU, L. 1993. Analytical searching vs. browsing in hypertext information retrieval systems. *Canadian Journal of Information and Library Science 18*(4), 1-13.

RICE, R., MCCREADIE, M., and CHANG, S. 2001. *Accessing and Browsing Information and Communication.* Cambridge, MA: The MIT Press.

SUTCLIFFE A. and ENNIS, M. 1998. Towards a cognitive theory of information retrieval. *Interacting with Computers 10*, 321-351.

TEEVAN, J., ALVARADO, C., ACKERMANN, M.S., and KARGER, D.R. 2004. The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search. In *Proceedings of the ACM Conference on Human Factors in*

Computing Systems, 415-422.

VASILYEVA, E., PECHENIZKY, M., TESANOVIC, A., KNUTOV, E., VERWER, S., and DE BRA, P. 2010. Towards EDM Framework for Personalization of Information Services in RPM Systems. In *Proceedings of the 3rd International Conference on Educational Data Mining*, 331-332.

WHITE, R.W., DUMAIS, S.T., and TEEVAN, J. 2009. Characterizing the influence of Domain Expertise on Web Serach Behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 132-141.

WITTEN, I. H. and FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, second edition. San Francisco, CA: Elsevier.

WOOLF, B. P., SHUTE, V. J., VANLEHN, K., BURLESON, W., KING, J., SUTHERS, D., BREDEWEG, B., LUCKIN, R., BAKER, R.S.J.d., and TONKIN, E. 2010. *A roadmap for Education Technology*. Monograph prepared for the Computing Community Consortium, Washington, DC. Retrieved from http://www.cra.org/ccc/docs/groe/GROE%20Roadmap%20for%20Education%20Technology%20Final%20Report.pdf

ZHAI, C. and LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 334-342.

ZHANG, X., ANGHELESCU, H.G.B., and YUAN, X. 2005. Domain Knowledge, search behavior and search effectiveness of engineering and science students. *Information Research 10*, 217.