

# A Task-oriented Approach to Search Engine Usability Studies

Isak Taksa

Baruch College, Zicklin School of Business, Department of Computer Information Systems, New York, USA  
Email: isak\_taksa@baruch.cuny.edu

Amanda Spink

Queensland University of Technology, Faculty of Information Technology, Brisbane, Australia  
Email: ah.spink@qut.edu.au

Robert Goldberg

Queens College, Department of Computer Science, Flushing, USA  
Email: rrg@solomon.cs.qc.edu

**Abstract**— Usability is a multi-dimensional characteristic of a computer system. This paper focuses on usability as a measurement of interaction between the user and the system. The research employs a task-oriented approach to evaluate the usability of a meta search engine. This engine encourages and accepts queries of unlimited size expressed in natural language. A variety of conventional metrics developed by academic and industrial research, including ISO standards, are applied to the information retrieval process consisting of sequential tasks. Tasks range from formulating (long) queries to interpreting and retaining search results. Results of the evaluation and analysis of the operation log indicate that obtaining advanced search engine results can be accomplished simultaneously with enhancing the usability of the interactive process. In conclusion, we discuss implications for interactive information retrieval system design and directions for future usability research.

**Index Terms**—usability, search engine, search tasks, query formulation, query refinement

## I. INTRODUCTION

The ISO standard defines usability as the extent that a user can utilize a product effectively, efficiently and with satisfaction in achieving specific goals. Further, the standard defines effectiveness as the accuracy and completeness with which users achieve specified goals; efficiency is measured by how much resources are expended in achieving the goals; and satisfaction is the freedom from discomfort, and positive attitudes towards the user of the product. While the standards conceptualize these three criteria with general product goals in mind, usability of a search engine could be defined in more specific terms. Effectiveness is the extent that the search engine returns documents relevant to the user's query.

Based on "Evaluating Usability of a Long Query Meta Search Engine", by Isak Taksa and Amanda Spink, which appeared in the Proceedings of the 40th Annual International Conference on Systems Sciences (HICSS, 2007), Hawaii, USA, January 2007, © 2007 IEEE

Efficiency would be how many refinement searches the user would be required to employ in obtaining an accurate information need. To fulfill satisfaction, the search engine should allow the user to express the search query in a natural language.

Even if the users know how to express their information needs (IN) in natural language they frequently face a "language" barrier while trying to convert this knowledge into a few exact terms to formulate an adequate search query. Numerous studies confirm users' persistence in using short queries [21, 41]. Another drawback of shorter, unfocused queries is the number of results returned by the search engine. Overwhelmed by the task of sifting through a massive volume of returned search results, users frequently examine just the first page (top 10 results) [32, 39] and continue with a new, reformulated, yet still short query, hoping to find more relevant results. This unstructured approach leads to time waste that is caused by the frequent finding/re-finding and subsequent accepting/discarding of the same documents. To further complicate the problem, very few users are familiar with or use Boolean operators [19] or phrases [46] to improve quality of search results. Additionally, commercial search engines were designed to find information, while leaving users on their own when they try to store and organize the search results [2] or discover new knowledge [25, 49]. In contrast, the long query meta search engine (LQMSE), the focus of this research, incorporates a range of novel functions and addresses many of the above concerns. To perform the actual Web search the LQMSE uses Google.

The remainder of this paper is organized as follows. In Section II we present an overview of the current state of usability research, its challenges and accomplishments. The discussion in Section III continues with a detail description of the long query meta search engine, its tasks and semi-automated tools to assist the user. Objectives of this research are stated in Section IV. The research design and results are described in details in Section V. Section

VI discusses implication of this research for search engine research and development. Conclusions and directions for further research are outlined in Section VII.

## II. RELATED USABILITY STUDIES AND DESIGN

### A. *Historical roots*

History provides an interesting perspective and an extra dimension for understanding the purpose of usability studies. Usability, even as currently understood, has its roots in the military minds of the 1970s. (They seemed to have coined the term usability.) [43] The U. S. military was seeking a solution to the problem of training members of the armed forces how to use complex electronic hardware based on reading the corresponding technical manuals and whether the manuals favored a particular gender or background. This was especially critical at that junction of time when the draft was winding down and the military was beginning an increased and open-access recruiting effort of civilians. [43]

An example of the change of approach in training was the operating manual of the M1 tank was limited to four ideas per page and was heavily reliant on a picture-based training. The result was that the manual grew over thirty-fold from one-hundred pages to over three-thousand. [43] It was on this basis that usability was adopted by the software industry. (Indeed, the popularity of this approach and the accessibility of its information, facilitated creation of the "M1 tank platoon" game by MicroProse Software, Inc. in 1989 [22]). It was not till the 1990s that usability became part of the HCI (Human-Computer Interface) movements. In this historical context, Nielsen's book "Usability Engineering" can be thought as the transitional point from traditional to contemporary views [31]. An interesting comparative study between HCI and usability can be found in Dillon [9].

### B. *Current views*

Researchers in the field of usability have developed a variety of diverse views on what usability is and how to study and measure it. Indeed, A recent paper on usability research [15], which examined usability measures from 180 published studies, concluded that choosing an appropriate and acceptable usability measure is a complex and difficult process. Some, when designing new systems, strictly follow ISO definitions of usability [18] and user-centered design [16], while others apply a user-centered approach to evaluate usability of existing search engines [25]. Some researchers branch out in search of new methodologies and metrics in measuring usability [37], while others try to keep abreast of new and evolving technologies [29].

Pace [21] suggests an alternative approach to usability testing. Instead of concentrating on the efficiency of getting the results, concentrate on the state of consciousness of the individual users and their sense of control and time during the experience. Whereas most studies use all data available to obtain relevance

judgments, Greisdorf and Spink [14] explore the usage of the median point of a distribution.

Many researchers and practitioners agree that context of use should be the driving force in the quest for accurate usability testing and measurements [8, 31]. Gabbard et al. in 2002 [12] gives a clear picture of the nexus between engineering disciplines that affect or deal directly with usability requirements. Development of an interactive system generally involves two major components: behavioral (external user-oriented) and constructional (internal software development). Usability engineering enables user interaction, both in terms of the way the interface looks as well as the behaviors that are provided. However, this engineering is predicated on two other engineering sub-disciplines: software and system engineering. Both of the latter comprise the constructional component of system design. Spink et al. [38] explores a user-centered approach to the evaluation of the Web search engine. They analyzed pre- and post-search questionnaires and search transaction logs of the Inquiries system.

### C. *Important factors*

An important source of information would be user-feedback. Due to the cost and time involved in evaluating this feedback, most major studies do not incorporate it. An interesting alternative is provided by Sharma and Jansen [36] who track subsequent user activities on the system such as saving to disk, bookmarking for further reference and printing (termed "implicit feedback"). Due to the explicit anonymity agreement of the participants of most studies, this information is not available.

Important elements in the users' satisfaction with the information retrieval process are ease of use, continuity of efforts and retention of earlier results. Some researchers built a flow theory to measure user's experience and satisfaction during information seeking activities [33]. Others measure effectiveness of the search engine that provides automated assistance during the search process [20] and detects and eliminates duplicate efforts (submitting same query, examining same search result, etc.) [48].

An important factor of engine usability is its ability to "forgive" errors and allow the user to go back and repeat as many tasks as necessary [44]. Since information retrieval tasks could be performed over a period of time, an engine's ability to retain and reuse earlier results is high on users' satisfaction list [42, 48], especially while conducting multi topic searches. Because search engines regularly return millions of hits, a search engine that filters search results for user relevance evaluation [23, 37] provides welcome help and improves user satisfaction.

### D. *Usability of a search engine*

The focus of this paper is the usability of a search engine and, therefore, we concentrated on studies that measure usability attributes applicable to various tasks of the information retrieval process. Several studies looked at efficiency and user satisfaction with the query formulation process [35, 48]. Others investigated the effectiveness of this process from a query reformulation

point of view. Query reformulation becomes more efficient if the user can see the search terms, which were used in previous search sessions [48]. Providing a larger search field to allow full query view makes the system more useful [34].

The experiment presented here parallels, in spirit, the design protocol used by Elspeth Golden, Bonnie E. John, Len Bass in 2005 [13]. As reported there, an experiment was designed to study the usefulness of the usability pattern USAP (Usability-Supporting Architectural Pattern) in modifying the design of software architectures to support a specific usability concern (there, a cancellation command missing from design).

As is established in the field of testing, a between-subjects design was used in the experiment. [10] This design is typical for a usability study, where data are collected on different measures such as completion time (and in the current research, how many queries are necessary to obtain the material). While that particular study can be critiqued, it does lay the basis for the development of further protocols.

Clearly, task completion time is a dependent variable since it depends on a number of factors including the complexity of each task and the experience level that each student has with the tools available. In order to concentrate directly on the intrinsic usability parameters, the levels of the participants were considered uniform and hence, not studied. Golden, John and Bass [13] divided the participants into three groups, each given an increasing amount of information regarding solving the task at hand.

Since the task in the present study significantly differs from that of the above, the exact manner in which the information was provided cannot be emulated. However, the overall design of dividing the students into three groups and each with different degrees of accessibility to the information at hand is the same and was inspired by their study protocol. Likewise, the between-subjects design was appropriate where the values of the dependent variable for one group of participants (for example, the group of participants who have most access to the information by querying the web from multiple platforms and multiple queries) are compared with the values for another group of participants (for example, the group of participants who only have time for one query on one search engine). Finally, in between-subjects designs, the data provided by each participant appears from one group only. See Spink et al. [28] who studied the degree of multitasking search and information task switching during multiple-query sessions.

#### *E. Conducting a usability study*

A very interesting reference on how to conduct a usability study is provided by Kevin Cheng in 2005 [7]. While the intent of that article was to instruct beginners ("moderators") on practical points about conducting the study, this document should be considered a "must" checklist for any level of study considered. Not all parts considered will be relevant to every study, but the following points of advice to participants are common to all studies are summarized here: (a) the participants must

realize that it is the study of the system and not "of them"; (b) while questions are encouraged, participants must realize that not every question can be answered; (c) the extent that the participant may use the system and the manner in which they can, must be clear to each group of participants; (d) to avoid stress, let participants be aware of any breaks and do not stress time limits, whether or not an official one exists. These points were incorporated into this research study design as well. Another useful set of eight guidelines is provided by WebCredible in 2006 [47].

A website that is dedicated to interface design and usability testing is [www.user.com](http://www.user.com). This site is run by Hal Shubin, a consultant in the field. This site collects together short essays on the key steps an academic researcher, usability engineer or industrial testing specialist should follow in order to conduct a test study. The phases of product design, usability testing as a part of user-interface design and obtaining expert product reviews are some of the essential components of a general study design that are discussed on that website. [16]

George Casaday in 1997 [6] is the first to identify that usability design would benefit from pattern architectures, an idea that was emerging around that time. This paper suggests that the creation of usable interactive systems be based on established software engineering patterns with traditional usability attributes. He provides examples of three pattern types: simple, intrinsic and circumstantial. Simple design can be expressed with only one attribute, intrinsic patterns require a combination of attributes and circumstantial patterns involve external constraints. Mahemoff and Johnston in 1998 [26, 27] extend that approach to include further patterns. Welie.com currently provides one-hundred and thirty patterns dealing with user needs, application needs and context of design and acts as an advanced portal to other libraries of pattern design.

The trichotomic approach of Casaday [6] formed the basis of the three groups of queries that were presented to the participant groups of this study. User situations that require information can be simple, requiring only one submission to the search engine, paralleling the simple pattern. Other queries are intrinsically complex (intrinsic pattern) and require multiple queries to clarify or to narrow the result set. Finally, there are circumstantial scenarios that span a number of disciplines or focuses where more than one specialized search engine would help the user obtain the preferred information.

#### *F. A recent study*

At this point, it would be appropriate to cite Bertolucci's PC World who recently conducted a comprehensive study ("shoot-out") on public search engines. [4] Multiple rounds of queries were submitted and each search engine was judged based on accuracy of response, simplicity of interface and how easy it is to use. What is particularly interested to this study is the equal attention given to image/video searches in addition to the standard queries for text.

The stated purpose of that study was whether Google is the best search engine. There is no dispute that Google

is the most popular search engine. Nielsen NetRatings in 2007 [30] document that Google has 53.7% of all submitted queries with a whopping 3.9 billion queries handled in the month of January 2007. In addition, it had a 40.6% YOY (year-on-year) growth. However, the article gives the impression that all major search engines today have reasonably simple interfaces and a high-level of accuracy of results. But, as immediately critiqued by Bradley [5], some important factors were missing from this study: level of experience of the user and single versus meta-search tools. Furthermore, while one search engine may be overall the best, but an individual search engine can be expert within a certain discipline or application.

### III. LONG QUERY META SEARCH ENGINE

Before we proceed with a description of our study and experiments we will briefly discuss conventional information retrieval and how the long query meta search engine changes the established conventions. Information retrieval using a commercial search engine (Google in our discussions and experiments) is a single search event that consists of several tasks. The user starts with information needs, usually expressed at length in natural language. Based on prior experience and domain knowledge the user selects a few search terms to formulate the search query (see Table I, task 1, Commercial SE) and then submits it to the search engine (task 4).

TABLE I.  
SEARCH AND RETRIEVAL TASKS: COMMERCIAL VS. LQMSE  
(M- USER INITIATED MANUAL TASK, A-AUTOMATED ENGINE TASK)

M / A	Commercial Search Engine Event	TASK #	Long Query Meta Search Engine Process	M / A
M	Query formulation	1	Query formulation	A
		2	Query reformulation; Phrase creation	M
		3	Determining filtering criteria	M
M	Submission to SE	4	Specifying control parameters for multiple query formulations; submission to SE	M
M	Search results examinations	5	Search results examinations and ranking	M
M	Storage and management of search results	6	Storage and management of search results	A
		7	Knowledge discovery	A

The search term selection process is inexact and search results could easily become skewed. The user then examines search results returned by the engine (task 5) and then proceeds to either bookmark the relevant site,

store its copy on a hard drive, or just discard it (task 6). If search results are not relevant (or if the search engine does not return enough relevant results) the user starts a new search event with some mental recollection of search terms used and sites retrieved in prior search events. Users rarely keep notes of queries and associated search results, so it is a frequent occurrence that similar or identical queries are submitted several times, and the user retrieves, examines, and stores the same results repeatedly. Furthermore, all tasks in this process are manual.

In contrast, the LQMSE redefines all of the above tasks, while also introducing new ones. The search event becomes a search process that automates many functions, thus leaving control of the process in user's hands. It also allows repetition of tasks with the same or different parameters. Instead of converting information needs (IN) into only a few search query terms, the user enters the entire IN description into the search window of the search engine (Fig. 1).

Figure 1. Long Query Search Field

The engine parses the IN and creates an ordered list of all words (terms) in the IN (excluding "stop" words). The order of words depends on two frequencies: the frequency of the word in the Google collection (number of documents that contain that word) and the frequency of the word in the original IN.

Query Refine		
Words in Your Queue		
Priority	Word	Remove
1	<a href="#">analysis</a>	<input type="checkbox"/>
2	<a href="#">design</a>	<input type="checkbox"/>
3	<a href="#">systems</a>	<input checked="" type="checkbox"/>
4	<a href="#">computerized</a>	<input type="checkbox"/>
5	<a href="#">systematically</a>	<input type="checkbox"/>
6	<a href="#">transforming</a>	<input type="checkbox"/>
7	<a href="#">functioning</a>	<input type="checkbox"/>
8	<a href="#">feeding</a>	<input type="checkbox"/>
9	<a href="#">accomplished</a>	<input type="checkbox"/>

Figure 2. Long query refinement

The ordered list is the initial query formulated by the engine (see Table I above, task 1, LQMSE). The user is presented with this ordered list for editing, i.e. deleting/correcting misspelled words, changing the order of the words or inserting new words. This query reformulation task (task 2) can be repeated any time throughout the search process (Fig. 2).

If the user is aware of phrases that are common to the domain, those could be added to the list of search terms submitted to the search engine (Fig. 3).

In task 3 (see Table I above), the user can specify the sites that will be filtered-in/out. This excludes the processing and evaluation of unwanted sites while ensuring the inclusion of potentially relevant sites.

Figure 3. Phrase builder

For example, the user only wants to include (filter-in) sites that offer instructional materials related to Systems Analysis & Design (SA&D) (words like *tutorial*, *notes*, *lectures* appearing in text summary or in the URL) and only from US institutions of higher learning (the *.edu/* in its URL) but wants to exclude (filter-out) commercial sites (*.com/* appearing in the URL) that offer the same material (Fig. 4).

Figure 4. Include/exclude filter settings

Figure 6. Final results (partial view)

Once all subqueries are submitted, all search results are filtered according to filter-in/out criteria and combined into a final search result list, which is ordered depending

The heart of this engine is its ability to create multiple subquery formulations. The user is presented with the option to decide on a number of subqueries (generated from terms in the ordered list, which was created in task 1 and edited in task 2 and submitted to the search engine). In task 4 the user specifies two control parameters  $n$  – number of words (top  $n$  words from the ordered list in task 1) that are used to generate conjunctive subqueries and  $r$  – minimum number of words in each subquery. This task could be visualized as spreading a fish net where  $n$  represents the size of the net while  $r$  specifies its mesh. The bigger the  $n$ , the wider the net used to “catch” (retrieve) potentially relevant sites.  $r$  on the other hand defines the depth of search (see Figure 5 below). Smaller  $r$  will provide a shallow search, meaning more results are considered for inclusion, larger  $r$  means deeper search where less results are found.

Figure 5. Subquery formulation parameters

The process of subquery formulation is quite straightforward: create various  $nCr$  combinations, or simply, create **all** possible search queries consisting of at least  $r$  terms from the list of  $n$  terms. The number of such combinations is calculated as follows

$$\sum_{i=r}^n nCi = \sum_{i=r}^n \frac{n!}{(n-i)! * i!} \quad (1)$$

For example, for  $n = 8$  and  $r = 5$  the total number of subqueries is 93. Recent research [45] demonstrates that the number of subqueries could be reduced without significant degradation of results.

While examining the site, the user may decide to store or bookmark (Fig. 6 above) the site (by assigning a relevance rank) or just discard it (by leaving the Rank box blank).

The engine stores the ranked results (task 6) and updates two lists of URLs: relevant (accepted) results and non-relevant (rejected) results. URLs in both lists are used to prevent earlier results (relevant and non-relevant) from appearing in subsequent search results, thereby saving the time and effort of processing and examining results more than once.

The final task (task 7) analyzes URLs in both lists for frequent appearance of common words or abbreviations in the domain name and/or path/file name thereby suggesting potential criteria for further filter-in/out choices. Figure 7 is an example of a file of accepted or relevant sites (called hicss.yes and conversely the file for rejected sites is called hicss.no).

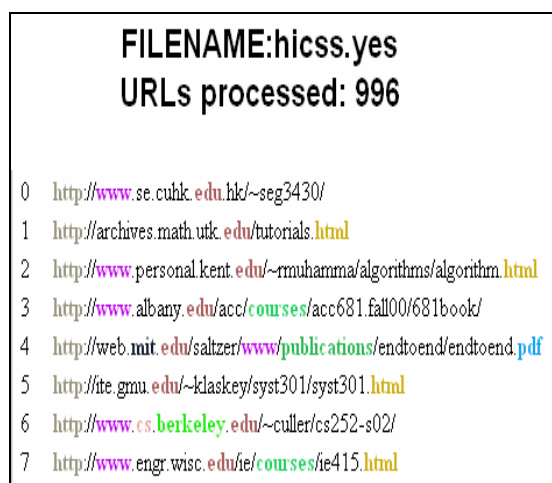




Figure 7. List of relevant results (all URLs)

The color-coding and the color legend (see Figure 8) allow for easy visualization of commonalities in a list of URLs.

954	S	F	edu
512	S	F	www
317	S	F	html
221	S	F	cs
199	S	F	mit

Figure 8. Color legend

For example, observing that there are 199 sites from **mit**, the user may want to investigate other common terms in URL addresses of relevant **mit** sites. By clicking button  (filter results) the user can observe further URL subdivision for all **mit** sites (Fig. 9) in original relevance order with a color legend similar to one in Figure 8. The button  provides a list of **mit** sites in alphabetical order.

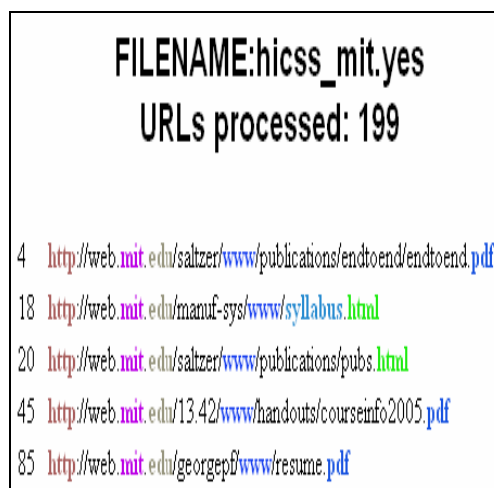


Figure 9. List of relevant results (selected URLs)

#### IV. RESEARCH OBJECTIVES

The classic information retrieval model created a representation of the document and translated an information need into a query and tried to match the query to the representation. Already early on, researchers cited by Spink and Jansen [39] challenged this assumption that the user would be required to represent the information need in phraseology understandable by the search engine. They asked why technology cannot provide an interface to the user so that the information need could be expressed in a direct ordinary manner instead of controlled query format imposed by the system. In terms of usability, traditional models were not efficient because forcing the user to submit queries in specific formats requires extra resources to learn the method and because of this discomfort, the user was not completely satisfied.

Bates [3] suggested an alternative usage model ("berrypicking") for understanding how the user obtains the information need in the absence of a natural language interface. Bates argues that the user would be interested in an evolving search, first obtaining a general piece of information on the topic and then refining the search in an iterative fashion, obtaining more and more pieces of information on their topic till the user is satisfied with the result obtained. Here too, the usability criteria of efficiency and satisfaction are not fulfilled because of the amount of time required by the user to obtain relevant results.

This research paper suggests a multiple query formulation (MQF) that allows the user to input their query in a natural language format to its fullest extent. Instead of the user having to wait throughout the inherent sequential process of Bates, MQF does all of the multiple searches prior to the return of the final result and that result benefits from integration of all of the intermediary results retrieved by the multiple intermediary searches. In a certain sense, the MQF simulates the "berrypicking" by considering various subsets of keyterms provided by the user's wordy natural language query. The users are free of



uneasiness (hence, satisfied) because the query is expressed in their own words and efficient because the search engine does all of the work subsequent to the user's single query submission.

Usability measures are traditionally classified into three major categories: effectiveness, efficiency, and satisfaction (we follow classification suggested in recent, comprehensive research [15]). Our objectives are to apply some of the conventional measures in each category to usability analysis of the long query search engine. Specifically, in the effectiveness category, *completeness* (user's ability to complete assigned tasks) and *quality of outcome* (relevance of retrieved results); in the efficiency category, *time to complete* (time required to complete the assignment) and *usage patterns* (user's participation in individual retrieval tasks); and in the satisfaction category, *control* (ability to control the outcome) and *learnability and retention* (stress-free learning to learn and hard to forget).

## V. RESEARCH DESIGN

This section provides the details about the experimentation studied in this research paper. The background for this study and design is presented in section two above. The current discussion is divided into three subsections: participants (subsection A), experiments (subsection B) and data analysis (subsection C).

### A. Participants

The study was conducted on the campus of an international university during the summer session of 2005. Three sections of a multi-section advanced Systems Analysis & Design (SA&D) course were randomly selected. Successful completion of this course required some computer competency. Though research proves that user-competency affects usability results [11, 28], for the purpose of our experiments we assumed that all students performed at the same level. The number of students in each section (group) was approximately the same (30, 32 and 35 students).

### B. Experiments

A three-part experiment was designed to collect data for usability analysis. In the first week (of a 6-week course) the students received the following two-paragraph description of the concept of SA&D.

*"The examination of a problem and the creation of its solution. Systems analysis is effective when all sides of the problem are reviewed. Systems design is most effective when more than one solution can be proposed. The plans for the care and feeding of a new system are as important as the problems they solve" [1].*

*"Systems analysis and design, as performed by systems analysts, seeks to analyze data input or data flow*

*systematically, processing or transforming data, data storage, and information output within the context of a particular business. Furthermore, systems analysis and design is used to analyze, design, and implement improvements in the functioning of businesses that can be accomplished through the use of computerized information systems"*[24, pp. 6-7].

The first part of the experiment - **Conventional benchmark IR**, was to use the above description to create search queries and to retrieve results relevant to each group's assignment. Each group received an individual assignment to find 20 relevant sites with specific instructions on how to search the Web. Additionally, each participant was asked to rank the relevancy of all selected results using the ordinal relevancy scale (from 1 for non-relevant to 9 for very relevant). We did not use the typical "relevant/non-relevant" discrete scale so as to ensure that students actually open and read each site. To gauge the students' approval of their own search results we asked another student to conduct an impartial relevancy evaluation. The following are the groups' assignments and execution instructions:

**Group A:** Use as many commercial search engines (SE) and formulate as many queries (Q) as needed to collect a list of sites (URLs) that are similar to the two-paragraph description above (excluding textbooks sellers);

**Group B:** Use one commercial SE and formulate as many queries as needed to collect a list of sites (URLs) that address the academic issues of SA&D (e.g. scientific papers) discussed in the two-paragraph description above;

**Group C:** Use one commercial SE and formulate one successful query that will provide 20 distinct educational sites that offer instructional SA&D information (tutorial, lecture notes, etc.).

The second part – **long query meta search engine IR**, which took place in the second week, was to repeat the original assignment while now using LQMSE. Even though students were not familiar with the engine, only **Group C** received an hour of introductory instruction, explaining all seven tasks (as specified in Table I above). The other two groups were shown the screen shots for each task for identification purposes only. During the actual experiments, students from **Group B** were allowed to have one-on-one consultation without giving any benefit to other students in the group. No students in any group were allowed to consult with each other.

It is important to mention that this search engine does not have any kind of on-line help, except for Task 6 which instructs the user on how to rank the relevancy of sites the user wants to save for future use. After the experiment, students in all three groups submitted written requests for consultation specifying the task number in question. Observers collected and tabulated these requests and then repeated the introductory tutorial (given initially to Group C) followed by a Question and Answer session.

The third part – **motivated long query meta search engine IR**, which took place in the sixth week, was to repeat the second part of the experiment under new conditions. Students did not see, discuss, or use the new engine. Before the experiments, the students were advised to select only highly relevant search results. Again, no consultation of any kind was allowed during the experiments.

### C. Data analysis

Results of the first experiment, presented in Table II, show that students are familiar with commercial search engines, are willing to experiment with search queries and spend time to obtain relevant results

TABLE II. CONVENTIONAL IR RESULTS

Group	# of SE (Min/Max)	# of Queries (Min/Max)	# of Terms (Min/Max)	Time spent (minutes) (Min/Max)	% of shared results
A	3/5	5/11	3/7	65/90	11
B	1	2/14	2/6	45/90	45
C	1	1	4/9	50/85	67

To format the data series in the subsequent charts (Fig. 10 through Fig. 13) we used standard Excel's patterns with the following fill effects: *light vertical* to depict results for Group A (multi-query/multi-platform), *narrow horizontal* for Group B (multi-query/single-platform), and *large checker board* for Group C (single-query/single-platform).

We identified two attributes in the effectiveness category: completeness and quality of outcome. Similar to any commercial search engine, when the user presses the "Search" (or "Go") button to begin the search, it is guaranteed that the search will be completed. Even though LQMSE consists of many tasks, the built-in defaults allowed users in all groups to successfully complete the search.

While it was an easy assignment for **Group A** (find similar results), students in **Group B** (find scientific papers) had to figure out a way to tweak the engine to retrieve more specific results. They used observers' consultations extensively, but only relied on the query reformulations option, and shied away from the filter option. On the other hand, **Group C** (find instructions material), having benefited from the introductory lecture, experimented with both options.

TABLE III. RELATIVE RELEVANCY RANKING (SELF VS. IMPARTIAL)

	Part 1	Part 2
Group A	1.75	0.6
Group B	1.6	1.2
Group C	0.9	0.8

Fig. 10 demonstrates the usage of engine defaults to achieve 100% completion of the search assignment. It closely correlates with the students' requests for

consultation after completing the second part of the experiment (Fig. 11).

To measure quality of outcome, we compared the relevancy results from the first and second parts of the experiments. Table III above reflects the fact that impartial evaluation always produces a lower relevancy rank than self-evaluation, and that the discrepancy is consistent among groups and experiments.

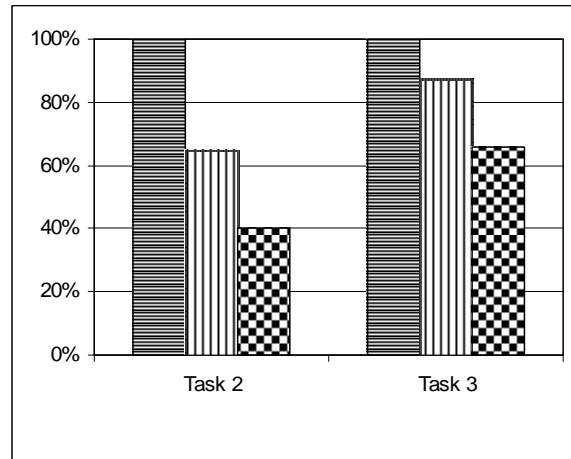


Figure 10. Using the default options to complete the search assignment

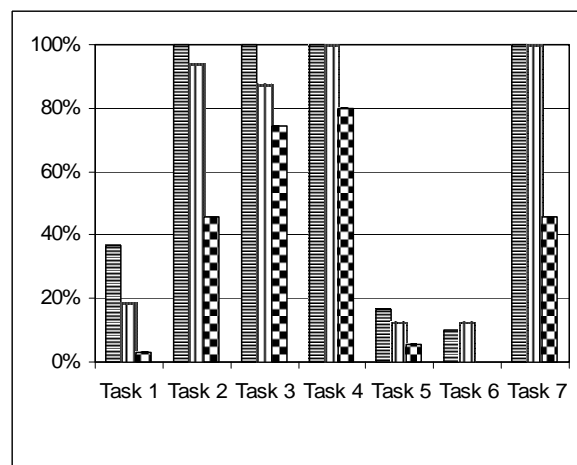


Figure 11. Requests for consultations

In the efficiency category we measured time to complete and usage patterns. Time to complete (we show minimum reported time), as depicted in Fig. 12, depends on the desired quality of the search results.

**Group A** with the easiest assignment completed Part 2 much faster than Part 1. However, when quality requirements were amplified in Part 3, the time to complete increased as well.

On the other hand, **Group B** and **Group C** reported longer time to complete Part 3 vs. time reported in Part 1. This could be explained by the many iterations of Tasks 2, 3 and 4 performed by the user in order to improve the quality of the final search results.

It is important to emphasize, that all tasks (except for Task 5) could be performed autonomously using engine defaults.



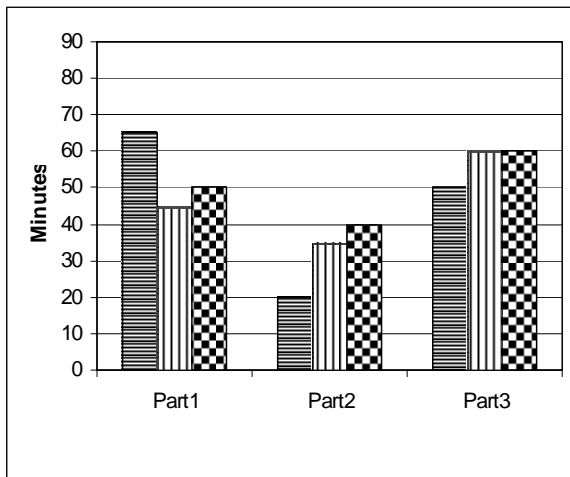


Figure 12. Time to complete

Fig. 13 demonstrates the average number of iterations performed by a student in each group. It is interesting to note that students in Group C, who received the most instructional time, were the most frequent users of Task 6, which requires more understanding and potentially improves the final results quicker. On the other hand, Task 4 where the affects on the final results are not obvious was the least used.

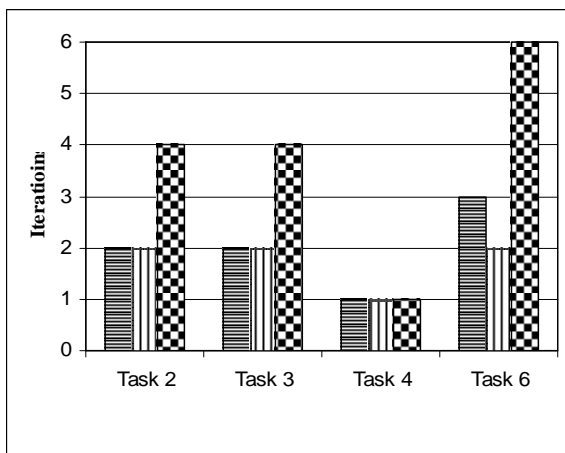


Figure 13. Average iterations per task

In the final satisfaction category we identified two attributes: control and learnability. The first, control, reflects the users' perception of how their action can influence the final search results. Table IV shows a dramatic increase in the users' utilization of tasks that control the quality of final search results (when motivated by a higher grade in the course).

TABLE IV. UTILIZATION OF USER CONTROL (PART 2 VS. PART 3)

Group	Task 2	Task 3	Task 4	Task 6
A	0/58	0/61	0/32	0/86
B	23/68	14/66	4/29	0/64
C	28/133	18/146	3/40	0/221

The users' ability to retrieve quality results after minimal exposure to a new engine and after a gap of four weeks is a good measure of learnability and retention (the second attribute of satisfaction). When asked to submit a request for consultation again, after Part 3 was completed, the students' response was largely in the vicinity of 3% for Tasks 4 and 6, and close to 0% for all other tasks.

## VI. IMPLICATIONS FOR SEARCH ENGINE DEVELOPMENT

Our usability results allow designers of search engines a look at users' behavior, needs and strategy throughout various tasks of the information retrieval process. For example, the *completeness* measure confirmed a need for a well constructed system of defaults to allow even a novice user to complete a search process. On the other hand, the *learnability* measure demonstrated the user's capacity (and motivation) to try new, unexplained and experimental functionality.

While this is true for many new technological innovations, the *usage-pattern* measure proved again that the users' participation in an optional task depends on their clear understanding of the costs and benefits associated with the task. The *control* measure confirmed that users armed with this understanding will attempt to manipulate search results and manage the flow of the process.

Our usability results suggest that search engine designers should introduce more functionality that assist the user during the lengthy, comprehensive, and often imprecise information retrieval process. Developers, through on-line or context-sensitive help, should make clear the costs and benefits of using, misusing or not using available functionality. Finally, users should be empowered to control search, retrieval and management of search results throughout the information retrieval process.

## VII. CONCLUSION AND FURTHER RESEARCH

After examining numerous usability measures, we applied selected measures to evaluate the usability of a long query meta search engine. Our initial results demonstrate that it is possible to select a generalized set of usability measures to evaluate a specialized search engine. Furthermore, we captured results that are significant to the design and development of new search engines.

Our research is still in progress and we plan to expand it in several directions. Additional usability measures will be explored. The experiments will be designed to separately examine objective and subjective usability categories.

In order to obtain more generalized results the meta-search engine will use additional underlying search engines. Finally, the research will benefit if users are divided into groups according to search experience and domain of interests.

## ACKNOWLEDGMENT

This work was partially supported by grants from The City University of New York PSC-CUNY Research Award to Baruch and Queens Colleges.

## REFERENCES

- [1] Answers.com, "Systems design," accessed January 11, 2007, <http://www.answers.com/topic/systems-design>
- [2] A. Aula, N. Jhaveri and M. Kiki, "Information search and re-access strategies of experienced web users," In *Proceedings of the 14th international conference on World Wide Web*, Japan, pp. 583–592, 2005.
- [3] M. Bates, "The Design of Browsing and Berrypicking Technique for the Online Search Interface," *Online Review*, Vol. 13(5), pp. 407–431, 1989
- [4] J. Bertolucci, "Search Engine Shoot-Out," *PC World*, (April 25, 2007) accessed May 17, 2007, <http://www.pcworld.com/article/id,130979/article.html>
- [5] P. Bradley. (2007). Return To The Search Engine Shoot-Out, accessed May 30, 2007, <http://searchengineland.com/070503-040333.php>.
- [6] G. Casaday, "Notes on a pattern language for interactive usability," *Proceedings of the 1997 Conference on Human Factors in Computing Systems (CHI '97)*, Atlanta, Georgia, pp. 289–290, 1997.
- [7] K. Cheng, "Beginner's Guide to Moderating a Usability Study," accessed June 23, 2007 <http://www.ok-cancel.com/archives/article/2005/06/beginners-guide-to-moderating-a-usability-study.html>
- [8] C. Cool and A. Spink, "Issues of Context in Information Retrieval (IR): an Introduction to Special Issue," *Information Processing and Management*, Vol. 38(5), pp. 605–611, 2002.
- [9] A. Dillon, "Beyond usability: process, outcome and affect in human computer interactions," *Canadian Journal of Library and Information Science*, Vol. 26(4), pp. 57–69, 2002.
- [10] Ergosoft Laboratories, "What is a between-subjects design?" 2001–2003, accessed June 15, 2007 [http://www.ergolabs.com/between\\_subjects\\_design.htm](http://www.ergolabs.com/between_subjects_design.htm)
- [11] L. Faulkner and D. Wick, "Cross-user analysis: Benefits of skill level comparison in usability testing," *Interacting with Computers*, Vol. 17(6), pp. 773–786, 2005.
- [12] J. L. Gabbard, J. E. Swan II, D. Hixa, M. Lanzagortac, M. Livingstonb, D. Brown, S. Julier, "Usability Engineering: Domain Analysis Activities for Augmented Reality Systems," *The Engineering Reality of Virtual Reality 2002*, A. Woods, J. Merritt, S. Benton, M. Bolas, Editors, *Proceedings SPIE Volume 4660, Stereoscopic Displays and Virtual Reality Systems IX*, pp. 445–457, 2002.
- [13] E. Golden, B. E. John, L. Bass, "The value of a usability-supporting architectural pattern in software architecture design: a controlled experiment," *Proceedings ICSE*, pp. 460–469, 2005
- [14] H. Greisdorf and A. Spink, "Median Measure: An Approach to IR System Evaluation," *Information Processing and Management*, Vol. 37(6), pp. 843–857, 2001.
- [15] K. Hornbæk, "Current practice in measuring usability: Challenges to usability studies and research," *International Journal of Human-Computer Studies*, Vol. 64(2), pp. 79–102, 2006.
- [16] Interaction Design, Inc., "Web design, UI design & Usability Testing," accessed July 18, 2007, <http://www.user.com/index.htm>
- [17] ISO 13407:1999, "Human-centered design processes for interactive systems," accessed January 14, 2007, <http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=21197>
- [18] ISO 9241-11:1998, "Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability," accessed from <http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16883&ICS1=13&ICS2=180&ICS3> on January 17, 2007.
- [19] B. J. Jansen and C. M. Eastman, "The Effects of Search Engines and Query Operators on Top Ranked Results," In *Proceedings of International Conference on Information Technology: Computers and Communications*, Las Vegas, 2003, pp. 135–139.
- [20] B. J. Jansen and M. D. McNeese, "Evaluating the Effectiveness of and Patterns of Interactions With Automated Searching Assistance," *Journal Of The American Society For Information Science And Technology*, Vol. 56(14), pp. 1480–1503, 2005.
- [21] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing and Management*, Vol. 42(1), pp. 248–263, 2006.
- [22] B. Jones, "Review of 'M1 Tank Platoon'," (October 2, 2006) accessed July 30, 2007, <http://www.mobgames.com/game/dos/m1-tank-platoon/reviews/reviewerId,80356/>
- [23] M. Kiki and A. Aula, "Findex: improving search result use through automatic filtering categories," *Interacting with Computers*, Vol. 17(2), pp. 187–206, March 2005.
- [24] K. Kendall and J. Kendal, *Systems Analysis and Design*. Pearson-Prentice-Hall Incorporated: Upper Saddle River, New Jersey, 2005.
- [25] C-H. Li and C-C. Kit, "Web Structure Mining for Usability Analysis" In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 309–312, 2005.
- [26] M. J. Mahemoff and L. J. Johnston, "Pattern Languages for Usability: An Investigation of Alternative Approaches," in Tanaka, J., Editor, *Proceedings of the 1998 Asia-Pacific Conference on Human Computer Interaction (APCHI '98)*, Shonan Village, Japan, pp. 25–31, 1998.
- [27] M. J. Mahemoff and L. J. Johnston, "Principles for a Usability-Oriented Pattern Language," in *Proceedings of the Australian Computer Human Interaction Conference (OZCHI '98)*, Adelaide, Australia, pp. 132–139, 1998.
- [28] G. Meiselwitz and G. Trajkovski, "Effects of Computer Competency on Usability and Learning Experience in Online Learning Environments," in *Proceedings of Seventh ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, pp. 339–342, 2006.
- [29] A. Monk, "Noddy's guide to usability," *Interfaces*, Vol. 50, pp. 31–33, 2002.
- [30] Nielsen, "Nielsen/NetRatings Announces January U.S. Search Share Rankings," accessed May 21, 2007, [http://www.nielsen-netratings.com/pr/pr\\_070228.pdf](http://www.nielsen-netratings.com/pr/pr_070228.pdf)
- [31] J. Nielsen, *Usability engineering*. Academic Press: San Diego, California, 1993.
- [32] S. Ozmutlu, A. Spink and H. C. Ozmutlu, "A Day in the Life of Web Searching: an Exploratory Study," *Information Processing and Management*, Vol. 40(2), pp. 319–345, 2004.
- [33] S. Pace, "A grounded theory of the flow experiences of Web users", *International Journal of Human-Computer Studies*, Vol. 60(3), pp.327–363, March 2004.

- [34] D. E. Rose, "Reconciling Information-Seeking Behavior with Search User Interfaces for the Web," *Journal of the American Society for Information Science and Technology*, Vol. 57(6), pp. 797–799, 2006.
- [35] B. Schmitt and S. Oberländer, "Evaluating and Enhancing Meta-Search Performance in Digital Libraries," In *Proceedings of the 3rd International Conference on Web Information Systems Engineering (WISE.02)*, pp.93–104, 2000.
- [36] H. Sharma and B. J. Jansen, "Automated evaluation of search engine performance via implicit user feedback," In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 649–650, 2005.
- [37] M. Shepherd, C. Watters and A. Marath, "Adaptive User Modeling for Filtering Electronic News," In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, Hawaii, 2002, Volume 4, pp. 102b.
- [38] A. Spink, "A User Centered Approach to evaluating Human Interaction with Web Search Engines: an Exploratory Study," *Information Processing and Management*, Vol. 38(3), pp. 401–426, May 2002.
- [39] A. Spink and B. J. Jansen, *Web Search: Public Searching of the Web*. Kluwer Academic Publishers: New York, NY, 2004.
- [40] A. Spink, M. Park, B. J. Jansen and J. Pedersen, "Multitasking During Web Search Sessions," *Information Processing and Management*, Vol. 42(1), pp. 264–275, 2006.
- [41] A. Spink, D. Wolfram, B. J. Jansen and T. Saracevic, "The public and their queries," *Journal of the American Society for Information Science and Technology*, Vol. 52(3), pp. 226–234, 2001.
- [42] T. Sumner and M. Dawe, "Looking at Digital Library Usability from a Reuse Perspective," In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 416–425, 2001.
- [43] J. Sumser (Editor), "Usability," Electronic Recruiting News, an interbiznet.com production, (November 24, 1999), accessed July 29, 2007 <http://www.interbiznet.com/ern/archives/991128.html>
- [44] A. Sutcliffe, "Assessing the Reliability of Heuristic Evaluation for Website Attractiveness and Usability," In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, Vol. 5, pp. 137, 2002.
- [45] I. Taksa, "Predicting the Cumulative Effect of Multiple Query Formulations," In *Proceedings of the IEEE International Conference on Information Technology: Coding and Computing, ITCC'05*, Vol. II, pp. 491–496, April 2005.
- [46] H. Topi and W. Lucas, "Searching the Web: Operator Assistance Required," *Information Processing and Management*, Vol. 41(2), pp. 383–403, March 2005.
- [47] WebCredible, "8 guidelines for usability testing," (April, 2006) accessed July 27, 2007 <http://www.webcredible.co.uk/user-friendly-resources/web-usability/usability-testing.shtml>.
- [48] B. M. Wildemuth, "Evidence-based practice in search interface design," *Journal of the American Society for Information Science and Technology*, Vol. 57(6), pp. 825–828, 2006.
- [49] H. Wu, M. Gordon, K. DeMaagd and W. Fau, "Mining web navigations for intelligence," *Decision Support Systems*, Vol. 41(3), pp. 574–591, March 2006.

**Isak Taksa** is an Associate Professor in the Department of Computer Information Systems at Baruch College of the City University of New York (CUNY). His primary research interests include information retrieval, knowledge discovery and text and data mining. He has published extensively on theoretical and applied aspects of Information Retrieval and Search Engine technology in journals including *Information Retrieval* and *Journal of the American Society for Information Science*.

**Amanda Spink** is Professor in the Faculty of Information Technology at the Queensland University of Technology and Co-Leader of the Information Science Cluster. Her primary research includes: basic, applied, industry and interdisciplinary studies in information science, information behavior, cognitive information retrieval; Web retrieval, including relevance, feedback and multitasking models. Professor Amanda Spink has published over 300 journal articles, refereed conference papers and book chapters, and 5 books. She is a member of the numerous journal editorial boards including: *Information Processing and Management*, *Journal of Documentation*, *Journal of Information Systems Education* and *Webology*.

**Robert Goldberg** is a tenured Professor of Computer Science at Queens College of the City University of New York. He holds his doctorate in Computer Science from the Courant Institute of New York University, conducted as an ONR graduate fellow there. He currently serves on the editorial board of the *International Journal of Intelligent Hybrid Systems* and on the review board of the *ACM Computing Reviews* journal. He co-edited a series of Special Issues on Developmental Mathematics for the *Mathematics and Computer Education* journal, co-authored a book on *Multiobjective Optimization*, published by Springer-Verlag, New York, and has written a number of articles on information technologies.