Transparent Queries: Investigating Users' Mental Models of Search Engines

Jack Muramatsu
Wanda Pratt
Information & Computer Science
University of California, Irvine
jmuramat@ics.uci.edu, pratt@ics.uci.edu

ABSTRACT
Typically, commercial Web search engines provide very little feedback to the user concerning how a particular query is processed and interpreted. Specifically, they apply key query transformations without the user's knowledge. Although these transformations have a pronounced effect on query results, users have very few resources for recognizing their existence and understanding their practical importance. We conducted a user study to gain a better understanding of users' knowledge of and reactions to the operation of several query transformations that web search engines automatically employ. Additionally, we developed and evaluated Transparent Queries, a software system designed to provide users with lightweight feedback about opaque query transformations. The results of the study suggest that users do indeed have difficulties understanding the operation of query transformations without additional assistance. Finally, although transparency is helpful and valuable, interfaces that allow direct control of query transformations might ultimately be more helpful for end-users.

Keywords
Interfaces, User Studies.

1. INTRODUCTION
Search engines often interpret and transform a user's query during the retrieval process. These processes profoundly affect both the search results and the users' ability to understand the relationship between their query and the returned results. Yet, few users are aware of those transformations or of the influence of those transformations on the sometimes mysterious results returned. Moreover, each search engine chooses its own combination of transformations, and might modify their choice without any indication of that change to the user. The application of these types of query transformations coupled with the common lack of feedback, leads users to develop inaccurate mental models of search engines.

Mental models contain varying degrees of detail and accuracy; yet, they guide users' operation, expectations, and understanding of tools [1]. When users' mental models differ from the actual underlying system model, people often become frustrated and fail to accomplish their tasks. By transforming the user's query without providing any feedback on those modifications, web search engines interfere with users' formation of accurate mental models and thus contribute to the users' inability to find the desired information.

We have developed a system called **Transparent Queries** (**TQ**) that uses lightweight visualization mechanisms to address this problem and provide the missing feedback. TQ makes the normally opaque transformations and processing of a user's query more transparent by illustrating the transformations in a visually annotated presentation of the query.

This work focuses on the usability problems that arise when search engines process and modify users' queries without their knowledge. We use the term **query transformations** to refer to these automatically applied processes**,** although they are heterogeneous in their nature and operation. Because most commercial search engines do not specify exactly how they transform users' queries or how they use those transformations to rank documents, we relied on a well-known summary of commercial web search engine features and underlying operations [2]. We verified this information through a series of manual test queries. Based on the summary of features, we chose four query transformations as candidates both for the user study and the TQ feedback mechanisms: application of a default logical operator, stop word removal, term suffix expansion, and term order sensitivity. These particular query transformations are not a conclusive list, but they serve as a starting point for exploring the related usability issues.

In this study, our goal was both to understand the mental models that users form of web search engines and to determine whether TQ encourages users to develop mental models that accurately match the underlying system model.

2. RELATED WORK
Most studies of searchers have focused on understanding or modeling users' search strategies and behaviors [3]. Choo et. al.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.

[4] used surveys, interviews, and client-side logging to characterize and model knowledge workers' information seeking behaviors on the Web. Navarro-Prieto et. al. [5] compared cognitive strategies of high and low experience Web searchers. The researchers found that, unlike experts, novices do not create search plans and strategies; rather they rely heavily on the results from the search engines. Borgman [6, 7] explicitly studied the mental models of users of Boolean retrieval systems. She noted that many systems implicitly apply a default connective operator (AND or OR), and she concluded that this implicit choice can be problematic and confusing for the naïve user. These previous studies point to the need both for further studies of searchers' mental models and for interfaces that provide feedback to help searchers form accurate models.

Others researchers have also promoted the use of feedback in search interfaces. Both in his general user interface guidelines [8] and his recommendations for search interfaces [9, 10], Shneiderman explicitly recommends that user interfaces offer informative feedback. Anick et. al. [11] report that visible query transformations in their Boolean search system simplified the task of query reformulation and made users more confident in the results of their search. Koenemann and Belkin [12] found that the degree of visibility and interactivity for a relevance feedback system positively influenced retrieval performance. We have also adopted Koenemann and Belkin's useful classification of three levels of interactivity: opaque, transparent, and penetrable. The **opaque** level represents an interface that does not provide any indication of the underlying transformations that the system performs. Opaque is the standard level of feedback commonly provided by modern commercial web search engines. In contrast, **transparent** interfaces provide visible feedback on the automatically applied transformation. Finally, **penetrable** interfaces provide both feedback as well as a means for users to control or adjust the transformations. We used this naming scheme in the creation of our system, Transparent Queries, and we refer to these levels of interactivity throughout the paper.

In contrast, Beaulieu and Jones [13] discuss potential pragmatic limits of increased feedback and control over query operations. Specifically, they raise the possibility that increased feedback and user control might increase users' cognitive loads. Thus, their work is a reminder that providing more feedback and more control will not necessarily improve a user's ability to harness a particular information retrieval system. Careful design and research are necessary to realize the potential benefits of such approaches.

3. TRANSPARENT QUERIES IMPLEMENTATION

Transparent Queries is implemented as a Java2 applet. The TQ applet is embedded in a three frame web interface (see Figure 1). One frame is allocated to the TQ applet, one contains navigational links to various Web search engines, and the last frame displays the currently selected search engine. A small amount of Javascript glue is used to connect Transparent Queries to the search engine frame. Specifically, a Javascript routine periodically checks to see if the URL of the search engine frame has changed. If the URL has changed, it is passed to TQ by way of a public method. TQ compares the received URL with those associated with supported search engines. If a match is found, the query terms are extracted and TQ proceeds to construct and display appropriate lightweight feedback for the user's query. All feedback provided by TQ consists of a concise textual message describing the operation of a given query transformation and a listing of the query terms entered by the user. Unless otherwise stated, TQ presents the feedback in blue text, which helps to set it off from the query terms that are displayed in black text. Several of the feedback mechanisms also make use of slight visual annotations of the query terms themselves. The following sections will discuss each of the four supported query transformations as well as provide specific details on the TQ feedback used for each transformation.

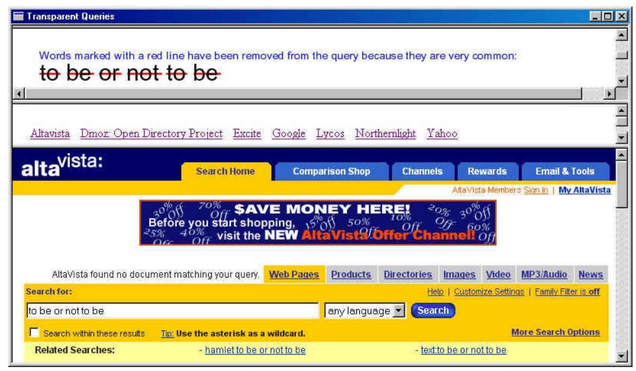

Figure 1: Transparent Queries interface for the user study. TQ is in the top frame, navigational controls are in the middle frame, and the bottom frame displays the currently selected search engine. This screenshot illustrates TQ feedback for the Stop Word Removal task and the associated query: *to be or not to be.*

3.1 Automatic Boolean Operator

Although it is unlikely that any commercial web search engine uses an underlying Boolean model, many search engines do employ Boolean logic at the query level. Specifically, many search engines choose a default Boolean operator if the user does not specifically provide one. In many cases, there is no obvious indication that such a choice exists (if it does) nor is there an indication that such a choice has been made without the user's input. Many search engines typically provide a simple web form for entering a list of query terms, but once a user submits a query, search engines decide whether all query terms must be present in any retrieved and ranked document (i.e., what we refer to as a Boolean AND) or whether a document is retrieved and ranked even if it contains only a subset of the query terms (i.e., Boolean OR). There are some exceptions to this situation. A few search engines include a way for users' to control the Boolean operator used as part of their primary search interface, and many more provide this option in a separate *advanced* search interface.

The choice of Boolean operator can dramatically influence the number and type of results returned, and in extreme cases, can produce either lots of results or no results at all. Transparent Queries attempts to provide feedback that will help users understand the resultant effect of a particular search engine's default Boolean operator. For OR search engines, Transparent Queries uses the following textual explanation "Query matches documents that have one or more of the following words.", and redisplays the user's query with blue OR's in between each term. For AND search engines, Transparent Queries displays the following textual explanation "Query matches documents that contain ALL of the following words", and redisplays the user's query with a blue AND separator between each query term.

3.2 Stop Word Removal

Although few web search engines perform stop word removal, it is a standard technique used in bibliographic search systems to decrease the index size and increase precision [14, 15]. Transparent Queries' feedback for stop word removal consists of the following textual explanation: *Words marked with a red line have been removed from the query because they are very common*. Additionally, each query term is listed below the textual explanation, and stop words are marked with red strikethrough lines (see Figure 1). Google performs removal of overly common words and provides both feedback about this transformation and offers users a way of overriding this transformation. The inclusion of this transformation in both the user study and TQ itself is an attempt to gain greater empirical insight into how users' perceive this query transformation and the degree to which they are able to understand related feedback.

3.3 Term Suffix Expansion

Many bibliographic search systems also use stemming both to transform the query and to create the document index. Although most web search engines do not perform stemming, some studies have shown that stemming improves retrieval performance [16]. We chose to focus on Term suffix expansion, a variant of stemming, because it is one of the few automatic stemming-like processes currently used by web search engines. In **term suffix expansion**, the search engine treats each query term as if it was written as the regular expression *term**, where * matches zero or more alphanumeric characters. Transparent Queries uses the following feedback message for term suffix expansion on Yahoo: *Query also includes words that start with words you entered.* Also, it shows the query terms with rotating suffixes (in blue text) to convey the effect of the suffix expansion (see Figure 2). For example, TQ's animation cycles through run*s*, run*ning*, and run*ner*.

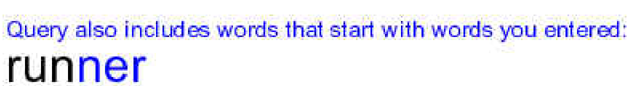

Figure 2: Transparent Queries feedback for term expansion. TQ provides feedback by rotating through the possible endings or suffix expansions for the query terms. Only one frame of the suffix animation is shown in this static capture.

3.4 Term Order Sensitivity

Google performs a combination of query transformations which we will refer to as term order sensitivity[16]. Google pays attention to both term proximity for matches and exact phrase matches. This processing results in term order sensitivity in that Google places greater weight on matches that occur in close proximity and in the same order listed in the query. TQ's feedback for this query transformation consists of the following textual explanation "Query will attempt to match the ORDER of your query words" followed by a listing of the query words.

4. USER STUDY

Although anecdotal evidence suggests that automatic query processes may pose usability problems, few published studies detail these problems or investigate user's conception of the search process. Therefore, we chose to study users' interpretations of opaque query transformations. The results from this study should serve as a starting point for future work on understanding the potential usability problems surrounding particular query transformations.

The second goal of this study is to evaluate Transparent Queries' suitability for reducing the uncovered usability problems. Our evaluation explored the suitability of the general approach used by TQ as well as the clarity of the specific explanations provided by TQ.

4.1 Study Design

To meet both of our study goals and maximize subjects' exposure to the automatic query transformations, we chose to ask subjects to perform predetermined (canned) queries using prespecified search engines. Counting on subjects to stumble upon the targeted query transformations through their own searching would have been very risky and extremely time intensive. The basic approach we used was to have the user issue one query across two carefully chosen search engines for each of the four targeted query transformations. We selected queries and search engines that would exaggerate and highlight the effect of the targeted query transformation. The order of the four tasks was randomized for each subject to minimize the effect of a systematic bias due to knowledge transfer or other unanticipated effects. In addition, the order in which the pair of search engines were used for each of the tasks was randomized. Each subject received identical directions at the beginning of the session.

The complete study consisted of single sessions with 14 participants. Only one subject was a computer professional. Each subject participated in a single session that lasted roughly 45 minutes and was broken down into two phases: the first phase explored users' mental models, reactions, and coping strategies for the targeted query transformations; and the second phase evaluated the suitability of TQ as an explanatory device. For the first phase, subjects used a standard web browser to perform the query tasks. During the second phase, subjects used a Transparent Queries augmented web interface (see section 3).

During the first phase, we asked subjects to describe the kinds of results they expected before executing the task's query. Our goal was to gain insight into the subject's existing mental models of search engine operation. The user then executed the query and then we asked for their opinions or judgement on the returned results. Essentially, we allowed subjects to discover the effects of the targeted transformation in terms of differential results. This approach guarantees that the subject experiences the effects of the targeted query transformations and maximizes the chance that they will notice the difference in query results and attempt to reconcile it. Our questions about subjects' expectations and their judgements of query results are a form of open-ended interviewing. The general nature of the questions was intended to assist subjects in expressing their thoughts and understandings of the query tasks without biasing or leading their judgments. We relied on qualitative data collection because we wanted to gain a rich understanding of how users perceive, explain and cope with potentially confusing search engine operations – specifically the relationship between a query and the returned results. Additionally it was not clear what quantitative measures would shed sufficient light both on subjects' mental models of search engine operation and their reaction to the confusing situations they would encounter during the study.

The second phase focused on evaluating the usability and utility of the Transparent Queries system. The same search tasks used in the first phase were also used in the second phase. The primary difference is that rather than ask subjects for their expectations about query results, we instructed users to execute the task queries and then pay special attention to the explanation provided by TQ. Subjects were asked to relate their understanding of the TQ explanation and to comment on whether or not it helped to explain the returned results. The next section details the four query tasks.

4.2 Query Tasks

As mentioned in the previous section, we employed four specific query tasks, one for each targeted opaque query transformation. In the discussion that follows, we provide details on each query task including the specific query used and the pair of search engines used.

4.2.1 Automatic Boolean Operator Task

For the automatic Boolean operator task, subjects used Northern Light and Excite to execute the following three-term query: *scuba snorkel arooba*. The chosen two search engines vary in terms of their default Boolean operator: Northern Light defaults to AND logic while Excite defaults to OR logic. Such a query might be issued if someone was interested in water sports in Aruba, but neglected to use the correct spelling. For this task we decided that the best way to maximize the effect of the different Boolean logic was to use a query that returned many documents with an OR engine and returned very few (preferably zero) documents with an AND engine. This stark difference in terms of results for the same query would confront users with a difference in search engine operation and would provide them an opportunity to reflect upon possible reasons for the seeming discrepancy. However, the pervasive use of full text indexing and an ever expanding collection of extremely diverse documents on the web made it difficult for us to formulate a query that returned zero or very few hits. After repeated attempts, we decided to use a misspelling, arooba, as a way of creating the desired type of query. We notified subjects of the misspelling immediately before this particular task.

4.2.2 Stop Word Removal

As with all our tasks, we wanted a query that would result in conspicuously different results between a search engine that performed stop word removal and one that did not. We chose the common query, *to be or not be* for this search task. The two search engines for this task were Altavista, which performs stop word removal, and Lycos, which does not perform this query transformation. Because each and every query term in this example is a stop word in Altavista (verified manually) , the query returns zero results with that engine, but Lycos returnsseveral relevant results.

4.2.3 Term Suffix Expansion

For this task, we wanted a query that showcased the difference between Yahoo's use of term suffix expansion and another similar directory-based search engine, Dmoz: Open Directory Project, that searched only on the exact query terms. We chose to use a simple one-term query for this task: *run*. The idea here was that many people might expect information about related athletic activities such as runn*ing*. However, only Yahoo would return such information because Dmoz focuses only on the exact query terms and as a consequence returns documents such as pages related to entertainment acts with *run* in their title.

4.2.4 Term Order Sensitivity

This task features two queries *fire boat* and *boat fire* and uses both the Google and Lycos search engines. This task takes advantage of the slightly different semantic meanings of the two queries, a ship used to fight fires, and a disaster involving a ship on fire. Because of it's term sensitivity Google's highest ranked results match each query quite well. Lycos does not pay attention to term order and returns the same results for each query. The differences between results returned by the search engines are far more subtle than that returned for the other three search tasks. For this reason, we use the two reversed queries to make the differences more conspicuous.

5. MENTAL MODELS OF SEARCH ENGINES

As subjects performed the query tasks with the standard non-augmented web browser, they encountered the effects of the four query transformations in the form of varying search results for identical queries.

5.1 Mental Models related to Boolean Operator

For this task, subjects issued the *scuba snorkel arooba* query to both Northern Light and Excite. Since *arooba* is a misspelling of Aruba and Northern Light uses a default AND Boolean operator, Northern Light does not return *any* results for this query. Because Excite uses a default OR Boolean operator, it returns several thousand results containing *scuba* and/or *snorkel*.

Eight of fourteen subjects thought that the query *scuba snorkel arooba* would return many results related to both the scuba and snorkel query terms but due to the misspelling very few results mentioning arooba (or Aruba). This expectation tends to suggest a mental model leaning more towards a Boolean OR interpretation than a Boolean AND interpretation. This is not to say that these subjects expected Northern Light and Excite to use a Boolean OR operator nor that they even understood the difference in those particular terms. This expectation may have been largely driven by the combination of an acknowledged misspelling of arooba and a belief that search engines usually return at least some results (often an overwhelming number). Accordingly, these subjects expected the results they received with Excite (OR) but were surprised by the lack of results from Northern Light (AND).

Five subjects were unable to come up with an accurate explanation of the vastly different results between Northern Light (nothing) and Excite (several thousand results containing a mix of scuba and/or snorkel). These subjects thought that the varying results were likely the product of some key difference in search engine operation but that was the extent of their insight. One such subject echoed this sentiment: "Maybe it's just a different search engine that focused on something else. I don't know". Six subjects were able to come very close to creating a model that accurately described the key differences between the OR and the AND search engine's operations. Of these, one subject decided that Northern Light (AND) must be looking "for all three words together" because of the following expectation, "I would imagine that scuba and snorkel would return some match". Three of these six subjects arrived at their mental model only after comparing the results from both search engines after this query task was completed. Finally, three subjects demonstrated mental models that closely resembled a notion of Boolean AND but did not demonstrate models resembling Boolean OR.

These subjects' reactions and attempted explanations for the Boolean Operator task suggest that this query transformation results in confusion for end-users. Many subjects were not able to come up with even the beginnings of an explanation or mental model for this task's query results. Other subjects required the conspicuous and rather unrealistic back-to-back comparison provided by the study design before they produced a mental model that closely mirrored the differences between the two default Boolean operators. Given these results, it is unlikely that users would be able to correctly attribute differences in search results across search engines caused by a different default Boolean operator to this typically opaque but rather central query transformation. After getting feedback through Transparent Queries, one subject even mentioned that he now finally understood why he often got such poor results from Excite and that he had previously just assumed that it was just a poor or outdated search engine.

5.2 Mental Models related to Stop Word Removal

For this task subjects submitted the following query to both Altavista and Lycos: *to be or not to be*. Altavista returns zero hits for this query as each term of the query is treated as a stop word and therefore the transformed query amounts to a null query. Lycos returns a few targeted results dealing with Shakespeare and Hamlet.

Ten of fourteen subjects expected to get information about Shakespeare and/or Hamlet for the *to be or not to be* query. Only two of these subjects explained the query results with a mental model that approximated the notion of stop words. common words". Two users explained the unexpected query results by reasoning that Altavista was lacking in its literature content. Another user responded with a slightly more vague model: "the databases (Altavista and Lycos) are not the same".

Six of the fourteen subjects recognized that the query *to be or not to be* might be problematic and result in "random" results. These users specifically commented that they had low expectations for results, because the words were just too "common" or because the search just was not precise enough. This result suggests that these subjects have a mental model concerning the need for specific queries. Of these six subjects, five of them demonstrated mental models that explained the differential results as a direct result of the problematic nature of the query terms (too short or too common). However, only one of these subjects suggested that the query terms were actually being dropped from the search.

Nine of the fourteen users were unable to explain the query results in a way that even remotely resembled the removal of stop words. Given the results for this query task, it is likely that the opaque transformation of removing stop words would confuse many users.

5.3 Mental Models related to Term Order Sensitivity

For this query task, subjects issued two queries *fire boat* and *boat fire* to both Google and Lycos. Lycos completely ignores the order of the query terms and thus treats the two queries as identical and returns identical results for each query. Google pays attention to both term proximity and exact phrase matches. Thus, Google returns different results for the two queries.

Only five of the fourteen subjects expected different results between the *fire boat* and *boat fire* queries. These expectations corresponded closely to the concept of a *fire boat* as a vehicle used to fight fires and of a *boat fire* as a disaster occurring on a boat. Several subjects appeared to be focusing on the individual query terms rather than seeing the query as a phrase. Thus, these subjects did not anticipate that switching the order of the terms would have an effect on the query results.

Although five of the fourteen subjects were able to explain the varying results between the two engines in a manner that captured some essence of term order sensitivity, many of these mental models were vague and incomplete. One such vague model was provided by a subject:

> "Lycos separates the two words and searches for the meaning, instead of what you're looking for. Google understands the meaning of the phrase".

This subject's very rough and incomplete mental model incorporates some notion of the difference between the way the two search engines perform the queries. The notion of a "phrase" can be seen as incorporating Google's use of automatic proximity processing (similar to an automatic NEAR operator). In contrast, one subject provided a clear albeit humorous explanation for this task's query results: "I think they, the little men in there, look at the order of the words."

Three subjects expressed a preference for results where the two terms appeared adjacent to one another, but not necessarily in a specific order. These subjects also thought that search engines should list results with the terms appearing "together" ahead of results where the terms were "separate". In other words, these subjects had an expectation that search engines would or should automatically apply some query transformation that takes term distance into account when presenting results. One subject echoed these expectations in his comments about results from Lycos:

> "That's weird, the one with *fire* and *boat* together is fifth. I thought that it listed the best match first… where the words appeared together".

In analyzing this data through repeated passes, the common practice of highlighting or bolding occurrences of matching query terms in the result summaries appeared to drive many subjects' understanding of the difference between the way Google and Lycos were processing this tasks queries. Specifically, subjects noticed that although both the terms "fire" and "boat" appeared in the result summaries provided by Lycos, they were typically separated rather than adjacent to one another. Conversely, these subjects noticed that the two query terms appeared adjacent to one another in Google's result summaries. Several subjects spoke of this in terms of searching for the terms "together" as opposed to separately. Thus, result summaries along with the practice of bolding matching terms, appear to be critical feedback for providing users with some basis for forming a mental model of how a particular search engine deals with multiple term queries. Furthermore, this data would tend to suggest that dynamic query summaries which are based on the occurrences of the actual query terms help not only in providing a better understanding of the content or subject matter of documents, but they may also promote the formation of roughly functional mental models of key aspects of search engine operation.

5.4 Mental Models related to Term Suffix Expansion

This task required subjects to issue the simple *run* query using both Yahoo and Dmoz: Open Directory Project. The key difference in query operation is that Yahoo performs term suffix expansion while Dmoz matches only the exact term entered. For this query, Yahoo returns documents containing *run* but also returns documents containing *running, runner* and *runs*. Dmoz returns only documents containing the exact term *run;* thus, none of its returned results deal with athletic content (i.e. *running*). All fourteen subjects expected that *run* would return information about athletics and sports, and most subjects mentioned running specifically. Subjects were also very aware that they would probably receive other information, and several mentioned that they anticipated information on movie titles and also other uses of the word such as in "run a computer program". Given these expectations, it was not surprising that subjects were generally satisfied with results from Yahoo. However, many subjects (nine) were very surprised with results from Dmoz. Specifically, they noticed that Dmoz did not return any results related to athletic activities such as running. One subject's comments neatly describe the surprise and confusion echoed by many subjects:

> "No, it's horrible. How can you not have at least one [running – athletics related result] out of the top twenty?"

Ten of fourteen subjects provided some explanation of the surprising query results for this task, but four subjects were not able to offer any explanation of the query results. Of the ten subjects who provided explanations, only three subjects demonstrated mental models that closely approximated the term expansion transformation. Interestingly, two of these three subjects noticed that Yahoo highlighted *run* even when it was a part of *running*. Coupled with Dmoz's highlighting of the plain *run* term, these subjects were able to conclude that Dmoz was only looking for the exact term, and in contrast, Yahoo was looking for other forms of *run*. The remaining seven subjects demonstrated mental models that differed greatly from the actual operation of Yahoo and Dmoz for this task. The majority of these subjects (six subjects) decided that Dmoz must not have much information (if any) on athletics. Although this is clearly not the case, subjects seemed earnest in their production of this mental model. This model seems driven by unfamiliarity with Dmoz and the strong expectation of information about the sport of running. The one remaining subject attempted to explain the query results in terms of "proper nouns".

Subjects' expectations that search engines return information about running from a query on *run* is not very surprising, but it is potentially problematic because very few of the popular commercial search engines perform any stemming-like query transformation. This result would tend to suggest that search engines should reconsider using some form of stemming to bring operation in line with users' expectations.

6. EVALUATION OF TRANSPARENT QUERIES

The second goal of our user study was to evaluate both the usability and utility of Transparent Queries. Specifically we focused on both the individual feedback mechanisms used by Transparent Queries and the basic lightweight feedback concept used by TQ. We present specific findings related to individual feedback mechanisms for the targeted query transformations and then present findings related to TQ's overall usability and utility.

6.1 Boolean Operator Feedback

In order to illustrate the potential utility and shortcomings of TQ feedback in helping users to come up with an accurate mental model of query operation, we discuss the experiences of two subjects who initially did not demonstrate an understanding of Boolean AND and Boolean OR.

During the first phase of the study, one subject commented that she had never heard of Northern Light but she had heard of Excite. She concluded that Excite was a superior search engine and that they "know more stuff". This reasoning represented her mental model as to why she had received no results from Northern Light. After viewing the TQ feedback, this subject was able to quickly replace her previous model with one that closely captured the differences in operation between Boolean AND and OR. Thus, TQ helped this subject to improve her mental model about this particular query transformation.

Another subject provided a similar reasoning for the discrepant results during the first phase: "Northern Light must be older". Although she seemed to understand the feedback for the OR operator, the AND feedback confused her. Specifically, she didn't understand why TQ was saying "Query matches documents containing ALL of the following terms: (query terms)" when in fact she did not receive any results. She thought that TQ was indicating that NL had found matches for all three words including the misspelled term. This is a subtle point and seems to indicate that users might misinterpret TQ feedback as an indication of what the search engine found rather than what it was looking for. It also suggests that TQ might be able to provide better feedback if it was designed so that it could take the actual returned results into account when presenting feedback to the user.

For the majority of subjects who were able to produce a mental model closely approximating the difference between the use of the Boolean operators, TQ confirmed their understanding. In one case, the subject was able to use the TQ feedback to improve the accuracy of his mostly correct mental model. This subject understood that Excite was looking for one or more of the terms, but had decided that it was ranking scuba ahead of snorkel based on the order of the returned results. After seeing the TQ feedback, this subject decided that there was no preference for one term over the other.

6.2 Stop Word Removal Feedback

With the benefit of the TQ feedback, most subjects were able to understand that Altavista had removed all their terms from the search. However, many thought that this wasn't too helpful partly because they didn't understand why the fact that a term was "too common" should matter and partly because TQ did not offer any suggestions on an alternative strategy or next step that would generate the results they were seeking.

6.3 Term Order Sensitivity Feedback

A few subjects were confused by the TQ feedback for Google on this task. Mainly they didn't quite understand what was meant by "order". However, the remainder of the subjects decided that the TQ feedback meant that Google was paying attention to the order of the terms while Lycos was not.

6.4 Term Suffix Expansion Feedback

Based on subject reactions, the TQ feedback for the term suffix expansion was the most readily understood and most appreciated of the various feedback messages displayed by TQ. The TQ feedback for this query transformation is more closely tuned to the specific query than the other TQ feedback messages. Specifically, the animated suffixes were based on and directly related to the single query term *run*. We believe the query specificity of the feedback is the primary reason for its success and popularity among subjects.

6.5 TQ Design Issues

The user study revealed several potential areas for improvement in the way TQ provides feedback to users. Despite our attempts to inform subjects that TQ provides feedback about how a search engine processes their query, a few subjects had trouble realizing that the TQ feedback for a query was directly tied to how the current search engine was processing the query. They appeared to view the TQ feedback as a type of generalized help about query formation techniques or something similar. We believe that rewording the queries so that the search engine becomes the active agent of the matching such as "Lycos will attempt to …" might help alleviate some of this confusion. Other minor problems included complaints about the use of future tense "will attempt to match" rather than past tense for explanations. These results suggest that wording of the textual explanations is a very delicate design task.

Furthermore, several subjects expressed a desire for Transparent Queries to provide a way of allowing the user to change the way a query is processed. So a user might be able to use a selection box embedded in the TQ interface to change from using stem expansion to just using the exact entered terms or vice versa. This is basically the key difference between transparent and penetrable interfaces. [12] One possible extension of this work would be to add query revision tools to TQ so as to provide a penetrable interface for both understanding *and* submitting queries.

7. LIMITATIONS AND FUTURE WORK

Although we had few subjects in this small-scale qualitative study, we did find a substantial mismatch between most users' mental models of web search engine operation and their actual operation. However, we did not prove that users' inaccurate models lead to poor retrieval performance. A large-scale study is needed to determine the effects of transparent or penetrable interfaces for query transformations on search precision and recall. We envision patterning such a study after Koenemann and Belkin's evaluation of the differences between opaque, transparent and penetrable interfaces for query reformulation [12].

Our study relied on contrived queries, and thus provides no insights on the prevalence, severity, or consequences of automatic query transformations. Further studies could use web log data to determine the frequency of these query transformations during users' regular searching activities. Such a study could help prioritize efforts to address the related usability issues in terms of both revised search engine design and augmentative tools such as Transparent Queries.

One significant challenge for carrying out the lightweight feedback approach employed by Transparent Queries is the need for providing simultaneous feedback for multiple concurrent query transformations. Transparent Queries currently provides feedback for a single transformation used in processing the user's query. Future designs must address this limitation.

Finally, subjects' comments during the evaluation of Transparent Queries strongly suggest that transparency was certainly helpful but not sufficient in helping users cope with the use of opaque query transformations. Future interfaces should provide a straightforward way for users to leverage their newfound understanding of query transformations and allow them to select and choose which query transformations they would like to apply to their queries. Thus, although transparent interfaces are promising, penetrable interfaces seem to offer better potential for helping users apply query transformations to satisfy their information needs.

8. CONCLUSIONS

Based on the study data, many users clearly have multiple misconceptions about how search engines process their queries. In the absence of feedback on those query transformations, users come up with naïve and erroneous mental models of search engine operation. This study design made the differing query transformations of search engines much more conspicuous than such differences would be in everyday use, but many users still did not understand how the search engines were transforming their queries. Thus, we conclude that many users will not be able to understand why they often receive erratic and confusing results from search engines. The study results strongly suggest that the opaque operation of query transformations represent a substantial barrier for users in their attempt to understand how various search engines process queries. As one subject attempted to understand confusing search engine results related to an opaque query transformation, he commented out of frustration that:

> "it's almost as if they [search engines] tell you what they want to give you, you want to find something but no, this is what *we* want you to have"

Furthermore, poor mental models of search engine functionality tend to lead users into making poor decisions regarding their search techniques. For instance, users might decide that a search engine that uses an opaque default Boolean AND operator (such as Northern Light) is really just outdated or bad and therefore should be avoided.

For search engine developers, user studies such as ours can point to features that users expect. In our study, we found that users expect some form of stemming, even though few systems support it. Thus web search engines should seriously consider adding this feature.

The evaluation data suggest that Transparent Queries helps users learn about the differences in how search engines process their queries, and thus discover inconsistencies between their model of search engine behavior and the actual underlying operation. The results of the evaluation also suggest that providing lightweight, concise, yet helpful and meaningful feedback on query transformations is a formidable and delicate task. Accordingly, Transparent Queries was much more successful in conveying the operation of term suffix expansion than in providing information about the use of a default Boolean operator. It is clear that more research needs to be done on the most effective way of informing users about opaque query transformations.

Our study results indicate that users need to understand some of the practical differences among the key query transformations that various search engines use. Armed with this knowledge, users both will be better able to choose search engines that meet their needs based on objective rather than potentially naïve observations and will know enough to look for ways to specify desired query transformations. This study naturally raises the question of how much transparency and penetrability is really appropriate or sufficient. The answer or the beginnings of an answer lie in future studies of the utility and usability of search interfaces that provide users with both feedback and control over currently opaque query transformations. Our work strongly suggests the need for more work along these dimensions. However, it should not be misread as a call for complete or unfettered transparency or penetrability. We recognize both the importance of search engine designers' skill and intuition as well as the potential for information overload from too many options and too much information about the countless numbers of query transformations. Based on the limited evaluation of Transparent Queries, we conclude that transparent and penetrable interfaces have the potential to balance the need for understanding query operation and the desire for search tools that maximize a user's productivity.

9. REFERENCES

1. Norman, D; *The Design of Everyday Things*. 1988, New York: Doubleday.
2. Notess, GR; Search Engine Showdown: A User's Guide to Web Searching. 2001.
3. Hölscher, C and Strube, G; Web Search Behavior of Internet Experts and Newbies. in *The Ninth International World Wide Web Conference*. 2000. Amsterdam
4. Choo, CW, Detlor, B, and Turnbull, D; Information Seeking on the Web - An integrated model of browsing and searching. in *Proceedings of the Annual Meeting of the American Society for Information Science (ASIS)*. 1999. Washington DC
5. Navarro-Prieto, R, Scaife, M, and Rogers, Y; Cognitive Strategies in Web Searching. in *Proceedings of the 5th Conference on Human Factors & the Web*. 1999
6. Borgman, CL; Why are online catalogs hard to use? Lessons learned from information-retrieval studies. *Journal of the American Society for Information Science*, 1996. 37(6): p. 387-400.
7. Borgman, CL; The User's Mental Model of an Information Retrieval System. in *SIGIR-85: Proceedings of the eighth international conference on Research and Development in Information Retrieval*. 1985.Pages 268 - 273
8. Shneiderman, B; *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Second Edition ed. 1992, Reading, MA: Addison-Wesley.
9. Shneidermann, B, Byrd, D, and Croft, WB; Clarifying Search: A User-Interface Framework for Text Searches., in *D-Lib Magazine*. 1997.
10. Shneiderman, B; A framework for search interfaces. *IEEE Software*, 1997. 14(2): p. Inclusive Pagination: p. 18-20.
11. Anick, PG, et al.; A Direct Manipulation Interface for Boolean Information Retrieval via Natural Language Query. in *SIGIR-90: Proceedings of the thirteenth international conference on Research and development in information retrieval* ,. 1990.Pages 135 - 150
12. Koenemann, J and Belkin, N; A case for interaction: A study of interactive information retrieval behavior and effectiveness. in *CHI 96 Human Factors in Computing Systems*. 1996.pp. 205-212
13. Beaulieu, M and Jones, S; Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting with Computers*, 1998. 10(3): p. 237-48.
14. Luhn, HP; A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Develoopment*, 1957. 1(4).
15. Salton, G and McGill, MJ; *Introduction to modern information retrieval*. McGraw-Hill Computer Science Series. 1983, New York: McGraw-Hill, Inc.
16. Lennon, M, Pierece, D, Tarry, B, and Willett, P; An evaluation of Some Conflation Algorithms for Information Retrieval. *Journal of Information Science*, 1981. 3: p. p. 177-183.

.