Automatic Identification of User Goals in Web Search

Uichin Lee
University of California
Los Angeles, CA 90095
uclee@cs.ucla.edu

Zhenyu Liu
University of California
Los Angeles, CA 90095
vicliu@cs.ucla.edu

Junghoo Cho
University of California
Los Angeles, CA 90095
cho@cs.ucla.edu

ABSTRACT

There have been recent interests in studying the "goal" behind a user's Web query, so that this goal can be used to improve the quality of a search engine's results. Previous studies have mainly focused on using manual query-log investigation to identify Web query goals. In this paper we study *whether* and *how* we can automate this goal-identification process. We first present our results from a human subject study that strongly indicate the feasibility of automatic query-goal identification. We then propose two types of features for the goal-identification task: *user-click behavior* and *anchor-link distribution*. Our experimental evaluation shows that by combining these features we can correctly identify the goals for 90% of the queries studied.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*; H.4.m [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement, Experimentation, Human Factors

Keywords

Web search, user goals, query classification

1. INTRODUCTION

Given the impact of search engines on the Web users' experience, improving the quality of search results has become the holy grail of search engine operators [1, 2, 3, 4, 5]. As part of this endeavor, there has been a recent interest in identifying the "goal" of a user during a search [6, 7, 8], so that the identified goal can be used to improve page ranking [2, 3, 7], result clustering [9, 10, 11] and answer presentation [12, 13].

In their seminal studies, Broder [6] and Rose and Levinson [8] have independently found that the goal of a user can be classified into at least two categories: *navigational* and *informational*. A query is considered *navigational* when a user has a particular Web page in mind and is primarily interested in visiting the page. *Informational queries*, on the other hand, refer to the queries where the user does not have a particular page in mind or intends to visit multiple pages to learn about a topic. In their studies, Broder [6] and Rose and Levinson [8] identified the goal of queries through user surveys and manual query-log investigation and proposed the automatic user-goal identification as an open research problem.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-046-9/05/0005.

In this paper, we study *whether* and *how* we can identify the user goal automatically without any explicit feedback from the user. There are two main challenges in studying this problem:

- *Do most queries have a predictable goal?* A user's goal for a query is inherently subjective. Thus, the first question is whether it is ever possible to associate a query with a particular goal simply by looking at the query without any user feedback. For example, our user study shows that most users associate the query *bestbuy* with the official BestBuy Web site and consider the query navigational, while the user opinion on the query *Alan Kay* is evenly split. Some people want to visit the homepage of Alan Kay, while others want to read multiple pages related to Alan Kay in order to learn about his career and research given his recent reception of Turing Award. When the user opinion is evenly split, it will be clearly difficult for a search engine to reliably predict the goal of a user without collecting any further information from that user. Given the above sample queries, it will be highly interesting to study how many queries will have a predictable goal, and how many queries will be "unpredictable" in their goals and require further information from the user for reliable prediction.
- *What features can we use to identify the user goal?* For the queries with a predictable goal, what features can we use for prediction? Do we need to understand the semantic meaning of a query or are there simple yet effective features that we can exploit?

In this paper, we first assess the predictability of a query goal through a human subject study. We then propose *past user-click behavior* and *anchor-link distribution* as potential features for the goal prediction. In particular, we make the following contributions in this paper.

- In Section 2, we describe our human subject study, in which we ask 28 participants in the UCLA Computer Science Department to indicate their potential goals for 50 most popular queries issued from the department. The purpose of this human subject study is twofold: (1) We evaluate the feasibility of automatic user-goal identification by checking whether a

large number of queries have a predictable goal. (2) We build a benchmark set of queries and their goals, so that we can evaluate the effectiveness of automatic goal-identification methods.

The result of our study is very promising. Our study shows that the majority of queries have a predictable goal; most of our subjects agreed on a particular goal (either navigational or informational) for these queries. Furthermore, our study suggests that there may exist an easy method to identify the queries whose goals are difficult to predict. We elaborate more on these findings in Section 2.

- In Section 3, we propose two features for the prediction of a user goal: *past user-click behavior* and *anchor-link distribution*. The basic intuition is that if a query is navigational, users will primarily click on the result that the user has in mind. Therefore, by observing the past user-click behavior on the query, we can identify the goal. Similarly, if users associate a particular query, say *bestbuy*, with a particular Web site, say the official BestBuy Web site, then we expect that most of the links that contain *bestbuy* in the anchor will point to the official Web site. Therefore, by observing the destinations of the links with the query keyword as the anchor, we may also identify the potential goal of the query.
- In Section 4, we evaluate the effectiveness of our proposed features using the benchmark queries from our human subject study. Our study shows that each individual feature enables us to achieve an accuracy of about 80%. Combined together, we achieve an accuracy of 90%. We also compare the effectiveness of our features with existing methods.

2. RESULTS OF HUMAN SUBJECT STUDY

We start our discussion with the description of our human subject study, in which we try to (1) evaluate how many queries have clearly predictable goals and (2) build a benchmark query set against which we can evaluate our automatic identification mechanisms.[1]

Roughly, our benchmark set consists of 50 most popular queries issued to Google from the UCLA Computer Science Department.[2] To study whether the goals of these queries are predictable regardless of individual users, we asked 28 graduate students in the department to indicate their most probable goal if they issued each query.

We decide to limit our user survey to CS graduate students mainly because of their ease of access. However, we believe this restriction does not introduce a significant bias in our result, because the queries are also collected from the same department. Since our subjects are likely to be familiar with the queries, we believe that they are likely to provide the most probable goal for those queries.

In the rest of this section, we describe our human subject study in more detail. In Section 2.1, we describe the taxonomy of user goals used in our study. In Section 2.2 we explain the exact questionnaire design of user survey. In Section 2.3, we provide the main results from our survey.

[1]We have contacted researchers who have built their proprietary benchmark sets in the past [6, 8]. Unfortunately, due to legal and technical constraints, we could not obtain their benchmark sets.

[2]More precise description on how these queries were collected is given in Section 4.

2.1 Taxonomy of queries

In our study we use the following taxonomy of query goals, largely based on [6, 8]:

- **Navigational queries.** By asking a navigational query, e.g., *citeseer* or *bestbuy*, a user already has a Website in mind and the goal is simply to reach that particular site. Note that for such a query, the user may either have visited that site before, or just assume such a site exists. For a navigational query, typically users will only visit the "correct" Website they have in mind.
- **Informational queries.** By asking an informational query, e.g., *hidden markov model* or *simulated annealing*, a user is exploring Websites or Webpages that provide background knowledge about a particular query topic. For an informational query, typically users do not pre-assume a particular Website to be the single "correct" answer, and they are willing to click on multiple results.

Note that the taxonomies proposed in [6, 8] are more detailed than ours; both have third categories — *resource queries* in [8] and *transactional queries* in [6] — and the categories are refined further into smaller subcategories. Due to the lack of consensus on the third category and to make our classification task manageable, we mainly focus on the two categories, *navigational* and *informational*, described above. It will be an interesting future work to see whether further refinement of user goals can be done automatically.

Also note that given the above definitions, there exist two potential criteria for classifying a query either as navigational or informational. One criterion is whether the user has a particular Website in mind when the user issues the query. Another criterion is whether the user intends to look at only a single Website or to look at multiple sites in the search results. As we will see later, these two potential criteria caused some confusion in our user survey, for which we had to make a certain decision.

2.2 Questionnaire design

A good design of the survey questionnaire is crucial in collecting reliable results from our user study. In the following, we describe the exact questions that we used in our survey and how our questionnaire has been refined to our final form through multiple revisions.

In our initial design stage, we first evaluated whether it is appropriate to directly use the navigational-informational taxonomy in our questionnaire. For this purpose, we interacted with four participants, first educating them with the taxonomy, and then asking them to classify the 50 queries as either navigational or informational. Afterwards we interviewed each of them to gather descriptive intentions for some representative queries, and further compared such descriptive intentions with the final navigational/informational choices. From this comparison we realized that even if two participants had exactly the same descriptive intention, they might end up casting that intention into different navigational-informational choices. This confusion was mainly due to the two potential criteria that they could use to classify the user goal.

For example, a user might search a person's name in order to reach not only that person's homepage, but also some other related sites, such as the person's DBLP publication

page or news articles about the person. In this scenario, the people who used the first criterion (*"do you have a particular Webpage in mind?"*) classified the intention as navigational, because they perceived a particular Webpage (the person's homepage) and reaching that page was part of the goal. On the other hand, the people who used the second criterion (*"do you intend to visit multiple pages?"*) classified it as informational because their goal was to gather information from *multiple* sites including the person's homepage.

Realizing this potential ambiguity and the randomness in the user classification, we decided to ask our subjects to indicate their descriptive intentions directly. Based on their descriptive intentions, we then classify the goal of the queries ourselves. In particular, we decided to present the following three choices to our participants:

- **Choice 1**: I already have a particular Website (or Webpage) in mind, and my major interest is just to reach that site (page) through the search engine.
- **Choice 2**: I know there's a particular Website (or Webpage) corresponding to this query. However, my interest is not only to reach that site, but also to visit some other sites returned by the search engine.
- **Choice 3**: I have no particular Website (or Webpage) in mind. I am willing to click on multiple results returned by the search engine.

Note that under both criteria, Choice 1 is clearly navigational because the user intends to visit a *single* Website *that he has in mind*. Similarly, Choice 3 is clearly informational because the user intends to explore multiple Websites and no Website is pre-assumed to be the single "correct" answer. The ambiguous case is Choice 2; depending on which criterion we use, it can be classified as either navigational or informational. We explored both possibilities in our study, but due to space limit, we report the result when we use the second criterion (*"do you intend to visit multiple sites?"*) and classify Choice 2 as informational. We report the corresponding results when we use the first criterion in the extended version of this paper [14].

After this decision, we went through one more revision of the questionnaire by handing out a draft version to three participants, asking for their feedback, and rephrasing some of the descriptions and reordering the sequence of presentation based on the feedback. After this final revision, we distributed our questionnaire to 28 graduate students inside our department and collected the final results.

As a final note, we also asked participants to indicate their familiarity with each query in our survey form, by marking a query either as *familiar* or *unfamiliar*.

2.3 Manual classification results

Given our survey results, we can summarize the manual classification result of a query q into a single value $i(q)$ which is the percentage of participants who indicate its goal as informational.[3] For example, the $i(q)$ value for query "IEEE Explore" is 0.036, which means 3.6% of the 28 participants has an informational goal for this query and the other 96.4% has a navigational goal. Given this $i(q)$ representation, we can safely classify a query q as informational if $i(q)$ is close to 1, and similarly as navigational if $i(q)$ is close to 0. We refer to a query as *unpredictable* when the user opinion on the query is evenly split and its $i(q)$ value is close to 0.5 — when the goal depends on individual users, it may be difficult to predict a particular user's goal.

[3]In computing such statistics we discard queries that a participant indicated as unfamiliar.

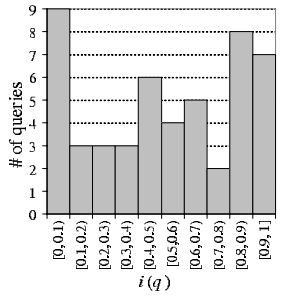


Figure 1: Query distribution along the $i(q)$ axis

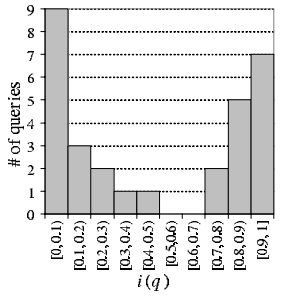


Figure 2: After removing software and person-name queries

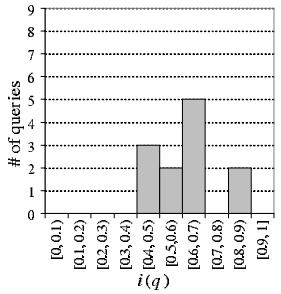


Figure 3: Distribution of the 12 software queries

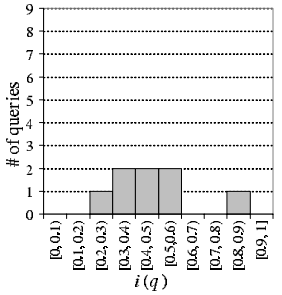


Figure 4: Distribution of the 8 person-name queries

We now present the $i(q)$ statistics for the 50 queries studied. Our main focus of this section is as follows:

- **Dichotomy or spectrum?** Do we observe clear separation between informational and navigational queries, or do we see a full spectrum of queries ranging from those that are clearly navigational, to those that are unpredictable, and eventually to those that are clearly informational?
- **Why unpredictable?** What are the unpredictable queries? Do they share any common features? What reasons cause such unpredictability? How can a search engine identify such queries and process them?

2.3.1 Dichotomy or spectrum

Figure 1 shows the distribution of the 50 queries along the $i(q)$ axis. For example, the leftmost bar shows that there are 9 queries with $i(q) \in [0, 0.1)$, which means that less than 10% of the participants indicate the informational goal for these queries. In other words, these 9 queries are "highly navigational."

Figure 1 suggests that, if we consider the 50 queries as a whole, a majority of queries have reasonably clear goals, but there is no clear dichotomy between informational queries and navigational queries. For instance, if we classify queries with $i(q) \leq 0.2$ as clearly navigational and those with $i(q) \geq 0.8$ as clearly informational, then 23 queries (46%) belong to the unpredictable region in between. In the next subsection we study those 23 queries in more detail.

2.3.2 Unpredictable queries

Our primary interest in these 23 queries is whether they share anything in common. To our surprise, 17 queries

out of the 23 queries (73.9%) belong to two topic categories, namely software names (e.g., "cygwin," "spybot," "ns2," etc.) and personal names (mostly computer science researchers inside or outside of our department). The other 6 queries have rather diversified topics, ranging from online services to news and events.

The above finding has led us to investigate all the software and person-name queries in our 50-query set, to study whether all such queries tend to be unpredictable. Among all 50 queries, 12 are software names and 8 are personal names. The $i(q)$ distribution for these two categories of queries are shown in Figures 3 and 4, respectively. The results show that 10 out of 12 (83.3%) software queries and 7 out of 8 (87.5%) person-name queries have their $i(q)$ values within $[0.2, 0.8]$, which suggests they are unpredictable.

Naturally we are interested in why software and person-name queries are unpredictable. To answer this question, we further interviewed six participants to collect anecdotal evidences behind their diversified answers. Following are the possible explanations obtained from the interview:

- **Software queries:** Given a software query, some participants chose Choice 1 (navigational) because they simply wanted to visit the official Website maintained by the software development team and they felt safer or more efficient to visit that site to download the latest version or fixes. Others chose Choice 2 or 3 (informational) because either (1) they were willing to click on any site as long as the site provides a downloadable version of the software, or (2) they were looking for comments, reviews or usage tips about the software hosted by sites other than the official one.
- **Person-name queries:** Participants who chose Choice 1 (navigational) for a person-name query were either (1) very familiar with that person and they knew exactly what to explore after they reach that homepage (e.g., to download research papers, reach their research groups, etc.) or (2) totally unfamiliar with the person so they just wanted to learn the basics of the person by visiting the personal homepage. Others chose Choice 2 or 3 (informational) to explore pages other than (or in addition to) a person's homepage, such as the person's DBLP publication page or news articles related to this person such as recent prizes, awards, etc.

By removing the 20 queries that are related to software and personal names, we obtain the distribution for the other 30 queries, as shown in Figure 2. We now observe clear separation towards the two ends, $i(q) = 0$ and $i(q) = 1$, which means that most of these 30 queries have predictable goals.

The following is a short summary of our main findings in this section:

- We observe that a large fraction of queries can be associated with a particular goal that most users agree on. These queries may be amenable to the automatic classification of the user goal.
- We observe that most of the "unpredictable" queries tend to belong to a few topic categories, such as software or personal names. Thus, it may be possible that a search engine can detect such queries using a topic-detection method [15] and treat them separately from other queries with predictable goals.

Given the ambiguity of the user goal for software and person-name queries, we will primarily use the 30 queries that are *not* software or person-name related, when we evaluate automatic goal-identification methods.

3. AUTOMATIC IDENTIFICATION OF QUERY GOALS USING VARIOUS FEATURES

In this section, we propose two categories of features for the automatic identification of the user goal: *past user-click behavior* and *anchor-link distribution*.

3.1 Past user-click behavior

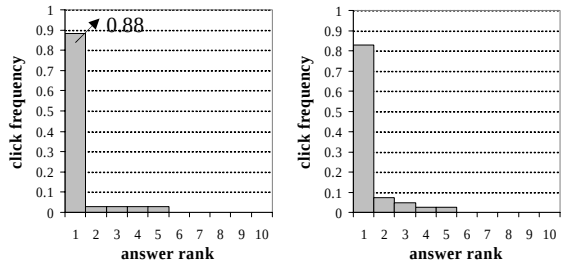
Click distribution. Our first feature is based on the intuition that the user's goal for a given query may be learned from how users in the past have interacted with the returned results for this query. If the goal of a query is navigational, then in the past users should have mostly clicked on a single Website corresponding to the one they have in mind. On the other hand, if the goal is informational, in the past users should have clicked on many results related to the query. Thus by observing how the results for a particular query have been clicked so far, we can tell whether the current user who issues that query has a navigational or an informational goal.

To formalize this idea, we introduce the notion of *click distribution* which captures how frequently users click on various answers. Given a query, its click distribution is constructed as follows: We first sort the answers to the query in the descending order of the number of clicks they receive from all users.[4] Afterwards we create a histogram where the i^{th} bin corresponds to the number of clicks accumulated on the i^{th} answer. We further normalize the frequency values so that these values add up to 1. For example, Figure 5(a) shows the click distribution for query *pubmed*. (Details about how the click data is collected is presented in Section 4.) The leftmost bar in the figure shows that for the query "pubmed," the top answer (www.ncbi.nlm.nih.gov/entrez/query.fcgi) got 88% of user clicks.

Given a query's click distribution, we can guess the goal for that query by investigating how that click distribution is skewed toward rank one. Intuitively, a highly skewed distribution suggests that a single answer is clicked much more often than others. Accordingly, the goal for the corresponding query should be navigational. On the other hand, a flat distribution suggests that the goal is informational. For example, from our benchmark set, we pick two queries that are clearly navigational: *pubmed* ($i(q) = 0.1$) and *UCLA library* ($i(q) = 0$), and we show their click distributions in Figure 5. We also show the click distributions for two queries that are clearly informational in Figure 6: *hidden markov model* ($i(q) = 1$) and *simulated annealing* ($i(q) = 1$). Apparently, distributions in Figure 5 are much more skewed toward rank one than those in Figure 6.

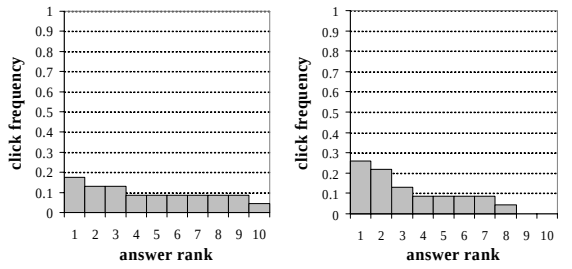
To predict a query's goal based on its click distribution, we summarize the distribution into a single numeric feature that captures how skewed the distribution is. Several standard statistical measurements exist to serve this purpose, including the mean, median, Skewness (the 3^{rd} central moment normalized by the standard deviation to the order of 3), and Kurtosis (the 4^{th} central moment normalized by

[4] For the vast majority of queries, this order is the same as the order in which they appear in the search result.



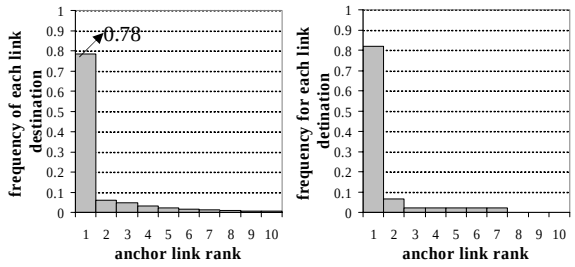
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 5: Click distributions for sample navigational queries



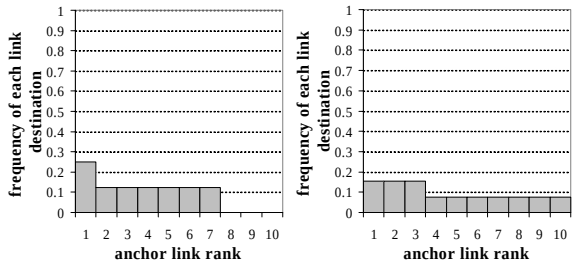
(a) hidden markov model ($i(q)$=1) (b) simulated annealing ($i(q)$=1)

Figure 6: Click distributions for sample informational queries



(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 7: Anchor-link distributions for sample navigational queries



(a) hidden markov model ($i(q)$=1) (b) simulated annealing ($i(q)$=1)

Figure 8: Anchor-link distributions for sample informational queries

the standard deviation to the order of 4) of a distribution. In Section 4 we experimentally evaluate the effectiveness of each of these measurements.

Average number of clicks per query. Besides click distribution, another feature embedded in the user-click behavior is how many results a user clicks on after the query is issued. Intuitively, for a navigational query, the user is most likely to click on only one result that corresponds to the Website the user has in mind. On the other hand, for an informational query, the user is most likely to click on several results. Therefore, we use the *number of clicks per query* as another potential feature based on user-click behavior.

One practical issue in using the user-click behavior is that a search engine needs to accumulate enough user clicks for a given query. Studies show that a large number of queries are issued multiple times, thus providing enough click data [16, 17]. For those queries without sufficient user-click data, search engines may use the feature that we propose in the next section.

3.2 Anchor-link distribution

Another feature that we may use is the destinations of the links with the same anchor text as the query.[5] For example, for a navigational query *pubmed*, a single authoritative Website exists (which is www.ncbi.nlm.nih.gov). As a result, if we extract all the HTML links with the anchor text *pubmed*, we expect to find that a dominating portion of these links point to that single Website; On the other hand, for a informational query *hidden markov model*, because of lack of a single authoritative site, we expect that the links with the anchor text *hidden markov model* point to a number of different destinations.

[5]An anchor is a piece of text surrounded by a pair of `<A HREF="..."> </A>` tags in an Web page, such as `<A HREF="http://www.ncbi.nlm.nih.gov">Pubmed</A>` where "Pubmed" is the anchor text and "www.ncbi.nlm.nih.gov" is the *destination link* for this anchor.

To formalize this idea, we introduce the notion of *anchor-link distribution*, similarly to what we did for user-click behavior. Given a query, its anchor-link distribution is computed as follows: First, we locate all the anchors appearing on the Web that have the same text as the query, and extract their destination URL's. Afterwards, we count how many times each destination URL appears in this list and sort the destinations in the descending order of their appearance. We then create a histogram where the frequency count in the i^{th} bin is the number of times that the i^{th} destination appears. Finally we normalize the frequency in each bin so that all frequency values add up to 1. Figure 7(a) shows a sample anchor-link distribution for query *pubmed*. (In Section 4 we will provide details about how we collect the anchor data via Web crawling.) The leftmost bar suggests that, 78% of the links with the anchor text *pubmed*, point to the top-ranked destination (www.ncbi.nlm.nih.gov).

For a navigational query, because of the existence of an authoritative answer, we expect the anchor-link distribution to be highly skewed toward rank one (which should correspond to the query's answer). On the other hand, the anchor-link distribution for an informational query should be more flat because of the lack of consensus regarding which Website provides the most authoritative answer. Again, in order to verify this intuition, we show the anchor-link distributions for four sample queries in Figure 7 and Figure 8. We can observe a clear distinction in the skewness of the anchor-link distributions between navigational queries and informational queries. In Section 4 we will experimentally evaluate how effective it is to use the mean, median, Skewness and Kurtosis of the anchor-link distribution in predicting query goals.

A practical concern in applying the anchor-link distribution is *link spams* and *mirror sites*. Sometimes, people create massive number of links to a Website that is not directly relevant to the anchor text in order to gain higher ranking in search results. Also, a Website may be mirrored at

multiple locations and each mirror may have similar numbers of links from other sites. Link spams and mirror sites, therefore, may distort the anchor-link distribution and introduce undesirable noise for our purpose. We did not observe any noticeable noise from link spams or mirror sites for our benchmark queries, but existing techniques for spam and mirror detection [18, 19, 20, 21] may be used to avoid any potential issue.

4. EVALUATING THE EFFECTIVENESS OF THE PROPOSED FEATURES

In the previous section we have proposed several features to predict the goal of a query. In this section we experimentally evaluate the effectiveness of these features using our benchmark query set. In Section 4.1, we describe how we obtain the feature values for the evaluation task. In Section 4.2, we study the effectiveness of our proposed features when they are used individually. In Section 4.3 we show how much the prediction accuracy improves when multiple features are combined. Finally in Section 4.4, we compare the effectiveness of our proposed features with those proposed in previous research.

4.1 Description of dataset

In this section we describe in detail how we select the queries for our study and how we prepare various feature values for each query. As we discussed in Section 2, we use the 30 queries that are *not* software or person-name related for our evaluation due to the ambiguity of the goals of these queries.

Collection of queries and click-through data. As we briefly mentioned in Section 2, our benchmark queries are the 50 most popular queries issued to Google from the UCLA Computer Science Department. In order to obtain these queries and the corresponding click-through behavior, we installed a packet recorder at the central router of our department, which handles all IP packets coming to/leaving from our department. For a period of 6 months (April 2004 till September 2004), this recorder captured the headers of all outbound HTTP requests, from which we could obtain Google queries and the click-through data. During this 6 months, 147,744 unique queries were issued from the department, and each query was issued 1.60 times on average.

In selecting 50 queries for our human subject study, we considered two options: (1) picking 50 random queries and (2) picking the 50 most popular queries. We decided to pick the popular ones, because it is relatively easier for our participants to judge on popular queries issued by many users, instead of some random queries that are issued once or twice by a single person. In addition, to avoid any potential bias introduced from the queries issued by a single user, we picked only the queries that were issued from at least 3 different IP addresses. On average, our 50 benchmark queries were issued by 19.6 IP addresses, with the maximum being from 64 IP addresses ("citeseer").

We can associate the users' click-through data with a particular query issued to Google using the *Referer* field of each HTTP header. Details are omitted for brevity. In our dataset, each of our benchmark queries got an average of 42 user clicks, which was sufficient for our evaluation.

Anchor data. To create the anchor-link distributions for our queries, we crawled 60,824,009 pages from the Web, starting from the Web sites listed in the Open Directory Project.[6]

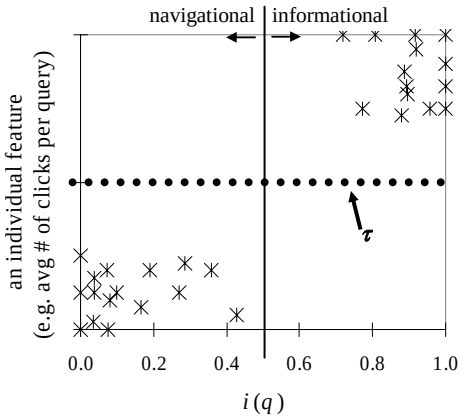


Figure 9: Hypothetical goal-prediction graph for an effective feature

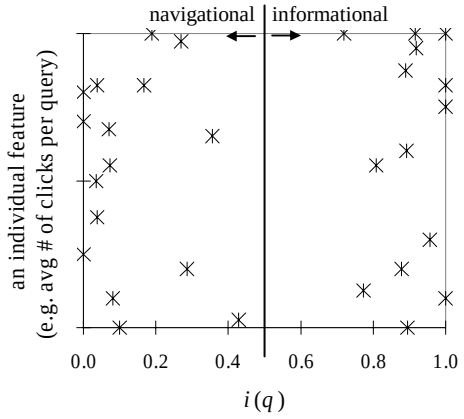


Figure 10: Hypothetical goal-prediction graph for an ineffective feature

After this data collection, we scanned the 60 million Web pages to identify the anchors that have the same text as our benchmark queries. On average, we could find 3,169 matching anchors for each of our benchmark queries. There are about 10 queries that we cannot find sufficient anchors to create their anchor-link distributions. We think this is largely due to our requirement that the anchor texts and the queries must be exact matches. It will be an interesting future work to relax this requirement and use partial matching methods. Currently for these queries we will mainly depend on the user-click behavior data to detect their goals.

4.2 Evaluation of individual features

In this section we investigate the effectiveness of our individual features in predicting the goal of a query.

Goal-prediction graph. To help readers assess the predictive power of individual features, we plot *goal-prediction graphs* in this section. We first explain how we can interpret a goal-prediction graph using two hypothetical graphs in Figure 9 and Figure 10. In the graph, the x-axis is the $i(q)$ value for each query, and the y-axis is the feature value for that query. For our discussion, we assume the feature is the *average number of clicks per query*.

If this feature is effective in predicting the user goal, we

[6] `www.dmoz.org`, which is claimed to be the largest, most comprehensive human-edited directory of the Web.

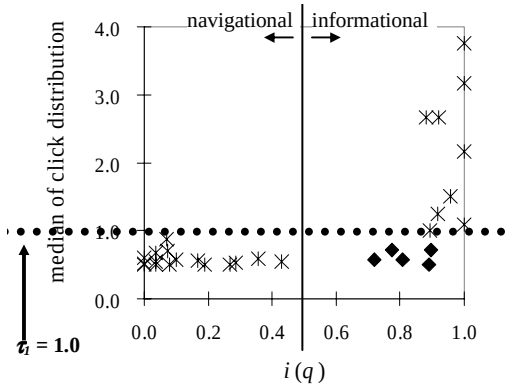


Figure 11: Median of click distribution

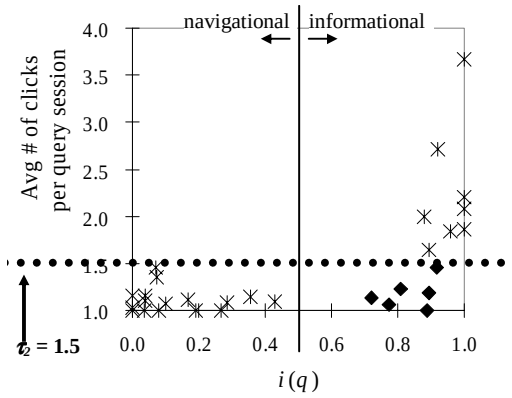


Figure 12: Avg # of clicks per query

expect that its value will be small for navigational queries and large for informational queries. Figure 9 shows the goal-prediction graph when this is the case. In the graph, all navigational queries ($i(q) < 0.5$) have small feature values and, thus, fall into the lower left corner of the graph. In contrast, all informational queries ($i(q) > 0.5$) have large feature values and are clustered around the upper right corner. Given this clear separation between navigational and informational queries, we can predict the goal of a query using the following simple criterion:[7]

$$\text{goal} = \begin{cases} \text{navigational} & \text{if feature value} < \tau \\ \text{informational} & \text{otherwise} \end{cases} \quad (1)$$

where the τ value is selected based on the expected distribution of the feature values for navigational and informational queries.

Given this criterion, we classify all queries below the dotted τ line of Figure 9 as navigational and everything above as informational. Figure 10 shows a goal-prediction graph when the feature is ineffective. Because there is no clear separation between navigational and informational queries, we cannot find a clear threshold value for the goal prediction.

In summary, the goal-prediction graph helps us to visually assess the predictive power of a feature by looking at the separation between navigational and informational queries.

Click distribution. We first compare the effectiveness of the four features based on the user-click distribution: *mean*, *median*, *Skewness* and *Kurtosis*. For comparison, we (1) plot the goal-prediction graphs for the features and (2) perform

[7]If a feature is negatively correlated with $i(q)$, the condition should be reversed.

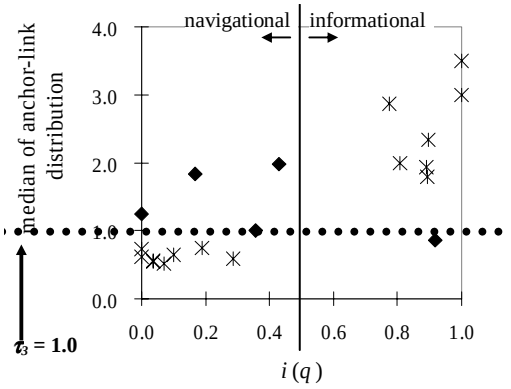


Figure 13: Median of anchor-link distribution

linear-regression analysis [22] to measure the correlation between the $i(q)$ values and the individual feature values. From this comparison, we observe that the three features — mean, median, and Skewness — show similar predictive power; the overall shape of the goal-prediction graphs is very similar, and with reasonable settings for the τ value, all three features show about 80% prediction accuracy.

For example, Figure 11 shows the goal-prediction graph for the median of the distribution. As the threshold value, we use $\tau_1 = 1.0$ based on the following intuition: For most navigational queries, the vast majority of users simply click on the page that they have in mind, so more than 50% of clicks go to the rank-one page. Thus, the median is typically one or less for navigational queries. Under this threshold setting, we get an accuracy of 83.3%; we correctly classify 25 queries (shown as stars in the figure) and misclassify the other 5 (shown as diamonds). Interestingly, we observe that most of the misclassification occurs for informational queries when we use the features based on the user-click distribution.

Average # of clicks per query. In Figure 12, we show the goal-prediction graph for the *average number of clicks per query*. For this feature, we set the threshold value at $\tau_2 = 1.5$ for the following reason: Navigational queries tend to receive only one click in most of the cases, and informational queries typically get more than one. $\tau_2 = 1.5$ is the middle point between the two. Under this setting, the average number of clicks yields an accuracy of 80%. We can see that the predictive power of the number of clicks are almost identical to that of the median shown in Figure 11. The general shape of the graphs is almost identical and misclassification occurs for informational queries.

Anchor-link distribution. We now examine the effectiveness of the anchor-link-distribution-based features. Again, we compare the mean, median, Skewness, and Kurtosis of the distribution using the goal-prediction graph and linear regression analysis and find that the mean, median, and Skewness show similar effectiveness in predicting the user goal; all three show the prediction accuracy of roughly 75%.

As an example, we show the goal-prediction graph for the median in Figure 13. We use the threshold value $\tau_3 = 1.0$ for the same reason discussed before; most of the links point to the single "authoritative" page for the given anchor. The median of the anchor-link distribution yields 75% accuracy.[8]

[8]In Figure 13 we only show 20 queries that have sufficient anchor data to derive their values for this feature.

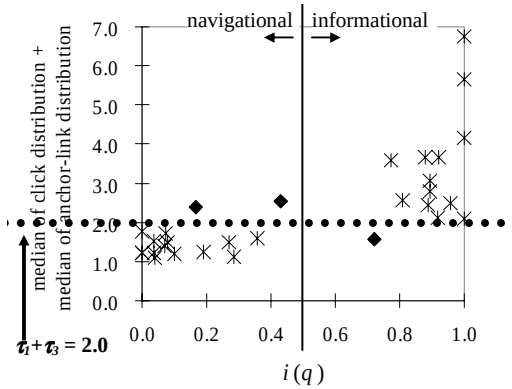


Figure 14: Combining median of click distribution and median of anchor-link distribution

Interestingly, we observe that when we use the features based on the anchor-link distribution, most of the misclassification occurs for navigational queries. For example, in Figure 13, most of the diamonds (misclassification) are in the navigational region.

4.3 Combination of multiple features

In this section we study how much the prediction accuracy improves when we consider multiple features. A number of different methods exist for combining multiple features in making a final decision (e.g., decision-tree method [23], and support vector machine [24]). In our current study, we examine the effectiveness of the following linear combination and defer the study of other methods as future work:

$$f = w_1 \cdot f_1 + w_2 \cdot f_2 + \cdots + w_n \cdot f_n$$

where f_i is the ith feature and w_i is the weight given to the ith feature. Again, we use the goal-prediction graph and the linear-regression analysis to evaluate the effectiveness.

As expected, combining features based on the same information does not increase accuracy. For example, the combination of the median and the Skewness of the click-link distribution results in the same overall accuracy. The accuracy improves only when we combine the features based on different information.

For example, we show the goal-prediction graph for the equal weight combination of the medians of the user-click and the anchor-link distributions in Figure 14. That is,

$$f = (\text{median of click distribution}) + (\text{median of anchor-link distribution})$$

Individually, we use the threshold values $\tau_1 = 1$ and $\tau_3 = 1$, so we use $\tau_1 + \tau_3 = 2$ for the combined threshold. Under this setting, the graph shows the overall accuracy of 90%. Comparing with the accuracy of using each individual feature, this result clearly indicates that combining multiple features is beneficial.

4.4 Comparison with prior work

In this section we compare the effectiveness of our features with the features proposed in a previous study [7]. In this study, Kang and Kim postulated that navigational-query terms appear more often in the anchor text and on the home pages of Web sites, compared to informational-query terms. Based on these hypotheses, they proposed the following features for automatic goal prediction:

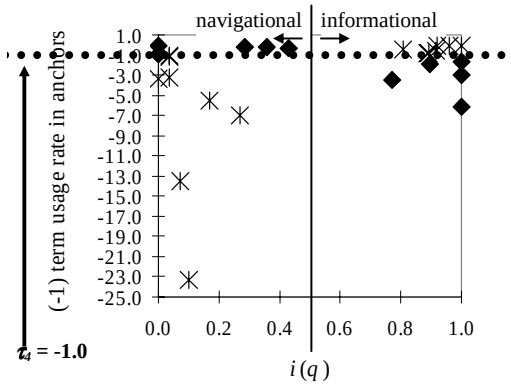


Figure 15: Anchor usage rate

- **Anchor usage rate.**[9] From a collection of pages downloaded from the Web, we count how many times the terms in each query appear in the anchor text and in the overall document collection. If the terms appear more often in the anchor text, the query is considered navigational.
- **Query term distribution.** We partition the set of downloaded Web pages into two collections: the *homepage collection* and the *content-page collection*. The homepage collection consists of the homepages of Web sites (i.e., Webpages with a root URL such as http://www.bestbuy.com). All other pages belong to the content-page collection. Given a query, we compare how many times the terms in the query appear in each collection. If the terms appear more frequently in the homepage collection, the query is considered navigational.
- **Term dependence.** This feature can only be applied to multi-term queries. The hypothesis of this feature is that if the co-occurrence of multiple terms in a particular query show more dependence in the homepage collection than in the content-page collection, the query is more likely to be navigational. The authors use *mutual information* to measure the dependence.

To evaluate the effectiveness of these three features, we build the homepage and the content-page collection from the 60 million pages that we downloaded from the Web, following the guideline provided in [7]. Using these collections, we compute the feature values for our 50 benchmark queries and plot the goal-prediction graphs in Figures 15 through 17. For all three features, the graphs do not show clear separation between the navigational and the informational queries. The highest accuracy is 60% when we use the anchor usage rate with the threshold value $\tau_4 = -1.0$.[10]

We also compare the effectiveness of the three features in predicting the $i(q)$ value using the linear-regression analysis. More precisely, we model relationship between a feature value x and $i(q)$ as

$$i(q) = \beta_0 + \beta_1 \times x.$$

Under this model, if a feature x predicts the $i(q)$ value well, then $\beta_1 \neq 0$. Thus, we make the null hypothesis that $\beta_1 = 0$ and validate this hypothesis by computing the p-value of

[9] The exact formulas of their features are quite complex. We only provide a high-level intuition of their proposed features.

[10] Since the anchor usage rate feature is negatively correlated with $i(q)$, in plotting the goal-prediction graph we flip the sign for this feature.

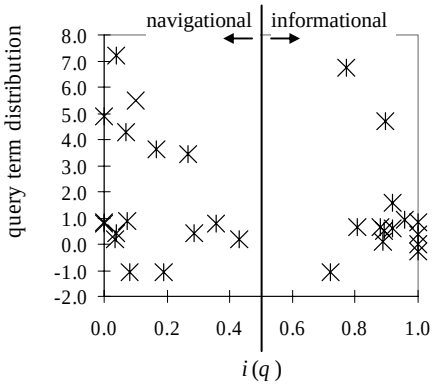


Figure 16: Query term distribution

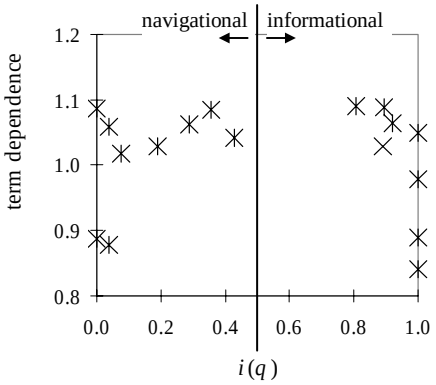


Figure 17: Term dependence

each feature. As a common practice, the null hypothesis is rejected when p-value<0.05 [25, 22], which indicates that the feature is effective in predicting the $i(q)$ value.

Figure 18 shows the result of this regression study. In the table, we also show the results for two of our proposed features for comparison. The result suggests that the three features proposed in [7] may not be very effective in predicting the user goal; the null hypothesis is accepted for all three features, indicating that they do not show strong correlation with the $i(q)$ value.

Given these results, we further investigate why the three features proposed in [7] are not very effective by manually looking at the feature values for some of our benchmark queries. We briefly summarize our main findings as follows:

- Query term distribution and term dependence are two similar features that rely on the *difference* of the query term distributions between the the homepage collection and the content-page collection. However, we find that the navigational and the informational queries in our benchmark do not exhibit consistent difference in our collection. For example, a clearly navigational query *ucla library* (whose $i(q) = 0$) appears more frequently in the content set (0.025% of the documents) than in the homepage set (0.015% of the documents), yet the situation is reversed for another navigational query *bestbuy* (which appears in 0.021% of the homepage set and in 0.0054% of the content set). For the term-dependence feature, we observe that, in most cases, terms in a query are as independent in the homepage set as they are in the content set, regardless of whether the query is navigational or informational (as shown in Figure 17).

Feature	p-value	Response to the null hypothesis $\beta_1 = 0$ (significance level 0.05)
query term distribution	0.2900	Accept
term dependence	0.6520	Accept
anchor usage rate	0.3078	Accept
median of click distribution	0.0141	Reject
median of anchor-link distribution	0.0001	Reject

Figure 18: Results of simple linear regression

- The anchor-usage rate assumes that the terms of navigational queries appear more often in anchors than in Web pages. We observe a number of instances for which this assumption seems invalid. For example, for informational queries *hidden markov model* and *simulated annealing*, they appear 2.9 and 6.1 times *more often in anchors* than in Web pages, respectively. The ratio for a navigational query *bestbuy* is 3.3, which is smaller than that of *simulated annealing*.

5. RELATED WORK

There is a large body of work on Web user's searching behavior and Web query statistics [26, 16, 27, 28]. A rather comprehensive review of such studies can be found in [29]. These studies are mainly concerned about the *general characteristics* of Web queries, while our concern is to learn *the goal* behind a Web query and identify the goal automatically.

Our work is inspired by recent studies by Broder [6], and Rose and Levinson [8] on Web query goals. By manually inspecting search engine query logs, the researchers have found that the query goals belong to a few categories such as navigational, informational, resource or transactional. They have further reported the percentage of Web queries that belong to each category from the manual inspection process.

To the best of our knowledge, the work by Kang and Kim [7] is the only published work on automatic identification of query goals. In that paper they proposed to explore the occurrence patterns of query terms in Web pages in order to detect the goal of a query as either navigational or informational. As we have shown in Section 4.4, we believe our proposed features are much more effective than the term-occurrence-pattern-based features.

In [2, 3] researchers have demonstrated that it is feasible to improve search engines' performance by applying specialized ranking mechanisms for navigational and informational queries. In the studies, the researchers assume that the queries' goals are already given. Our study can be beneficial to this thread of work by providing an automatic mechanism to predict the goal of a user.

Our study is also related to recent research on analyzing users' clicking behavior after they issue a Web query [30, 31]. The main focus in these works is to detect *similar Web queries* based on the similarity of user-click behavior for these queries. In [32], Kraft and Zien have also analyzed anchor texts for the purpose of *Web query refinement*. This work is based on the observation that Web queries and anchor texts are highly similar [33], and additional terms appearing in anchor texts are good candidates to append to the original query and to make the search more specialized.

6. CONCLUSION

In this paper we studied the automatic identification of a user goal for a Web query. Through a human subject study, we first showed that about 60% of the queries we

studied have "predictable" goals independent of users. This study further suggested that for the other 40% of the queries with less predictable goals, a search engine may be able to employ simple techniques to detect and handle them separately. We then proposed two categories of effective features in identifying the goal of a query: past user-click behavior and anchor-link distribution. Our experimental evaluation showed that using a combination of the proposed features we can correctly identify the goals for 90% of the queries studied. We also experimentally compared our proposed features with those investigated in previous research. Our results showed that our features clearly outperformed the existing features.

One limitation of our study is that our experiment was conducted on a potentially-biased dataset: queries from the CS department may show a technical bias and are likely to be well crafted and potentially work related. Therefore, some of the characteristics that we observed may not be true of user queries in general. While we believe our two features will be effective for predicting user goals even for general queries, it will be interesting to see how some of our observations may change for a larger dataset.

7. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0347993. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] D. Hawking and N. Craswell. Overview of the TREC-2001 Web track. In *Proceedings of the Tenth Text REtrieval Conference (TREC-10)*, 2001.

[2] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR '01*, 2001.

[3] T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving web pages using content, links, URLs and anchors. In *Proceedings of the Tenth Text REtrieval Conference (TREC-10)*, 2001.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh Int'l. World Wide Web Conf.*, 1998.

[5] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.

[6] A. Broder. A taxonomy of Web search. *SIGIR Forum*, 36(2), 2002.

[7] I. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of ACM SIGIR '03*, 2003.

[8] D.E. Rose and D. Levinson. Understanding user goals in Web search. In *Proceedings of the Thirteenth Int'l. World Wide Web Conf.*, 2004.

[9] H. Zeng, Q. He, Z. Chen, W. Ma, and J. Ma. Learning to cluster web search results. In *Proceedings of ACM SIGIR '04*, 2004.

[10] Oren Zamir and Oren Etzioni. Grouper: a dynamic clustering interface to web search results. In *Proceedings of the Eighth Int'l. World Wide Web Conf.*, 1999.

[11] Vivisimo search engine. http://vivisimo.com/.

[12] C. Olston and E. H. Chi. Scenttrails: Integrating browsing and searching on the world wide web. *ACM Transactions on Computer-Human Interaction*, 10(3):177–197, September 2003.

[13] M. Chen, M. Hearst, J. Hong, and J. Lin. Cha-Cha: A system for organizing intranet search results. In *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems*, 1999.

[14] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. Technical report, UCLA Computer Science, 2004.

[15] A. Sugiura and O. Etzioni. Query routing for web search engines: Architecture and experiments. In *Proceedings of the Ninth Int'l. World Wide Web Conf.*, 2000.

[16] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large Web search engine query log. *SIGIR Forum*, 33(1):6 – 12, 1999.

[17] Danny Sullivan. Searches per day. http://searchenginewatch.com/reports/article.php/2156461, 2003.

[18] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proceedings of VLDB '04*, 2004.

[19] T. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the Eleventh Int'l. World Wide Web Conf.*, 2002.

[20] J. Cho, N. Shivakumar, and H. Garcia-Molina. Finding replicated web collections. In *Proceedings of ACM SIGMOD '00*, 2000.

[21] K. Bharat and A. Broder. Mirror, mirror, on the Web: A study of host pairs with replicated content. In *Proceedings of the Eighth Int'l. World Wide Web Conf.*, 1999.

[22] J.L. Devore. *Probability and Statistics for Engineering and the Sciences*. Duxbury, 6th edition, 2004.

[23] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[24] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[25] D.D. Wackerly, W. Mendenhall III, and R.L. Scheaffer. *Mathematical Statistics with Applications*. Duxbury, 6th edition, 2002.

[26] C. Hoelscher. How Internet experts search for information on the Web. In *Proceedings of WebNet '98*, 1998.

[27] B.J. Jansen, A. Spink, and T Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2):207 – 227, 2000.

[28] A. Spink, B.J. Jansen, D. Wolfram, and T. Saracevic. From E-Sex to E-Commerce: Web search changes. *IEEE Computer*, 35(3):107 – 109, 2002.

[29] B.J. Jansen and U. Pooch. A review of Web searching studies and a framework for future research. *J. of the American Society of Information Science and Technology*, 52(3):235 – 246, 2001.

[30] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of ACM SIGKDD '00*, 2000.

[31] B.D. Davison, D.G. Deschenes, and D.B. Lewanda. Finding relevant Website queries. In *Proceedings of the Twelfth Int'l. World Wide Web Conf.*, 2003.

[32] R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceedings of the Thirteenth Int'l. World Wide Web Conf.*, 2004.

[33] N. Eiron and K.S. McCurley. Analysis of anchor text for Web search. In *Proceedings of ACM SIGIR '03*, 2003.