Functional Faceted Web Query Analysis

Bang Viet Nguyen
Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543
nguyenv3@comp.nus.edu.sg

Min-Yen Kan
Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543
kanmy@comp.nus.edu.sg

ABSTRACT

We propose a faceted classification scheme for web queries. Unlike previous work, our functional scheme ties its classification to actionable strategies for search engines to take. Our scheme consists of four facets of ambiguity, authority sensitivity, temporal sensitivity and spatial sensitivity. We hypothesize that the classification of queries into such facets yields insight on user intent and information needs. To validate our classification scheme, we asked users to annotate queries with respect to our facets and obtained high agreement. We also assess the coverage of our faceted classification on a random sample of queries from logs. Finally, we discuss the algorithmic approaches we take in our current work to automate such faceted classification.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Human Factors, Standardization, Measurement

Keywords

query classification, faceted classification, authority, temporal and spatial sensitivity, ambiguity, query analysis, search engines

1. INTRODUCTION

While information retrieval methods have provided algorithms that can scale beautifully to handling large volumes of documents and queries, our fundamental understanding of web queries remains quite primitive. In recent years, macroscopic studies of query logs have measured changes in average query length, query topic distribution and session lengths. While these statistics play an integral part in tuning search engine performance, such macroscopic studies alone have not provided insight into the information seeking process of web search engine users. On the other hand, microscopic studies have addressed this weakness by characterizing how users seek information on the web: what are users' intents, how do they modify queries and what types of hyperlinks do they follow. Microscopic analysis is necessarily descriptive and sheds light on the intrinsic nature of user queries. However, such descriptive studies may not translate to algorithmic strategies in handling queries. In order to build reliable implementations of query classification, researchers have needed to simplify or throw away parts of such detailed microscopic analyses to create coarse-grained, operational categories.

Such mismatches between macroscopic and microscopic analyses are problematic. We believe that the middle ground between these two types of analyses for web queries is currently missing. Such a study should tie together previous microscopic analyses on query analysis and information seeking in the context of the web. However, an ideal classification should also enable to do something useful with resulting classified queries - as modeling classification is often done. In the context of web query analysis, this means being able to act accordingly based on query classification and deciding on what information is relevant to the users' goals and ongoing information seeking process. For example, navigation queries that aim to locate a particular web site are best served by making it easy to jump directly to the site.

A key difference of this work from previous query analyses is to provide a purpose-driven functional classification unlike previous schemes that sought to understand user queries intrinsically, without posterior processes in mind. A contribution of this work is to propose a classification scheme that is specifically tied to actionable strategies that search engine can take when a query is classified.

Another aspect of our work is that we propose a faceted classification, rather than mutually exclusive categories as was given by earlier work. This was a crucial insight for us, modeled after relevant work in the library and information science (LIS) literature on reference query analysis and information needs. Rather than focusing on user goals directly, we propose a classification scheme that consists of multiple facets instead. Our scheme consists of four facets: ambiguity, authority sensitivity, temporal sensitivity and spatial sensitivity. A query is classified on the basis of each facet separately, and its facet classification invokes a specific handling actions.

In our scheme, user intentions are not modeled directly; rather they are inferable from the inventory of facet classifications. From a modeling standpoint, our approach can be seen as inserting an abstract layer between the query's surface form and the user intent. That is, instead of inferring the user's intent directly, we model the query as a collection of facets that then informs us of the intent. To connect our work with previous work, we have cast previous web query taxonomies and their example queries into our framework.

WWW2007, May 8–12, 2007, Banff, Canada.
.

For example, navigational queries are characterized by one particular setting of our facets.

The benefit of such query analysis is that query handling is modeled as primitives that correspond to facets, rather than to high-level user goals directly. For example, knowing that a user is trying to learn a concept (a high-level user intent) may not help change how a search engine deals with such a query, but knowing that a query is ambiguous (a facet) may help. For example, query expansion or relevance feedback might be appropriate actions for ambiguous queries.

Finally, to ensure that our work has a broad impact, we created our classification by limiting our analysis to simple query logs, without sequential, timestamp or session information. While other streams of information (e.g., clickthrough data, distribution of web page results) are certainly helpful in query classification, our approach is to extend to use of simple data to construct and replicate manual classification queries..

We introduce our faceted scheme in the next section and discuss its relation to related work in Section 3. In Section 4, we report on our manual query log mining results and the results of our human subject study to validate the stability of our proposed faceted classification scheme in Section 5. In this section, we discuss our findings on the actual distribution of web queries according to this scheme as well as the replicatabiliity of each facet. While not the focus of this paper, we sketch out algorithms for handling each facet and our ongoing work on creating automated methods for facet classification in Section 6.

2. FACETED QUERY CLASSIFICATION

In our scheme, queries are classified with respect to each facet. We first describe the four facets of our classification and then justify its construction.

- *Ambiguity* (Am): {Polysemous, General, Specific} Queries exhibit different levels of ambiguity [16], just as in natural language. We say a query is polysemous if there is more than one strong sense of a query in retrieving results. Single word homophones, acronyms or common person names are polysemous under this definition (e.g., "mustang", "ui", "j brown"). General queries have only one strong sense but relevant information may be scattered among various subcategories that do not correlate with each other. Examples include queries such as "health" (which correspond to distinct sub-categories like health industry, health advice, health experts), or "travel" (travel tips, a good list of travel places, travel agents). Specific queries are all other queries; these queries have a coherent set of relevant information to address it. Product queries such as "apple iphone" are thus examples of specific queries, as information on specifications, product availability and price are usually found in specific categories of web pages.

- *Authority Sensitivity* (Au): {Yes, No} Queries that require an authoritative, trusted answer are marked as sensitive to authority. Factoid questions in TREC [17] are examples of queries that are sensitive to authority. Certain queries may also implicitly require results from authoritative sites in which the site sought must be free of malicious intent (e.g., virus scan software - we don't want to download such software from disreputable companies or sites). In this respect, all navigational queries are ones that require authoritative results, as their intention is to go directly to the genuine website of interest.

- *Temporal Sensitivity* (T): {Yes, No} Queries marked as temporally sensitive should return different results when issued at different times, i.e. they need to take into account the time when the query was issued. The sensitivity may vary with the scope of the time period (vary by year, month, day). Thus "us president" would be temporally sensitive versus "us president 1956" would not, as the latter query has indicated a specific time period that makes it unambiguous.

  Note that also some queries are faddish: e.g., "tamagochi", "star wars kid". These queries describe some event or entity that was highlighted in current events. By default, we do not consider these temporally sensitive, since the results of such searches are stable after an initial period of fame.

- *Spatial Sensitivity* (S): {Yes, No} Queries that are spatially sensitive should return results that account for the user's location. Queries may be sensitive to the user's general region (city, country) or be sensitive to the user's exact location in the case of queries from mobile clients (e.g., "pizza", "nearest hotel"), relating more to commodities and services in M-commerce.

To create our classification, we used randomly sampled portions of the query log from the AllTheWeb European search engine from May 28^{th}, 2002, which was released to the research community by Jim Jansen. In creating our classification, we encountered several difficulties which we would like to discuss. Our goal was to create a classification that was usable, general and connected to previous work. These considerations greatly impacted the final design. As we discuss the connection with previous work in the next section, we detail the two other considerations here.

With respect to usability, we chose to limit each facet's values to a small set of discrete values. Although it can be argued that each facet should be graded along a continuum, continuous values are often hard to analyze, replicate or analyze from a manual standpoint. We believe that using discrete values makes our classification simple and more usable and does not forsake much representational power.

With respect to generality, we wanted our classification to be applicable to as many general types of web queries as possible. One consequence of this was that we restricted ourselves to query types distinguishable just from examining the query string itself.

We also note that the proposed facets are not exhaustive. Other interesting facets that we considered were personalized queries and subjective/objective queries. However, these were difficult to distinguish without using session data and examining results of the searches, and were thus discarded. Our query facets are also not independent - many of the facet combinations are naturally mutually exclusive. For example, a polysemous query may have several possible underlying interpretations, making it impossible to determine sensitivity to the remaining facets. Despite these shortcomings, we believe our classification is a step forward in web query analysis.

<Am=Specific, Au=No, T=Yes, S=No>
new lingerie fashion
top ten search engine ranking
french open
macintosh news
eurovision contest
<Am=Specific, Au=Yes, T=Yes, S=No>
tennessee birth records
onspec drivers download
air messenger pro download
download grand theft auto
world no tobacco day poster
download accellerator plus

Table 1: Sample queries from two bins. Most software/resource seeking queries fit into the second bin <Am=Specific, Au=Yes, T=Yes, S=No>.

As all facets are discrete, there is a limited set of $3 \times 2 \times 2 \times 2 = 24$ possible facet combinations to which we can assign queries. We term each such combination a *bin*. For example, one bin contains all the queries that possess the combination of facet settings *Ambiguity*=Specific, *Authority sensitivity*=No, *Temporal sensitivity*=Yes, *Spatial sensitivity*=No. For convenience we refer to bins by an ordered 4-tuple <Am, Au, T, S>, such that the example above is identified as <Specific, No, Yes, Yes>. As the facets are not entirely independent of each other, some bins are empty; queries that are polysemous are grouped into one single bin. In Table 1 we present two examples of such queries from the bins <Specific, No, Yes, No> and <Specific, Yes, Yes, No>.

2.1 Actions for facets

What type of actions should be prescribed given facet classification? Here, we give some quick algorithmic sketches and rationale; a detailed implementation is beyond the scope of this work. By default, *specific* and non-sensitive queries should not alter a search engine's best effort retrieval strategy. When facets display other values, we can then invoke an appropriate action:

- *General* and *polysemous* queries as defined in our *Ambiguity* facet are too general and underspecified. An appropriate disambiguation or query expansion should be used, either in a relevance feedback framework (in which the user chooses between alternatives)
- Queries that are *authority sensitive* need information from trusted sources. In these cases, we may favor sites that are known to be professionally edited or peer reviewed; or weight ranking to rely more heavily on graphical prestige favors (e.g., PageRank) at the expense of content based similarity.
- *Temporally sensitive* queries should seek to use recently indexed data, where the window of recency could be inferred from the timeframe of the temporal sensitivity (e.g., most recent year, month, date). With RSS, highly sensitive topics can be supported using just-built indices, with content pushed from publishing sites.
- *Spatially sensitive* queries, like temporal ones, would need to reweight results to favor resources highly ranked for the locality of the user.

3. RELATED WORK

Primitive research have attempted to understand web query features as well as users' search behaviors through transaction log analysis. Jansen *et al.* [8] (Excite study), Silverstein *et al.* [15] (Altavista study) reported various statistics on web queries using the respective search engine query logs, including number of terms per query, number of queries per user, etc. A complete review for other primary web search studies is presented in [7].

The intents of users behind web search queries are diverse and complex. Several key works that deal specifically with web queries logs have explored this issue. Broder's seminal work [1] revealed that web users' goals could be categorized into classes unlike general "information need". Broder proposed the first taxonomy for web search queries according to the underlying goals: Navigational for those queries which are intended to look up a specific web site that the user has in mind; Informational for those looking for any information on a topic; Transactional for those looking for services provided by any webpage (for example, downloading music). Rose and Levinson [14] furthered Broder's research, further refining Broder's sub-categories and broadening Broder's transactional query class with a more encompassing resource query class to include viewing, downloading and obtaining resources available on the Web. The paper also tried to determine which classification techniques are effective for this taxonomy. They analyzed some example queries, concluding that classification is feasible by examining query logs and clickthrough data. They also noted that ambiguous cases exist, that would be difficult to classify.

Initially, we wanted to refine the high-level web query typology of informational / navigational / resource popularized by Rose and Levinson [14]. However, we had a difficult time finding concrete subclasses that could be tied to different retrieval engine actions - this is in contrast to the hierarchical classification in their work that categorizes for user intent.

The careful reader will note that the facets we have proposed have been discussed previously in the literature in one form or another. Our contribution is thus to formalize and collect these aspects into one typology.

Ambiguity of web queries has been extensively studied, in particular research on query disambiguation. The understanding of ambiguity is thus very diverse. In particular, we note the work [16] which categorized distinct dimensions of ambiguity, modeled after a library agent interpretation of ambiguity (to address an Anomalous State of Knowledge – ASK). This categorization of ambiguity has a sufficient level of abstraction to describe our ambiguity facet.

Temporal sensitivity has been studied in several works: [2] analyzed temporal factor of web queries to find correlations between two given web queries. [5] discussed faddish characteristics of blogspace. They report on the faddish query phenomenon "getting intensely fashionable only for a short time" with respect to web logs (blog). Such a definition can be extended to certain classes of web queries, such as ones that are news related. [4] dealt specifically with the spatial facet, proposing "local" versus "global" location scopes, where a local scope is equivalent to spatial sensitivity and global scope to insensitivity. Query internal features, extensive gazetteer information and some distribution features are

all integrated as a supervised classification problem. Spatial sensitivity is further examined in [18], in which the spatial facet is characterized as a dominant location, defined subjectively as the location where most individual would know the answer to the query. Here, multiple live search results are analyzed with appropriate multiword tokenization to yield data for the classification decision.

To the best of our knowledge, authority sensitivity of web queries has not been previously investigated. The closest relevant work concerns the retrieval of reliable answers [13] and trusted question answering [12]. The latter study used knowledge from search engines to find reliable answers for factoid questions. However, both works did not investigate this attribute for web search queries.

3.1 Automatic classification

Once a holistic classification of web queries was published, attempts to automate the classification of queries were quick to follow. Kang and Kim [10] presented a comprehensive solution for classifying queries which is based on the taxonomy by Broder [1]. They used various features from the queries themselves (for example, part of speech) and from other collected data relating to the queries. They divided a set of collected web pages (WT10g) into 2 sets: DBTopic and DBHome. This division is based on the URL feature to decide whether a webpage is a site entry or not. Based on this set of data, various features are extracted, including the distribution of query terms, their mutual information and usage rate in hyperlink anchor text. The authors concluded that there are still many shortcomings with their approach in classifying queries. Lee *et al.* [11] built upon this work, conducting a human subject study to assess the feasibility of classifying query goals, using the resulting data for benchmarking later. Their work also introduced two new features for automatic classification: click distribution and anchor link distribution. Their work resulted in the current state-of-the-art accuracy of 90% for query classification between navigational and informational query classes.

Query classification is a prominent field outside of web query log analysis as well. In library science, Kan and Poo [9] examined the classification of known-item queries (queries that attempt to retrieve items seen or known before to the user; somewhat analogous to navigational queries) used query length, part of speech and language modeling combined using machine learning. In the information retrieval community, a central issue in question answering (which seeks an exact answer to natural language queries, e.g., "Katie Holmes" for "Who is Tom Cruise married to?") is in query typing. Here, classification of questions help the IR system pinpoint the type of answer to be expected (e.g., both "who" and "married to" indicate that a PERSON is an expected answer). Fine-grained classification of answer types and their correlated word patterns have been publicly released [6], pointing the way forward for future, more detailed query classification. A key difference in question answering is that it deals largely with well-formed natural language questions, unlike the typical keyword searches found in web query logs.

4. CLASSIFICATION DEMOGRAPHICS

A good query classification will cover most queries encountered and be able to differentiate queries with respect to its goals. To assess how well our classification covers actual queries, we sampled query logs and attempted to use our faceted scheme to classify these queries. For this purpose, we randomly sampled 100 new queries from another portion of the same AllTheWeb log. We organize the sampled queries into bins, where each bin is a combination of the values of the four facets as previously discussed. Non-English and sexually oriented queries were replaced by other random samples, in a similar manner to the work [1].

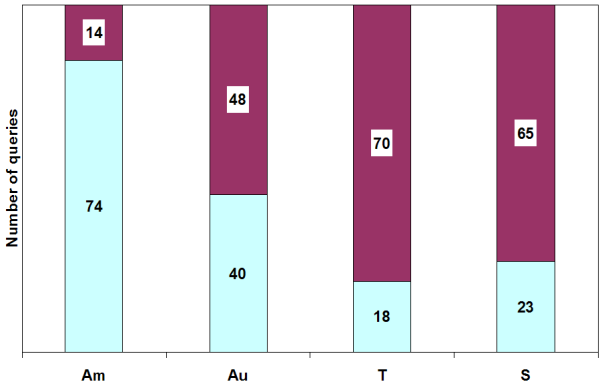

Figure 1: Distribution of queries per facet (polysemous queries were removed). For *Ambiguity*, the light blue portion represent *Specific* queries; dark purple, *General* queries. For *Authority*, *Temporal*, *Spatial*, light blue represents *Sensitive* queries; dark purple, *Not sensitive* queries.

Ambiguity	*Authority*	*Temporal*	*Spatial*	#
Polysemous	-	-	-	12
General	-	Yes	Yes	3
General	-	Yes	No	1
General	-	No	Yes	3
General	-	No	No	6
Specific	Yes	Yes	Yes	2
Specific	Yes	Yes	No	6
Specific	Yes	No	Yes	3
Specific	Yes	No	No	28
Specific	No	Yes	Yes	1
Specific	No	Yes	No	8
Specific	No	No	Yes	9
Specific	No	No	No	18

Table 2: Distribution of 100 Manually Sampled Queries

Table 2 gives the demographic results from our sampling. While only indicative (as we only sampled 100 queries), we observe that most queries are adequately defined (i.e., are specific) such that the expected search result is guessable. Such queries correspond to queries for which we can hope to glean user intent. On the other end of the scale, ambiguous (i.e., *general* in our definition) queries still make up over $1/10^{th}$ of the sampled queries. While this may be disheartening when interpreted as the result of poor formulation of information need, it should be noted that many other factors may be partially responsible: users may be following the same search trail from a previous session, or the query may be disambiguated based on other data unaccounted for

in our study (e.g., personalization or session data). Temporally and spatially sensitive queries account for about one fifth of the samples each, motivating why localized search is a high priority for commercial search products. Also, when queries are not polysemous, a large portion (40%) requires authoritative answers in retrieving results. In particular, among specific queries that were neither temporally nor spatially sensitive, more queries required answers or web pages from authoritative sites.

4.1 Mapping Facets to Rose and Levinson's work

How do our bins relate to classifications defined in other previous work? We take the detailed hierarchy and sample queries of Rose and Levinson as a representative standard, reproduced in brief in Table 3. Note that previous classifications only deal with well-formed, unambiguous queries, and do not acknowledge the problem of ambiguity as part of their classification schemes; we cannot infer user goals given polysemous or ambiguous queries. In contrast our classification explicitly enumerates this possibility, making it quantifiable and a target for automated classification.

We discover that certain classes of queries map easily to particular bins in our taxonomy. For examples, most of their query classes map to our two most general bins: <Specific, -, No, No>. Here, a "-" in a facet indicates any value is acceptable. Navigational and resource queries often have the additional requirement of needing an authoritative answer <Specific, Yes, No, No>. Specific classes of resource queries may further specialize for temporal sensitivity, for example, software related queries fit in the bin <Specific, Yes, Yes, No>, where the Yes setting for T is needed to account for newer or different releases and versions. Open-ended informational questions (also called reference or research questions in LIS) often do not require authoritative answers, while closed-ended questions (factoid questions as termed by IR literature) often do. For such closed-ended informational queries, a single website may contain all the necessary information to answer the query, thus the authority of the website is more important; answers to open-ended questions sometimes can only be synthesized after viewing many sources, hence the dependency on a single authoritative website is diminished.

5. INTERANNOTATOR RELIABILITY

A good classification should also lead to reproducible results among independent subjects. We conducted a human subject study to validate whether this characteristic holds for our proposed faceted classification. Our goal was to assess whether human subjects could label queries in a consistent manner given the facet definitions and labeled examples. We want to know how strongly evaluators agree on the taxonomy as a whole and on with respect to individual facets. We describe our experiment by first discussing how we chose queries to ask human evaluators to judge, and then discuss other experimental details.

5.1 Query Sampling

From our internal demographic survey, we see that queries are unevenly distributed among bins in our faceted classification scheme. Thus a random sample of queries would likely overrepresent some bins and underrepresent others. To ensure our classification is replicable with respect to all facets, it is more favorable to ensure an even distribution of queries representing bins. In this way, we can allocate equal effort in judging each facet.

Category	<Am, Au, S, T>
1. Navigational	<Specific, Y, N, N>
2. Informational	
2.1 Directed	
2.1.1 Closed	<Specific, Y, -, ->
2.1.2 Open	<Specific, N, -, ->
2.2 Undirected	<-, N, -, ->
2.3 Advice	<Specific, -, -, ->
2.4 Locate	<Specific, Y, -, Y>
2.5 List	<Specific, -, -, ->
3. Resource	
3.1 Download	<Specific, Y, Y, N>
3.2 Entertainment	<-, -, -, ->
3.3 Interact	<Specific, Y, -, ->
3.4 Obtain	<Specific, Y, -, ->

Table 3: Search Goal Hierarchy from Rose and Levinson's taxonomy mapped to possible bins from our faceted classification.

To ensure this, we selected queries in such a way that all values of the four facets are equally represented. For example, half of the non-polysemous queries were chosen to be temporally sensitive, while the other half were not. We selected a total of 75 queries to test, which we subdivided into five sets of 15 queries. In each set, values for all facets were equally represented, as above.

5.2 Experimental Details

25 volunteer subjects were then recruited from our institution by mass email that gave the high level details of the experiment. Subjects were all regular users of web search engines but all were not aware of or affiliated with this research or research on query analysis in general. Each subject was given a layman's definitions of the facets and examples of each value on a printout for reference, and then randomly matched to one of the five query sets. As such, we had five subjects annotating each query set.

After reading the proper instructions, subjects were directed to a web interface [1] where an individual query was shown (again from the raw query logs, no additional information was provided) and asked to answer questions about each facet. For each facet except Ambiguity, we asked the evaluators to rate on a scale from 1 (not sensitive) to 5 (sensitive). For the Ambiguity facet, we asked the evaluators to first choose between polysemous or not polysemous. If subjects chose not polysemous, they were further asked to rate the query on another Likert scale from 1 (specific) to 5 (general). When all judgments for a query are completed, the next query from the set is shown, until the subject completed all 15 queries. Upon completion of the survey, the study completion time was noted, and subjects were paid a token amount for their participation.

Note that in our study we opted to use a graded scale, rather than the minimalist binary/trinary scales as defined in our classification scheme. This was for two reasons. First, a graded scale gives more freedom for the evaluator to assign scores. A graded scale with a midpoint can capture

[1] http://wing.comp.nus.edu.sg/facetedQueries/

Facet	F-ratio	P-value
Ambiguity	0.4297	0.7871
Temporal	1.9353	0.1039
Spatial	0.7223	0.5770
Authority	4.7070	0.0010

Table 4: Variance in evaluators' judgments.

Facet	# Bipolar	# Non-bipolar
Ambiguity	63	4
Authority	57	10
Temporal	64	3
Spatial	65	2

Table 5: Distribution of queries as (bi)polar or continuous, per facet. Polysemous queries were removed.

when queries do not correlate with the target query; in contrast, the minimalist scale may artificially force subjects to assign ratings. Aside from testing for agreement and correlation, we wanted to see whether subjects agreed that our facets were bipolar. Second, the data collection can serve as training data for subsequent studies on automatic query classification, and here again, a rich representation of the data is valuable. The raw data collected by our study is also available from the survey's web interface.

5.3 Results

Ratings for each facet of each query were tabulated and analyzed. We have two objectives for our study. First, we want to measure the inter-annotator agreement between evaluators to assess the replicability of our scheme as well as the reliability of the collected data. Second, we want to assess whether subjects felt that the facets were bipolar, thus justifying our minimalist discrete values in our scheme.

For the first objective, we used ANOVA to measure the variance between users' judgments on the query set. ANOVA measures whether observations from multiple groups statistically differ. Table 4 shows the results. An F-ratio of 1.0 is perfect agreement, leading us to accept the null hypothesis (i.e., there is no variance between users' judgments). The P-value gives the probability that the variance happens by chance (random error), which tells us how reliable the statistics are: the higher P-value is (normally a threshold of ≤ 0.05 is chosen), the more likely that the variance is by chance and thus not significant.

The table shows that the F-values for *ambiguity*, *temporal sensitivity* and *spatial sensitivity* have low variance. This is further validated by the high P-values, showing that our evaluators do not significantly disagree with each other. Of these facets, *temporal sensitivity* shows the most variation, which may be caused by the ambiguity of the query itself (e.g., cases such as "movie press releases" where the averaged rating is 2.6, come close to the midpoint value of 3.0 between "sensitive" and "not sensitive").

A problematic facet is the *authority sensitivity*. Since the variance of evaluators' judgments also varies between queries (some queries had more polar judgments), this also suggest that we examine the *authority sensitivity* facet more selectively (perhaps by examining each query separately) should we use the data in future computational tasks.

Next, we report the query classification results judged by the evaluators. In Table 6 and 7, we illustrate the top distinctive (i.e., polar) queries for each end of the facets. The queries are sorted in decreasing order of value. We note that the more distinctive queries (as noted internally in our manual sampling) are also quite well replicated in evaluators' judgments.

We are interested to know whether subjects felt that facets had more of a bipolar nature than a continuous one. To evaluate this, we took the original ratings on the 5-point Likert scale and mapped to new values based on the formula $abs(3-x)$. This folds the judgments along the midpoint of the scale. Then, average scores close to 2 indicate a (bi)polar rating, scores close to 0 indicate a more continuous nature. Table 5 shows the demographics based on this assessment. If a query had an average judgment on this remapped scale below 1.0, it was considered non-polar, and polar otherwise. We can see that most queries are indeed judged as polar by subjects, with *authority sensitivity* having slightly less consensus.

6. AUTOMATIC FACET CLASSIFICATION

We briefly discuss the feasibility and possible approaches to automate the classification of the facets. In particular, because the other facets in our classification have been discussed by other works, we focus on our new facet of *authority sensitivity* which is part of our ongoing work.

Up to now we have been using just raw query logs; in automatic detection, we can leverage more data to do the classification. For *authority sensitivity*, we propose that this facet of a query can be inferred from the clustered keywords that would make up a potential answer in the search results. When a user issues a search that requires an authoritative answer or website (e.g., "us independence day"), we observe that the target answer (e.g., "july fourth") is mentioned repeatedly in the set of relevant documents. For general web queries that are not authority sensitive, we observe keywords are more uniform in distribution – as there is no single correct answer that is sought. Question answering techniques already take advantage of this fact [3], our extension is to apply this logic to quantify the extent of this repetition and to correlate it with authority. If keywords (as determined by words with high inverse document frequency, IDF) in the relevant passages show skewness – marked uneven distribution – we conclude that the query seeks an authoritative answer.

We have implemented a prototype system that uses the Google API to retrieve search results and more importantly the relevant passages (snippets). Given a web query, we obtain a list of the first n=10 search results. Each document's summary snippet is converted to a vector of tokens, where the tokens are stemmed to conflate similar terms together. We then extract the most important terms (keywords) by using the IDF weighting scheme. Frequencies of the keywords are also normalized to sum up to 1.0, resulting in a probability distribution of keywords for a given web query.

The skewness of this keyword distribution is calculated to decide authority sensitivity. The more skewed the distribution, the more likely we believe the query is attempting to retrieve an authoritative answer. To validate this approach we prepared 50 queries, manually tagged and evenly distributed between authority sensitivity and insensitivity.

Polysemous		General		Specific	
model	5/6	carrier	5.0	interest rates for car loans	1.0
red	5/6	law and order	4.33	french speaking countries	1.0
omega	5/6	harley	4.25	download grand theft auto	1.0
bondagedirectory	4/5	top 10 ranking	4.2	download accellerator plus	1.0
braun	4/6	presidents	4.0	top ten search engine ranking	1.0
player	3/5	school locator	3.83	first commercial bank canada	1.0
cruises	3/5	movie press releases	3.8	canada map	1.17
boston	3/5	message boards	3.8	philips light bulbs	1.17
		master	3.67	onspec drivers download	1.17
		xmovie	3.67	87 firebird fuel pump problems	1.17

Table 6: Ten most distinctive queries for *Ambiguity* judged by evaluators.

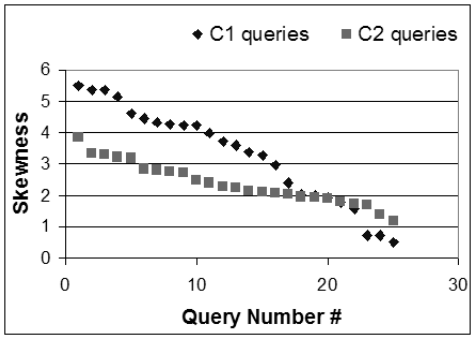

Figure 2: Comparison of skewness distribution of *authority-sensitive* vs. *not authority-sensitive* queries

Figure 2 arranges the two sets of 25 queries in descending order of skewness. We can see that skew values above 4 correspond only to authority sensitive queries and values below 2 are largely authority insensitive queries. Preliminary results show 75%, 60%, 69.6% for precision, recall and F_1 measures respectively. We take these preliminary results as indicative evidence that authority sensitivity can be automatically classified.

We close this section by briefly mentioning possible ways to classify the other facets.

For both *spatial* and *temporal sensitivity*, a possible simple approach is by examining query logs. If a query appears in the query log isolated as well as with temporal or spatial restrictor, then this is evidence that it is sensitive to that facet. For example "us president" would be temporally sensitive as a query log would contain this search as well as "us president 1956"; "cheap hotels" would be spatially sensitive as a query log would contain this search as well as "cheap hotels boston". In these examples, the time "1956" and location "boston" restrictions help us to classify the query as sensitive. This approach simplifies the earlier work in spatial classification [4, 18], using a subset of their features for a unified approach to both spatial and temporal classification.

There are many possible approaches for detecting *ambiguity*. We can rely on an analysis of the returned search results (as we did in authority sensitivity) to measure the diversity to distinguish between general and specific queries. External lexical databases, such as WordNet can also be used to detect polysemous queries.

7. CONCLUSION

We have proposed and evaluated a novel web query classification system. Our classification differs from previous works in two crucial fashions: 1) it is a faceted classification, in which queries are categorized along four general dimensions, and 2) it is a classification intended to lead to actions that an information system can take given a classified query. We believe that our classification can serve as a foundation for bridging current research on users' goals and search engine strategies that can assist users in web search.

Queries tagged as ambiguous, sensitive to authority, time or space lead us to propose respective actions of query disambiguation/expansion, and re-ranking search results based on authoritativeness, recency and locality. Our facets encompass and related to past query classifications. We assessed our classification by asking human subjects to classify sampled queries and found that ambiguity, temporal and spatial authority are well-defined. The remaining facet, Authority sensitivity, obtains lower agreement although an important facet in determining users' information needs.

While we showed possible methods for automating the detection of ambiguity, temporal and spatial sensitivity are possible using query logs, we believe authority sensitivity is most challenging. We thus outlined our current approach towards automating this aspect of the classification and in future work hope to further refine and evaluate our automated query facet classification techniques.

8. REFERENCES

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In *Proc. of the 14th int'l conf. on the World Wide Web (WWW '05)*, pages 2–11, New York, NY, USA, 2005. ACM Press.

[3] C. Clarke, G. Cormack, D. Kisman, and T. Lynam. Question answering by passage selection. In *Proc. of TREC 9*, pages 673–654, 2000.

[4] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *CIKM '03: Proc. of the 12th international conf. on Information and knowledge management*, pages 325–333, New York, NY, USA, 2003. ACM Press.

[5] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. of*

Authority Sensitivity Facet			
Not Sensitive		Sensitive	
rabbits	4.8	telemarketing laws	1.0
message boards	4.6	canada map	1.33
murals	4.6	first commercial bank canada	1.5
lap pools	4.5	insurance quotes	1.6
master bath	4.2	tennessee birth records	1.6
working holiday	4.17	virginia law	1.67
squall	4.0	law and order	1.67
emerald gardens	3.83	patent	1.83
harley	3.83	interest rates for car loans	1.83
movie press releases	3.8	onspec drivers download	1.83
Temporal Sensitivity Facet			
Not Sensitive		Sensitive	
what is the function of ram memory?	5.0	interest rates for car loans	1.17
download accellerator plus	4.8	top ten search engine ranking	1.17
watts amps volts	4.67	who is canada prime minister	1.2
vegetable wash	4.67	presidents	1.4
rabbits	4.6	stock historic results	1.4
emerald gardens	4.5	law and order	1.67
lap pools	4.5	tennessee birth records	2.0
master bath	4.4	video game review	2.0
gravity waves	4.33	inventions japan	2.0
first commercial bank canada	4.33	southern regional championships	2.0
Spatial Sensitivity Facet			
Not Sensitive		Sensitive	
onspec drivers download	5.0	emerald gardens	1.0
download accellerator plus	4.8	interest rates for car loans	1.17
master bath	4.6	independent escorts	1.17
what is the function of ram memory?	4.5	presidents	1.2
air messenger pro download	4.5	cruises	1.4
writer software	4.5	retirement communities	1.5
top ten search engine ranking	4.33	stock historic results	1.6
inventions japan	4.2	election speeches	1.6
tennessee birth records	4.2	law and order	1.67
download grand theft auto	4.2	memorial wall	1.67

Table 7: Ten most distinctive queries for the three sensitivity facets judged by evaluators.

WWW 2004, pages 491–501, New York, NY, USA, 2004. ACM Press.

[6] E. Hovy, U. Hermjakob, and D. Ravichandran. A question/answer typology with surface text patterns. In *Proc. of the Human Language Technology (HLT)*, 2002.

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

[8] 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.

[9] M.-Y. Kan and D. C. C. Poo. Detecting and supporting known item queries in online public access catalogs. In *JCDL '05: Proc. of the 5th joint conf. on Digital libraries*, pages 91–99, New York, NY, USA, 2005. ACM Press.

[10] I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *SIGIR '03: Proc. of the 26th int'l ACM SIGIR*, pages 64–71, New York, NY, USA, 2003. ACM Press.

[11] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. of WWW 2005*, pages 391–400, New York, NY, USA, 2005. ACM Press.

[12] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger. The role of context in question answering systems. In *CHI '03: extended abstracts on Human factors in computing systems*, pages 1006–1007, New York, NY, USA, 2003. ACM Press.

[13] J. J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger. What makes a good answer? the role of context in question answering. In *INTERACT*, 2003.

[14] D. Rose and D. Levinson. Understanding user goals in web search. In *Proc. of WWW 2004*, pages 13–19, 2004.

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

[16] N. Stojanovic. On analysing query ambiguity for query refinement: The librarian agent approach. In *ER 2003*, volume 2813 of *LNCS*, pages 490–505. Springer, DEC 2003.

[17] E. M. Voorhees. Overview of the trec 2003 question answering track. In *Proc. of TREC 2003*, pages 54–68.

[18] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting dominant locations from search queries. In *SIGIR '05: Proc. of the 28th int'l ACM SIGIR*, pages 424–431, New York, NY, USA, 2005. ACM Press.