Analysis and Automatic Classification of Web Search Queries for Diversification Requirements

Sumit Bhatia
Computer Science and Engineering
Pennsylvania State University
University Park, PA 16802, USA
sumit@cse.psu.edu

Cliff Brunk
Yandex Labs
299 S. California Avenue
Palo Alto, CA 94306, USA
cliff@team-yandex.ru

Prasenjit Mitra
Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802, USA
pmitra@ist.psu.edu

ABSTRACT

Search result diversification enables the modern day search engines to construct a result list that consists of documents that are relevant to the user query and at the same time, diverse enough to meet the expectations of a diverse user population. However, all the queries received by a search engine may not benefit from diversification. Further, different types of queries may benefit from different diversification mechanisms. In this paper we present an analysis of logs of a commercial web search engine and study the web search queries for their diversification requirements. We analyze queries based on their click entropy and popularity and propose a query taxonomy based on their diversification requirements. We then carry out the task of automatically classifying web search queries into one of the classes of our proposed taxonomy. We utilize various query-based, click-based and reformulation-based features for the query classification task and achieve strong classification results.

Keywords

Diversity, Query Classification, Web Search, Query Logs.

INTRODUCTION

Queries submitted to a Web Search Engine (WSE) typically consist of 2-3 terms and hence, seldom clearly specify the underlying information need of the user. In such a scenario, the WSE can minimize the probability of dissatisfaction of the average user by constructing and presenting a diverse set of search results to the user that covers different aspects underlying the original user query (Agrawal et al., 2009).

ASIST 2012, October 28-31, 2012, Baltimore, MD, USA.

Most current search result diversification approaches diversify results either *implicitly* by including documents in the result set that minimize redundancy and maximize novel information (Carbonell & Goldstein, 1998, Chen & Karger, 2006, Wang & Zhu, 2009) or *explicitly* by including documents corresponding to various aspects/sub-topics of the original user query (Agrawal et al., 2009, Santos et al., 2010a).

To improve performance, a commercial WSE/practical information retrieval system utilizes various query analysis techniques to identify and use *query specific* retrieval strategies in order to present an optimized result list for the query. For example, different methods may be used by the search engine for long queries (Bendersky & Croft, 2008), queries with commercial intents (Dai et al., 2006), queries that require localization of results based on user's location (Lu et al., 2010) etc. Most of the current methods of search result diversification however, treat all the queries as equal whereas not all the queries received by a search engine will benefit from search result diversification. Hence from a search engine's perspective, *it is crucial to differentiate queries that may potentially benefit from search result diversification from those that may not*. For queries that may potentially benefit from diversification, different queries may require different diversification strategies. For example, for an ambiguous query like *"java"*, the first crucial step is to identify different interpretations/meanings of the query and then accordingly present results corresponding to each of these different interpretations (Java programming language, island in Indonesia named Java etc.). On the other hand, for a query like *"java tutorial"*, where it is clear that the user is interested in information related to the programming language, the search engine should try to present tutorials of diverse nature (tutorials on different aspects of the Java programming language, different difficulty level, etc.). Further, some queries may require a more aggressive result diversification as compared to other queries (Santos et al., 2010b).

Motivated by these considerations, the research questions that we address in this work and our contributions are as follows:

1. How important search result diversification is from a Web Search Engine's perspective, i.e., for what fraction of web search queries, it may be beneficial to diversify search results?

All the queries received by a WSE may not benefit from diversification of search results. Intuitively, queries that are ambiguous or multi-faceted may benefit from search result diversification (Boyce, 1982). We expect such queries to have many different URLs being clicked by different users in the past. In order to identify such queries we use click entropy (Dou et al., 2007) which measures the spread of user clicks for a given query. Click entropy has also been used previously to identify ambiguous queries (Wang & Agichtein, 2010) and queries that can potentially benefit from personalization (Teevan et al., 2008) and diversification (Clough et al., 2009). We analyze query logs from a commercial web search engine consisting of more than 373 million query transactions and 87 million unique queries. Our analysis based on the queries' click entropy revealed that at least 20.35% of the query transactions in the logs may potentially benefit from search result diversification (ref. Section Query Log Analysis for Diversity). Further, we found that these query transactions account for only 0.53% of the unique queries in the logs indicating that by improving search results by employing diversification strategies for only a (relatively) small number of popular (frequent) queries, *the search engine can improve results and hence, user experience for almost one-fifth of query transactions.*

2. For queries that may benefit from search result diversification, are there any differences in the types of diversification requirements for these queries?

In order to find an answer to this question, we randomly sampled 500 queries from the queries that were identified as potentially benefiting from diversification. Based on a manual analysis of these queries and the associated documents shown and clicked for these queries, we propose four query classes from a diversity perspective:

(i) Ambiguous queries,

(ii) Unambiguous but underspecified queries,

(iii) Information gathering queries, and

(iv) Miscellaneous (ref. Section Query Log Analysis for Diversity).

3. Can we automatically classify web search queries according to their diversification requirements?

On receiving a user query, a WSE needs to first identify the nature/type of query so that an appropriate diversification strategy can be used. We train supervised classifiers for automatic web query classification as per their diversification requirements. We utilize four types of features for this task -- query based, features based on query request type, click based and reformulation based features. We achieve encouraging classification results with an overall classification accuracy of 72.35% and an overall F-1 measure of 0.735 (ref. Sections Automatic Query Classification and Experiments). The dataset used in this paper consisting of 500 queries and associated class labels is also being made available to research community.

RELATED WORK

The work reported in this paper is related to search result diversification, query log analyses and web query classification. In this section, we provide an outline of some of the representative research that is most closely related to our work.

Search Result Diversification

Maximum Marginal Relevance (MMR) (Carbonell & Goldstein, 1998) represents one of the earliest attempts for search result diversification. For a given user query MMR selects documents that are relevant to the user query as well as provide novel information when compared to previously selected documents. Chen and Karger (2006) argue that the strategy of returning as many relevant results as possible (the *Probability Ranking Principle (PRP)(Robertson & Jones, 1976)*) is not always optimal. Hence they put forward the idea of returning a set of documents that maximizes the probability of finding a relevant document in top-k documents. Agrawal et al. (2009) study the problem of diversifying search results of ambiguous web queries. They assume the availability of a taxonomy of information and that both queries and documents may belong to one or more categories in this taxonomy. The problem is formulated as an optimization problem that aims to maximize the probability of satisfying the average user. Gollapudi and Sharma (2009) describe an axiomatic framework that can be used for designing and characterizing diversification mechanisms. Santos et al. (2010a) proposed the xQuAD (explicit Query Aspect Diversification) framework that takes into account various *aspects* of an underspecified query. In the proposed framework, the different aspects of a given query are represented in terms of *sub-queries* and the documents are ranked based on their relevance to each sub-query. Welch et al. (2011) describe an algorithm for diversifying results of informational queries where the user's information need is satisfied by not one but multiple relevant documents. Santos et al. (2010b) propose a supervised selective diversification approach that trades off relevance and diversity on a per query basis. He et al. (2011) describe a clustering based framework for result diversification.

Query Log Analysis

Web search engine query logs contain a wealth of information about users' behavior, their information requirements and how users interact with the search engines. Hence, study and analyses of search engine logs

can provide useful insights about user requirements as well as weaknesses of the current state-of-the-art search engines. One of the first large scale analysis of web search engine query logs was presented by Silverstein et al. (1999). They analyzed logs of Alta Vista search engine consisting of approximately one billion search requests and 285 million user sessions. They noted significant differences between users of web search engines and users of traditional information retrieval systems. Specifically, queries issued to web search engines are much shorter, users generally see only the first result page and query reformulations are less frequent. Ross and Wolfram (2000) analyzed logs of Excite search engine and categorized most frequently co-occurring query term pairs into one or more of 30 subject areas. Beitzel et al. (2004) analyzed one week (26 December 2003 -- 1 January 2004) of logs from America Online (AOL) and found that average query length is 2.2 terms, roughly 2% of queries contain query operators and about 81% of users looked at only the first results page. Further, they also observed changes in frequency and popularity of topically categorized queries across the hours of the day. Jansen and Spink (2006) present a comprehensive comparison of nine different studies of search engine logs performed over a period of seven years. They found that many characteristics such as session length measured in number of queries, number of single term queries remain stable over different time periods and search engines, however, the number of users that only look at the first results page has increased over time which could be attributed to improvements in algorithms used by search engines. Clough et al. (2009) analyzed Microsoft Live Search logs and found that at least 9.5%-16.2% of the queries could benefit from diversification. The analyses of search logs presented in this paper differs from previous works in that in addition to analyzing the logs to identify how many queries can benefit from diversification methods, we also study differences among different queries in terms of their diversification requirements. Such an analysis provides insights about what different types of diversification strategies should the search engines use and how much can search result diversification methods benefit the users.

Query Classification

There has been much work on web query classification where queries are classified into certain target categories depending upon the application at hand. Broder (2002) in his seminal work developed a taxonomy for web search queries and categorized web search queries as informational, transactional and navigational queries. Kang and Kim (2003) describe methods to classify web queries into following three categories depending upon the user's intent -- *(i)* topic relevance task (informational queries), *(ii)* homepage finding task (navigational) and *(iii)* service finding task (transactional). A web query classification challenge was organized as KDD-CUP 2005 competition (Li et al., 2005) where participants were required to classify 800,000 web search queries into 67 predefined topical categories. Gravano et al. (2003) classified web queries as *local* and *global* depending upon whether the search engine should present localized results based on the users' geographical location. Local queries such as *"san francisco flower shop"* require the localized results whereas a global query such as *"java applet"* does not require geographical localization. The work by Wang and Agichtein (2010) is similar to our work in the sense that they use clickthrough information to classify queries into ambiguous and informational queries. However, the taxonomy of queries proposed in this work is different than the categories defined by them and in addition to clickthrough information, we also explore query level, URL level and reformulation based features for query classification.

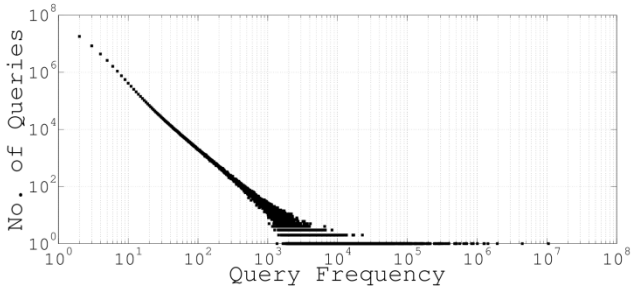

Figure 1. Plot showing distribution of query frequencies in the query logs. The distribution follows a power law with α =1.16.

DATA DESCRIPTION

We used query logs of a commercial web search engine that we were provided access to. The logs were for queries issued in the United States market. Table 1 summarizes various statistics about the dataset. The logs consist of more than 373 million query requests out of which there are about 87 million unique queries. Mean query length (in number of terms) for all the queries is 1.08 terms per query whereas considering only the unique queries, mean query length is 4.63 terms per query. Out of the roughly 87 million unique queries, about 5.5 million queries are single term queries. Figure 1 depicts the distribution of query frequencies as observed in the query logs which follows a power law with $\alpha = 1.16$. Of all the 87 million unique queries, roughly 47 million queries are issued only once.

QUERY LOG ANALYSIS FOR DIVERSITY

In this section, we explore what types of queries may benefit from search result diversification. In particular, our focus is on finding an answer to following questions.

1. What fraction of queries can be potentially benefited from diverse search results?
2. How the queries that may benefit from diversification differ in terms of their diversity requirements?

Query Statistics	
Number of queries	373,439,364
Number of unique queries	87,347,656
Mean query length (no. of terms)	1.08
Mean unique query length (no. of terms)	4.63
Number of unique single term queries	5,559,118
Number of queries issued only once	46,825,903
Reformulation Statistics	
Number of reformulations	21,616,189
Average number of reformulations per query	2.66
Number of queries that were reformulated in a session	14,288,180

Table 1. Characteristics of the query log data used in this work.

Identifying Queries that may Potentially Benefit from Diversification

Intuitively, queries that represent multiple information needs may potentially benefit from diversification of search results. Such queries should have a large number of different URLs being clicked by different users and hence, we use *click entropy* (Dou et al., 2007) to identify such queries in the logs. Click entropy measures the spread of URLs clicked for a given query and has been used previously to identify ambiguous queries (Wang & Agichtein, 2010) and queries that can potentially benefit from personalization (Teevan et al., 2008) and diversification (Clough et al., 2009).

Click entropy (CE) for a query q is defined as follows.

$$CE(q) = \sum_{d \in D_q} -P(d|q) log_2 P(d|q)$$

Here, D_q is the set of documents/URLs clicked by various users for query q. A higher value of click entropy indicates that users selected different documents for the given query indicating that different users were looking for different information when they used the given query and hence, indicates a potential for diversification. The idea here is to identify queries with high click entropies and observe the reasons for users clicking different URLs for the query.

In order to ensure that we have sufficient evidence for queries used for analysis, we considered only those queries that appeared in the logs at least ten times. That resulted in a total of 2,485,228 unique queries that appeared for a total of 199,208,177 times in the query logs. Click entropy was computed for all these queries and Figure 2 shows a scatter plot between query frequency and query click entropy for this set of queries. Each point on the plot represents a query with its frequency (log scale) on y-axis and its click entropy on x-axis.

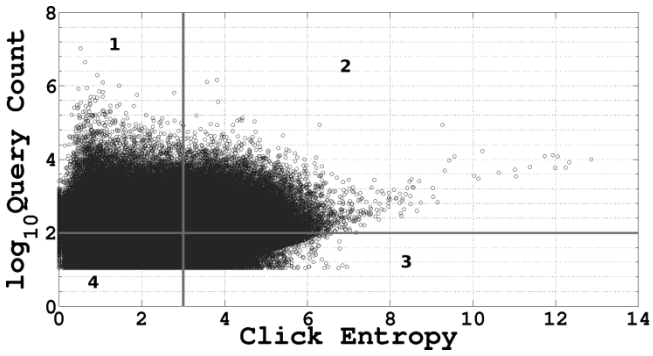

Figure 2. Scatter plot showing query frequency and associated click entropy as observed in the query logs. The plot is divided into four quadrants (see text for details). Quadrants marked as 1, 2, 3 and 4 refer respectively to HFLE, HFHE, LFHE and LFLE quadrants.

Next, for further analysis we divided the plot into four quadrants based on frequency and entropy values of queries by choosing a threshold frequency of 100 and a threshold entropy of 3. The threshold entropy of three has been used to weed out navigational queries and is in accord with values used in previous research (Clough et al., 2009). We were also interested in separating the queries based on their popularity (as measured by frequency of occurrence in the logs), hence we chose a threshold frequency of 100 times in the six months period to separate the queries into low and high frequency (popularity) categories. By using these thresholds, we can divide all the queries in the logs in following four categories:

1. Low frequency and low entropy (LFLE) queries
2. Low frequency and high entropy (LFHE) queries
3. High frequency and low entropy (HFLE) queries
4. High frequency and High entropy (HFHE) queries

Table 2 summarizes some other statistics about queries in each of the four categories. Queries in the LFLE class account for 2.24% of all the unique queries in the logs and appear roughly forty four million times in the query logs (11.83%). A large fraction of queries in this class are generally *long-tail* queries where the user is generally looking for a specific piece of information. E.g. "*ohio department of corrections", "mutual savings credit union"* etc. Many of the queries in this class are navigational queries. Queries in HFLE class are mostly navigational or transactional queries where the user is looking for a specific website (e.g. "*pogo"*, "*askjeeves.com"* etc.) or answers to some common questions (e.g. "*calories in strawberry"* etc.).

Query Class	Condition	No. of Unique Queries	No. of Times Query Issued
LFLE	Frequency ≤ 100, Entropy < 3	1,958,351 (2.24%)	44,183,993 (11.83%)
LFHE	Frequency ≤ 100, Entropy > 3	338,076 (0.39%)	10,734,720 (2.87%)
HFLE	Frequency > 100, Entropy ≤ 3	66,177 (0.08%)	78,998,631 (21.15%)
HFHE	Frequency > 100, Entropy > 3	122,624 (0.14%)	65,290,833 (17.48%)

Table 2. Four classes of queries based on frequency and click entropy values. The percentage values are with respect to the whole query log data.

Queries belonging to LFHE class are also generally quite specific. The reason for the high entropy values is due to the fact that these queries are generally *"literature survey"* type queries – the user is looking for various aspects of the query or a single document is not able to provide the complete information (e.g. "*peru facts"*). Queries in this class even though are relatively infrequent and lie in the long-tail, can potentially benefit from for diversification.

From Figure 2, we note that there are a number of queries that have high frequency as well as high click entropies (HFHE queries). Even though the number of unique queries in this class is small (0.14% of all the unique queries), the fact that these queries have high frequencies indicate that these queries are issued repeatedly by a considerable fraction of user population (17.48% of all the queries). Further, high click entropies indicate that the users are clicking on different URLs for these queries indicating, along with many other things, that the users' information needs are not being satisfied by the results shown by the search engine. Thus, improving search results for these queries is extremely crucial. These are the popular queries that have a high potential for diversification and hence, should be the prime focus of the search engine's diversification framework. In next subsection, we present an analysis of queries belonging to the high entropy categories (HFHE and LFHE).

Types of Diversification Requirements

In order to study and understand the reasons for high click entropies of the queries in HFHE and LFHE categories, we randomly sampled queries from the HFHE and LFHE classes and analyzed the URLs shown to the users and the URLs that were clicked by the users for these queries. Based on our manual analysis of the queries, we propose the following query classes:

1. Ambiguous queries (A): Ambiguous queries have more than one meaning. For instance, *"jaguar"* can mean both an animal and a car (and even an old Mac OS operating system). Further, a considerable fraction of these queries are the acronym queries such as the query *"iit"* which could refer to either the Indian Institute of Technology or the Illinois Institute of Technology. Sometimes, one meaning of the query may be more likely than another. For example, consider the query *"paris"* -- it can refer to the capital city of France or it can also mean the casino in Las Vegas, USA. For these types of queries, the search engine needs to ensure that the documents corresponding to the different possible interpretations of the query are presented to the user. For this purpose, a topic hierarchy such as the one provided by the Open Directory Project (ODP) (http://www.dmoz.org/) can be utilized.

2. Unambiguous but underspecified queries (U): These queries are unambiguous in the sense that the meaning of these queries is clear; there is only one way to "read" or "interpret" these queries. They refer to an unambiguous entity however, it is not clearly specified what the user wants to know about the entity. E.g., consider the query *"lady gaga"*. Here there is no ambiguity in what the query means but still it is not clear what the user wants to know about Lady Gaga -- does she want to watch the music videos, read news, find song lyrics, or purchase songs at the iTunes store? The user's intent is not explicitly specified. For such queries, the search engine needs to focus on discovering the underlying intents behind the underspecified query and accordingly create a result list to cover these different intents.

3. Information gathering queries (browsing) (I): These queries have a clear meaning and are sufficiently specified, but the user does not expect one result to answer his or her need. For example, consider the query *"peru facts"* or *"how to make cheesecake"* etc. The user prefers to see novel (new and non-redundant) information in different documents. The user expects to see many good results and browse them, collecting information. For such queries, the novelty and redundancy considerations are important.

4. Miscellaneous/None of the above (M): The queries that belong to this category correspond to download/watch movies online, download software for which the click entropy is high due to the fact that many of the URLs for these queries are spam/misleading leading a user to try different URLs till he gets the desired result. For example, for many *"download software"* type queries, the user may have to try many different URLs till a working URL is found.

AUTOMATIC QUERY CLASSIFICATION

As described in the previous section, the reasons for diverse clicks (or high click entropies) for different queries are different and hence, it is essential for a search engine to determine the type of query automatically so that the appropriate mechanisms can be utilized to construct the result list as per the requirements of the queries. In this section, we report results of experiments on automatically

classifying queries into one of the above described four query classes.

Feature Description

For automatic query classification we derive a number of features from the input query itself as well as utilize the information present in the query logs in the form of the URLs shown for the query in the past and click-through information as well as various reformulations for the query in the past. Some of the features used require query logs while some can be computed independent of the logs (see Table 3 for this information). The various features used are described in detail in this subsection and are summarized in Table 3.

1. Query Features: These features are derived from the input query and try to capture various characteristics of the query that may be indicative of its diversity requirements. For example, we expect underspecified queries to be shorter in length than well specified queries. Hence, we use number of terms and characters in the input query as features for the automatic query classification task. Likewise, we use clarity score of a query as one of the features as it has been shown to be a good measure for query ambiguity (Cronen-Townsend et al., 2002). The clarity score for a query Q is computed as follows:

$$clarityScore(Q) = \sum_{w \in V} P(w|Q) log_2 \frac{P(w|Q)}{P_{coll}(w)}$$

Here, V is the vocabulary generated from the corpus, $P(w|Q)$ represents the probability of the term w as computed from the query language model and $P_{coll}(w)$ represents the probability of the term w as computed from the corpus' language model. The clarity score, thus computes the K-L divergence between the query and collection (corpus) language models. If the query is highly ambiguous and could be satisfied by documents belonging to many different topics, then the probability distribution over words as computed by the query language model will be similar to the probability distribution over words as computed by the corpus language model, hence low divergence between the two language models and a lower clarity score. Likewise, for highly specific and unambiguous queries, the query will be satisfied by documents belonging to that specific topic and thus, the query language model will diverge from the collection language model which contains documents belonging to a wide variety of documents, and hence, a higher clarity score. In other words, ambiguous queries or queries that can be satisfied by a diverse set of heterogeneous documents have relatively lower clarity scores and unambiguous queries and queries that can be satisfied by homogeneous documents have relatively higher clarity scores.

2. Query Request Type Features: These features try to capture the nature of information/content that has been requested in the query. For example, presence of a URL in the query may indicate navigational intent whereas the presence of 5W1H words (what, where, when, why, who, how) may indicate a question being asked in the query (information browsing behavior). We were provided access to proprietary query request type classifiers used by a commercial web search engine and these were used to evaluate IsDownload, IsIMG, IsVid, IsPorn, IsTV features (see Table 3). Given an input query, these classifiers predict whether the results from a specific vertical are also relevant for the query or not and hence, using their predictions as features can help identify certain underspecified queries. For example, for the query *"madonna",* the classifiers tell us that this underspecified query can also benefit from results from image and video search verticals.

3. URL and Click Features: These features try to capture the differences between clickthrough patterns of queries belonging to different query classes. We use total number of URLs clicked for a query as well as the number of unique URLs for the query. We use click entropy that measures the spread of URLs clicked for a given query as well as domain entropy (Wang & Agichtein, 2010) that measures the spread of domains of all URLs being clicked for the query. The ClickSTD feature (ref. Table 3) is defined as the standard deviation of click frequencies of URLs being clicked for the query. For information browsing queries, we expect that the URLs shown for these queries should have a more uniform distribution of clicks as the users want to gather information by browsing multiple documents whereas for ambiguous and underspecified queries we expect the clicked URL distribution to be skewed towards URLs corresponding to the dominant intents behind the query.

4. Reformulation Features: Generally, if the user is not satisfied with the results shown by the search engine for the initial query, she may reformulate the query by either changing the query or by providing additional specifications by adding new terms to the query (Jansen et al., 2009). This is especially true for ambiguous and underspecified queries (Sanderson, 2008). In order to compute derive features based on this reformulation behavior, we first define a *search session* as consisting of all the transactions (query requests and clicks) performed by the same user (identified by a unique user id in the query logs) with in a time period of 15 minutes (typical length of a web search session as has been observed previously (Jansen & Spink, 2003)). Further we say that a query q_2 is a reformulation of query q_1 iff q_2 is issued after q_1 in the same search session as q_1 and q_1 and q_2 have at least one non-stopword in common. We process the query logs to identify search sessions and extract query reformulations from these sessions. This information is then used to compute reformulation based features such as number of different reformulations for a query, number of sessions in which the query is being reformulated, average number of additional words in all of the query's reformulations etc.

EXPERIMENTS

Data Preparation

We took a random sample of 500 queries from the queries belonging to the HFHE and LFHE categories and asked three human evaluators to assign the queries into one of the four query classes as described above. Each evaluator provided class labels for all the 500 queries and the final label of a query was decided by the majority vote. Queries that were assigned different labels by all the three evaluators were discarded. The numbers of queries belonging to the different classes as assigned by the evaluators are summarized in Table 4. We note that a majority decision was obtained for 454 queries (90.8%). The Fleiss' Kappa score for multi-rater agreement was 0.44, indicating moderate agreement.

Table 5 lists some representative queries for each of the four classes. The complete list of queries used in this work and the labels provided by evaluators are available upon request for research purposes.

Feature	Description	Require Logs	Type
	Query Features		
QueryLength	Number of words in query	No	Numeric
CharCount	Number of characters in query	No	Numeric
Clarity	Clarity Score (Cronen-Townsend et al., 2002) of the query	No	Real
QueryFrequency	Number of times query occurs in the search logs	Yes	Numeric
QueryIssueTime	No. of times query was issued in one of the four defined time periods (12AM-6AM, 6AM-12PM, 12PM-6PM, 6PM-12AM).	Yes	Numeric
	Query Request Type Features		
IsURL	Is there a URL in the query	No	Binary
IsDownload	If the query contains the word download	No	Binary
IsIMG	If the query contains request for images	No	Binary
IsVid	If the query contains request for a video	No	Binary
IsPorn	Is the query a porn query	No	Binary
IsQuestion	If any of the 5W1H words (what, where, why, when, who, how) present in the query	No	Binary
IsTV	If the query contains request for tv shows	No	Binary
IsFree	If the query contains the keyword *free*	No	Binary
ForeignQuery	If the query is a non-english language query	No	Binary
	URL and Click Features		
ClickFrequency	Total number of clicks for the query	Yes	Numeric
URLCount	Number of unique URLs that were clicked for the query	Yes	Numeric
Query-URL-CountRatio	Ratio of QueryFrequency and URLCount	Yes	Real
ClickEntropy	Click entropy of the query	Yes	Real
ClickSTD	Standard deviation of frequencies of URLs being clicked for the query	Yes	Real
DomainEntropy	Domain entropy (Wang & Agichtein, 2010) of the query	Yes	Real
	Reformulation Features		
NumReformulations	Number of different reformulations for a query	Yes	Numeric
ReformulationsInSession	Total number of sessions in which the query is being reformulated	Yes	Numeric
Reform-SessionRatio	Ratio of NumReformulations and ReformulationsInSession	Yes	Real
AvgReformIncrement	Average number of additional words in all of the query's reformulations	Yes	Real
AvgUniqueReformIncement	Average number of additional words in all of the query's unique reformulations	Yes	Real

Table 3. List of features used for automatic query classification and their description.

Experimental Protocol

We used implementation of different classifiers as provide by the Weka toolkit (Hall et al., 2009). We experimented with a variety of supervised classification schemes including decision trees, SVM, multi-layer perceptron classifier, naive bayes classifier and a logit model classifer. The performance of all the classifiers was comparable with the logit model classifier achieving the best performance. Hence, we report results obtained using the logit model classifier as the main focus of this work is not on comparing different machine learning algorithms but to study if we can automatically classify web search queries as per their diversification requirements. We used stringent ten-folds cross validation for experiments and the results reported are averaged over the ten folds.

Query Class	All Agree	Two Agree	No Agreement
A	26	18	--
U	83	83	--
I	59	91	--
M	55	39	--
Total	223	231	46

Table 4. Statistics about class labels as provided by the three evaluators.

Automatic Query Classification Results

Table 6 reports the results for automatic query classification task where we achieve an overall classification accuracy of 72.35%. Note that these numbers are for all the classes combined, hence in order to study the classifier performance for each class, we report separate results for each class in Table 7. We report precision, recall, F-score and are under the ROC curve for each of the four classes and also the overall metrics. We achieved an overall precision of 74.8% and a recall of 73.3%. We also note that the minimum F-Score of 0.659 is achieved for class **A** (Ambiguous queries) and maximum F-Score of 0.807 is achieved for class **M** (Miscellaneous queries). Table 8 reports the confusion matrix for the four classes. From the table we observe that it is relatively difficult to distinguish between the queries belonging to **I** and **U** classes. 38 queries belonging to I class were classified as belonging to class U and 27 queries belonging to U class were classified as belonging to class I.

Ambiguous(A)	motion
	nci
	being human
	safety
	up
Unambiguous but underspecified (U)	firefox web browser
	2011 buick regal
	anorei collins
	girls tattoos
	carmen villalobos
Information Browsing (I)	free psn codes
	best facebook statuses ever
	african american inventors
	chicken and rice recepies
	colon cancer symptoms
Miscellaneous (M)	3d pinball space cadet download
	teen wolf episode 4
	busted celebrity.com
	jersey shore season 4 episode 5 full episode
	www.chase.com

Table 5. Some examples of queries used in the experiments belonging to each of the four classes.

Classification Accuracy	72.35%
Kappa Statistic	0.6156
Mean absolute error	0.1897
Root mean square error	0.3108

Table 6. Classification results for the automatic query classification task.

Class	Precision	Recall	F-Score	ROCArea
A	0.711	0.614	0.659	0.918
U	0.657	0.783	0.714	0.842
I	0.747	0.727	0.736	0.893
M	0.931	0.713	0.807	0.923
Overall	0.748	0.733	0.735	0.883

Table 7. Classification results for the automatic query classification task for each individual query class.

	A	U	I	M
A	27	16	0	1
U	7	130	27	2
I	1	18	109	2
M	3	14	10	67

Table 8. Confusion matrix for the four classes. Entry *(i,j)* refers to the number of queries in class *i* that were classified as belonging to class *j*.

Conclusions and Future Work

We presented an analysis of web search queries as per their diversification requirements. Our analysis of logs of a commercial search engine (yandex.com) revealed that 0.53% (460,700) of all the unique queries in the logs are high entropy queries (HFHE+LFHE) and they account for 20.35% of all the query mass, i.e., one in five queries present in the logs can potentially benefit from search result diversification. Further, based on analysis of queries with high click entropy we proposed to classify web queries from the perspective of their diversification requirements into following four classes: ambiguous, unambiguous but underspecified, information gathering and miscellaneous. We also studied the problem of automatically classifying web search queries into these four classes. We utilized features described from the user input query, click-through information and query reformulations and achieved an overall precision of 74.8% and recall of 73.3% for the automatic query classification task. Our future will focus on developing query-specific diversification strategies.

ACKNOWLEDGEMENTS

We would like to thank Dr. Jim Jansen and Sujatha Das for providing valuable suggestions to improve this draft. Part of this work was performed while Sumit Bhatia was an intern at Yandex Labs. Part of this work was funded by National Science Foundation under Grant No. 0845487. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

Agrawal, R., Gollapudi, S., Halverson, A. & Ieong, S. (2009), Diversifying search results, *in* 'WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining', ACM, pp. 5–14.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. & Frieder, O. (2004), Hourly analysis of a very large topically categorized web query log, *in* 'Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval', SIGIR '04, ACM, New York, NY, USA, pp. 321–328. http://doi.acm.org/10.1145/1008992.1009048

Bendersky, M. & Croft, W. B. (2008), Discovering key concepts in verbose queries, *in* 'Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval', SIGIR '08, ACM, New York, NY, USA, pp. 491–498. http://doi.acm.org/10.1145/1390334.1390419

Boyce, B. (1982), 'Beyond topicality: A two stage view of relevance and the retrieval process', *Information Processing & Management* **18**(3), 105 – 109. http://www.sciencedirect.com/science/article/pii/0306457382900334

Broder, A. (2002), 'A taxonomy of web search', *SIGIR Forum* **36**, 3–10. http://doi.acm.org/10.1145/792550.792552

Carbonell, J. & Goldstein, J. (1998), The use of mmr, diversity-based reranking for reordering documents and producing summaries, *in* 'SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, NY, USA, pp. 335–336.

Chen, H. & Karger, D. R. (2006), Less is more: probabilistic models for retrieving fewer relevant documents, *in* 'Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval', SIGIR '06, ACM, New York, NY, USA, pp. 429–436. http://doi.acm.org/10.1145/1148170.1148245

Clough, P., Sanderson, M., Abouammoh, M., Navarro, S. & Paramita, M. (2009), Multiple approaches to analysing query diversity, *in* 'SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, NY, USA, pp. 734–735.

Cronen-Townsend, S., Zhou, Y. & Croft, W. B. (2002), Predicting query performance, *in* 'SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, NY, USA, pp. 299–306.

Dai, H. K., Zhao, L., Nie, Z., Wen, J.-R., Wang, L. & Li, Y. (2006), Detecting online commercial intention (oci), *in* 'Proceedings of the 15th international conference on World Wide Web', WWW '06, ACM, New York, NY, USA, pp. 829–837. http://doi.acm.org/10.1145/1135777.1135902

Dou, Z., Song, R. & Wen, J.-R. (2007), A large-scale evaluation and analysis of personalized search strategies, *in* 'Proceedings of the 16th international conference on World Wide Web', WWW '07, ACM, New York, NY, USA, pp. 581–590.

Gollapudi, S. & Sharma, A. (2009), An axiomatic approach for result diversification, *in* 'WWW '09: Proceedings of

the 18th international conference on World wide web', ACM, New York, NY, USA, pp. 381–390.

Gravano, L., Hatzivassiloglou, V. & Lichtenstein, R. (2003), Categorizing web queries according to geographical locality, *in* 'Proceedings of the twelfth international conference on Information and knowledge management', CIKM '03, ACM, New York, NY, USA, pp. 325–333. http://doi.acm.org/10.1145/956863.956925

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009), 'The weka data mining software: An update', *SIGKDD Explorations* **11**(1).

He, J., Meij, E. & de Rijke, M. (2011), 'Result diversification based on query-specific cluster ranking', *Journal of the American Society for Information Science and Technology* **62**(3), 550–571. http://dx.doi.org/10.1002/asi.21468

Jansen, B. J., Booth, D. L. & Spink, A. (2009), 'Patterns of query reformulation during web searching', *Journal of the American Society for Information Science and Technology* **60**(7), 1358–1371.

Jansen, B. J. & Spink, A. (2003), An analysis of web documents retrieved and viewed, *in* H. R. Arabnia & Y. Mun, eds, 'International Conference on Internet Computing', CSREA Press, pp. 65–69.

Jansen, B. J. & Spink, A. (2006), 'How are we searching the world wide web?: a comparison of nine search engine transaction logs', *Inf. Process. Manage.* **42**, 248–263. http://dx.doi.org/10.1016/j.ipm.2004.10.007

Kang, I.-H. & Kim, G. (2003), Query type classification for web document retrieval, *in* 'Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval', SIGIR '03, ACM, New York, NY, USA, pp. 64–71. http://doi.acm.org/10.1145/860435.860449

Li, Y., Zheng, Z. & Dai, H. K. (2005), 'Kdd cup-2005 report: facing a great challenge', *SIGKDD Explor. Newsl.* **7**, 91–99. http://doi.acm.org/10.1145/1117454.1117466

Lu, Y., Peng, F., Wei, X. & Dumoulin, B. (2010), Personalize web search results with user's location, *in* 'Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval', SIGIR '10, ACM, New York, NY, USA, pp. 763–764. http://doi.acm.org/10.1145/1835449.1835604

Robertson, S. E. & Jones, K. S. (1976), 'Relevance weighting of search terms', *Journal of the American Society for Information Science* **27**, 129–146.

Ross, N. C. M. & Wolfram, D. (2000), 'End user searching on the internet: An analysis of term pair topics submitted to the excite search engine', *Journal of the American Society for Information Science* **51**(10), 949–958. http://dx.doi.org/10.1002/1097-4571(2000)51:10<949::AID-ASI70>3.0.CO;2-5

Sanderson, M. (2008), Ambiguous queries: test collections need more sense, *in* 'SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, NY, USA, pp. 499–506.

Santos, R. L., Macdonald, C. & Ounis, I. (2010*a*), Exploiting query reformulations for web search result diversification, *in* 'WWW '10: Proceedings of the 19th international conference on World wide web', ACM, New York, NY, USA, pp. 881–890.

Santos, R. L., Macdonald, C. & Ounis, I. (2010*b*), Selectively diversifying web search results, *in* 'Proceedings of the 19th ACM international conference on Information and knowledge management', CIKM '10, ACM, pp. 1179–1188. http://doi.acm.org/10.1145/1871437.1871586

Silverstein, C., Marais, H., Henzinger, M. & Moricz, M. (1999), 'Analysis of a very large web search engine query log', *SIGIR Forum* **33**, 6–12. http://doi.acm.org/10.1145/331403.331405

Teevan, J., Dumais, S. T. & Liebling, D. J. (2008), To personalize or not to personalize: modeling queries with variation in user intent, *in* 'Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval', SIGIR '08, ACM, New York, NY, USA, pp. 163–170. http://doi.acm.org/10.1145/1390334.1390364

Wang, J. & Zhu, J. (2009), Portfolio theory of information retrieval, *in* 'SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval', ACM, New York, NY, USA, pp. 115–122.

Wang, Y. & Agichtein, E. (2010), Query ambiguity revisited: clickthrough measures for distinguishing informational and ambiguous queries, *in* 'Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics', HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 361–364. http://portal.acm.org/citation.cfm?id=1857999.1858054

Welch, M. J., Cho, J. & Olston, C. (2011), Search result diversity for informational queries, *in* 'Proceedings of the 20th international conference on World wide web', WWW '11, ACM, New York, NY, USA, pp. 237–246. http://doi.acm.org/10.1145/1963405.1963441