A Query Classification Scheme For Diversification

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

Cliff Brunk
Yandex Labs
Palo Alto, CA-USA
cliff@yandex-team.ru

Prasenjit Mitra
Information Science and Technology
Pennsylvania State University
University Park, PA-16802
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 diverse user expectations. 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 initial results of our efforts to study the diversification requirements of queries in a web search scenario. We use click entropy as a measure to identify queries that can potentially benefit from search result diversification and propose a query taxonomy based on their diversification requirements. We also present results of experiments to automatically classify queries into these categories.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Query formulation, Search Process; H.3.5 [**Information Storage and Retrieval**]: Online Information Services – Web-based services

General Terms

Human factors, Algorithms, Experimentation

Keywords

Query log analysis, query classification, query taxonomy, query ambiguity, search result diversification.

1. INTRODUCTION

Queries submitted to a web search engine typically consist of 2–3 terms and hence, do not always clearly specify the underlying information need of the user. In such a scenario, the search engine can present a diverse set of search results to the user so as to cover different aspects underlying the original user query. Most of the current approaches for search result diversification focus on including documents in the result set that minimize redundancy and maximize novel information [4, 5, 19] or by explicitly including documents corresponding to various aspects/sub-topics of the original user query [1, 14]. The current methods of search result diversification treat all queries as equal however, not all the queries received by a search engine may benefit from search result diversification. Hence from a search engine's perspective, it is important to differentiate queries that may potentially benefit from search result diversification from those that may not. Even for queries that may potentially benefit from diversification, some may require a more aggressive result diversification as compared to other queries [15]. Further, it is not clear what types of queries require what type of diversity as different queries may require different diversification strategies. For example, for an ambiguous query like *"java"*, the search engine should try to present results corresponding to the different interpretations of the query (programming language, place etc.) whereas, for a query like *"java tutorial"* where the user intent is clear, the search engine should try to present diverse documents that minimize redundancy.

In this work, we present an analysis of queries that may potentially benefit from result diversification and propose a query classification scheme from the perspective of diversification requirements of web queries. Our hypothesis is that queries for which users clicked many different URLs in the past may potentially benefit from diversification. We use click entropy [17] to identify such queries. Click entropy has been used previously to identify ambiguous queries [20] and queries that can potentially benefit from personalization [18] and diversification [6]. Based on a manual analysis of high click entropy 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. We also report results of automatic query classification experiments where we show how a query can be classified into one of the four above classes.

2. RELATED WORK

The work reported in this paper is related to search result diversification, query log analyses and web query classification. Since there exists a large body of work dealing with each of these problems, it is impossible to provide a comprehensive survey of all such works due to space considerations. In this section, we provide an outline of some of the representative research that is most closely related to our work.

2.1 Search Result Diversification

Maximum Marginal Relevance (MMR) [4] introduced by Carbonell and Goldstein represents one of the earliest at-

*Work done while the author was an intern at Yandex Labs.

DDR '12, Feb. 12, 2012, Seattle WA, USA.
.

tempts 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 [5] argue that the strategy of returning as many relevant results as possible (the *Probability Ranking Principle (PRP)*) 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. [1] 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 [7] describe an axiomatic framework that can be used for designing and characterizing diversification mechanisms. Santos et al. [14] 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. [21] 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. [15] propose a supervised selective diversification approach that trades off relevance and diversity on a per query basis.

2.2 Query Log Analysis

Web search engine transaction logs (or 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. [16]. 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 [13] 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. [2] 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 [10] 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. The analyses of search logs presented in this paper differs from previous works in that we analyze the logs to identify how many queries can benefit from diversification methods, what different types of diversification strategies should the search engines use and how much can search result diversification methods benefit the users.

2.3 Query Classification

There have been many works on web query classification where queries are classified into certain target categories depending upon the application at hand. Broder [3] 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 [11] 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 [12] where participants were required to classify 800,000 web search queries into 67 predefined topical categories. Gravano et al. [8] 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 [20] is most similar to our work in 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 and url level information for query classification.

3. DATA DESCRIPTION

We used roughly six months (179 days, from 17^{th} March 2011 to 11^{th} September 2011) of query logs of a commercial search engine. 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.

In the search logs used in this work, user sessions have already been identified. A session, as defined in the logs, consists of all the queries issued by the same user on a single day. There are roughly 49 million such unique sessions in the query logs we used and average session length is 7.83 queries per session. Note that this average length can be attributed to the long time duration of each session as well as many sessions containing thousands of queries corresponding to queries issued by automated bots. In order to filter such sessions, we only consider sessions containing ≤ 100 queries.

Query Class	Condition	Number of Unique Queries	Number of Times Query Issued
Low-Frequency, Low-Entropy (LFLE)	Frequency $\leq$ 100, Entropy $\leq$ 3	1,958,351 (2.24%)	44,183,993 (11.83%)
Low-Frequency, High-Entropy (LFHE)	Frequency $\leq$ 100, Entropy $>$ 3	338,076 (0.39%)	10,734,720 (2.87%)
High-Frequency, Low-Entropy (HFLE)	Frequency $>$ 100, Entropy $\leq$ 3	66,177 (0.08%)	78,998,631 (21.15%)
High-Frequency, High-Entropy (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.

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
Session Statistics	
Total Number of sessions	49,424,821
Mean session length (number of queries)	7.83
Total Number of sessions with frequency$\leq$ 100	49,368,180
Reformulation Statistics	
Number of unique queries that were reformulated in a session	8,113,711
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
Number of reformulations	23,449,703

Table 1: Characteristics of the query log data.

4. QUERIES THAT MAY BENEFIT FROM DIVERSIFICATION

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. Do different types of query differ in their diversity requirements? If yes, what are these types?

In order to find answers to these questions, we first use *click entropy* [17] to identify queries for which different users have clicked different URLs in the past. Click entropy has been used previously to identify ambiguous queries [20] and queries that can potentially benefit from personalization [18] and diversification [6].

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

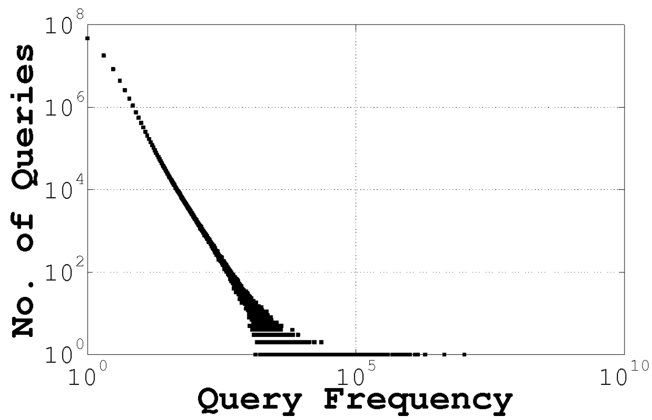

Figure 1: Plot showing distribution of query frequencies in the query logs. The distribution follows a power law with $\alpha = 1.16$.

$$CE(q) = \sum_{d \in D_q} -P(d|q) \log_2 P(d|q) \quad (1)$$

Here, D_q is the set of documents/URLs clicked by various users for query q.

A higher click entropy indicates that users selected different documents for the given query indicating that the query was used by users looking for different information 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.

Next, 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. Figure 2 shows a scatter plot between query frequency and query click entropies 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. We then divided the plot into four quadrants based on frequency and entropy values of queries. We chose a threshold frequency of 100 and a threshold entropy of 3. Table 2 summarizes some other statistics about queries in each of the four quadrants. Queries in the LFLE class account for 2.24% of all the unique queries in the logs and appear

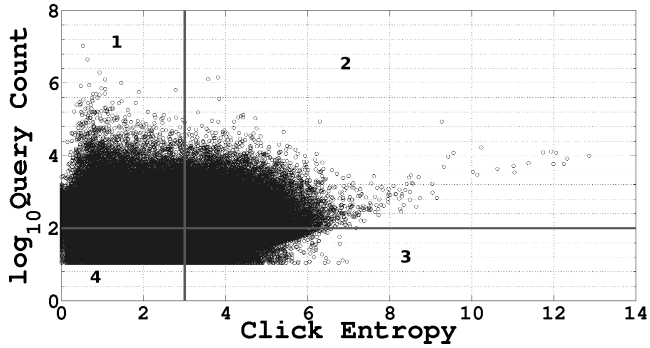

Figure 2: Scatter plot showing query frequency and associated click entropy as observed in the query logs. The plot is divided into four quadrants (see text for details). Quadrants marked as 1, 2, 3 and 4 refer respectively to HFLE, HFHE, LFHE and LFLE quadrants.

roughly forty four million times in the query logs (11.83%). A large faction 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 specific website names. Queries belonging to LFHE class are also generally quite specific. The reason for the high entropy values is due to the fact these queries are generally "literatue survey" type queries – 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`, `katie morgan` etc. 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.). 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). 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. Next, in order to come to a classification scheme for web queries from a diversity perspective, we randomly sampled and analyzed queries from the HFHE and LFHE classes. 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.

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 *"madonna"*. Here there is no ambiguity in what the query means but still it is not clear what the user wants to know about madonna – does he want to watch the music videos, read news, find song lyrics, or purchase songs at the iTunes store? The user's intent is not 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 softwares 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.

5. 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.

For automatic query classification, we used features tabulated and defined in Table 3. The features used can be grouped into two classes: *Query Features*, that are derived from the query alone and *Click Features* that are derived using the click-through information about the query present in the search logs.

5.1 Data Preparation

We randomly selected 500 queries belonging to the HFHE category and asked three human evaluators to assign the queries into one of the four query classes as described above.

Feature	Description	Type
	Query Features	
QueryLength	Number of words in query	Numeric
QueryFrequency	Number of times query occurs in the search logs	Numeric
NumReformulations	Number of different reformulations for a query	Numeric
ReformulationsInSession	Total number of sessions in which the query is being reformulated	Numeric
Reform-Session-Ratio	Ratio of NumReformulations and ReformulationsInSession	Real
IsURL	Is there a url in the query	Binary
IsDownload	If the query contains the word download	Binary
IsIMG	If the query contains request for images	Binary
IsVID	If the query contains request for a video	Binary
IsPorn	Is the query a porn query	Binary
IsQuestion	If any of the 5W1H words present in the query	Binary
IsTV	If the query contains request for tv shows	Binary
IsFree	If the query contains the keyword *free*	Binary
	Click Features	
ClickFrequency	Total number of clicks for the query	Numeric
URLCount	Number of unique URLs that were clicked for the query	Numeric
Query-URL-CountRatio	Ratio of QueryFrequency and URLCount	Real
ClickEntropy	Click entropy of the query	Real
ClickSTD	Standard deviation of frequencies of URLs being clicked for the query	Real

Table 3: List of features used for classification and their description

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.

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%).

5.2 Results

We used implementation of different classifiers as provide by the Weka toolkit [9]. We experimented with a variety of supervised classification schemes including decision trees, SVM, multi-layer perceptron classifier, naíve bayes classifier and a logit model classifer. The performance of all the classifiers was comparable with the logit model classifier achieving the best performance. Due to space constraints, we only report results for the logit model based classifier in Table 5. We used stringent ten-folds cross validation for experiments and the results reported are averaged over the ten folds. Table 5 reports results for each class and Table 6 reports the confusion matrix for the four classes. We achieved an overall precision of 72.4% and a recall of 70.7%. We also note that the minimum F-Measure is achieved for class **A** (Ambiguous queries) and maximum F-measure is achieved for class **M** (Miscellaneous queries).

6. CONCLUSIONS AND FUTURE WORK

	A	U	I	M
A	26	12	1	5
U	7	127	29	3
I	1	49	99	1
M	1	19	5	69

Table 6: 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*.

In this work we reported results of our initial efforts towards finding an answer to following questions: *(i)* what fraction of web queries can be potentially benefited from diverse search results? and *(ii)* do web queries differ in their diversity requirements? If yes, what are these types? Our analysis of logs of a commercial search engine revealed that 0.53% (460,700) of all the unique queries are high entropy queries (HFHE+LFHE) and they account for 20.35% of all the query mass, i.e, one in five web queries can potentially benefit from search result diversification. Further, based on analysis of popular 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 described results of automatic query classification experiments where we were able to classify queries into four categories with an overall precision of 72.4% and recall of 70.7%. The focus of our ongoing and future research is to employ a larger and diverse set of features to improve query classification.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14. ACM, 2009.

[2] S. M. Beitzel, E. C. Jensen, A. Chowdhury,

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
A	0.591	0.022	0.743	0.591	0.658	0.910
U	0.765	0.278	0.614	0.765	0.681	0.805
I	0.660	0.115	0.739	0.660	0.697	0.867
M	0.734	0.025	0.885	0.734	0.802	0.898
Overall	0.707	0.147	0.724	0.707	0.709	0.855

Table 5: Classification results for the query classification task.

D. Grossman, and O. Frieder. 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, pages 321–328, New York, NY, USA, 2004. ACM.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36:3–10, September 2002.

[4] J. Carbonell and J. Goldstein. 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*, pages 335–336, New York, NY, USA, 1998. ACM.

[5] H. Chen and D. R. Karger. 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, pages 429–436, New York, NY, USA, 2006. ACM.

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

[7] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 381–390, New York, NY, USA, 2009. ACM.

[8] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 325–333, New York, NY, USA, 2003. ACM.

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

[10] B. J. Jansen and A. Spink. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Inf. Process. Manage.*, 42:248–263, January 2006.

[11] I.-H. Kang and G. Kim. 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, pages 64–71, New York, NY, USA, 2003. ACM.

[12] Y. Li, Z. Zheng, and H. K. Dai. Kdd cup-2005 report: facing a great challenge. *SIGKDD Explor. Newsl.*, 7:91–99, December 2005.

[13] N. Ross and D. Wolfram. End user searching on the internet: an analysis of term pair topics submitted to the excite search engine. *J. Am. Soc. Inf. Sci.*, 51(10):949–958, August 2000.

[14] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 881–890, New York, NY, USA, 2010. ACM.

[15] R. L. Santos, C. Macdonald, and I. Ounis. Selectively diversifying web search results. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1179–1188. ACM, 2010.

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

[17] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2010.

[18] J. Teevan, S. T. Dumais, and D. J. Liebling. 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, pages 163–170, New York, NY, USA, 2008. ACM.

[19] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122, New York, NY, USA, 2009. ACM.

[20] Y. Wang and E. Agichtein. 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, pages 361–364, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[21] M. J. Welch, J. Cho, and C. Olston. Search result diversity for informational queries. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 237–246, New York, NY, USA, 2011. ACM.