Identifying Ambiguous Queries in Web Search

Ruihua Song[1, 2], Zhenxiao Luo[3], Ji-Rong Wen[2], Yong Yu[1], and Hsiao-Wuen Hon[2]
[1] Shanghai Jiao Tong University, Shanghai China
[2] Microsoft Research Asia, Beijing China
[3] Fudan University, Shanghai China
Contact: rsong@microsoft.com

ABSTRACT
It is widely believed that some queries submitted to search engines are by nature ambiguous (e.g., java, apple). However, few studies have investigated the questions of "how many queries are ambiguous?" and "how can we automatically identify an ambiguous query?" This paper deals with these issues. First, we construct the taxonomy of query ambiguity, and ask human annotators to manually classify queries based upon it. From manually labeled results, we find that query ambiguity is to some extent predictable. We then use a supervised learning approach to automatically classify queries as being ambiguous or not. Experimental results show that we can correctly identify 87% of labeled queries. Finally, we estimate that about 16% of queries in a real search log are ambiguous.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Natural Language*

General Terms
Experimentation, Languages, Human Factors

Keywords
Ambiguous query, query classification, broad topics, Web user study

1. INTRODUCTION
Some technologies like personalized Web search and search results clustering aim to improve users' satisfaction towards ambiguous queries from different perspectives. However, there is no sufficient study on ambiguous queries identification. Questions like "what percentage of queries are ambiguous?" and "can we automatically determine whether a query is ambiguous?" are still open. If we can estimate the percentage of ambiguous queries, we would know how many queries will be influenced potentially by the query ambiguity oriented technologies. If we can further identify ambiguous queries automatically, it is possible to apply such technologies for a particular kind of queries, instead of for all. We will try to answer such questions in this paper.

Identifying ambiguous queries is challenging for three reasons. First, there is no acknowledged definition and taxonomy of query ambiguity. Many terms related to this concept, such as "ambiguous query," "semi-ambiguous query," "clear query," "general term," "broad topic," and "diffuse topic." These terms are confusing in our investigation. Second, it is uncertain whether most queries can be associated with a particular type in terms of ambiguity quality. Cronen-Townsend et al. [1] proposed to use the relative entropy between a query and the collection to quantify query clarity, but the score is not easily aligned to concepts in human's mind. Third, even if ambiguous queries can be recognized manually, it is not realistic to label thousand of queries sampled from query logs. So how can we identify them in an automatic way?

In this paper, we first construct taxonomy for query ambiguity from the literature. We then assess human agreement on query classification through a user study. Based on the findings, we take a supervised learning approach to automatically identify ambiguous queries. Experimental results show that our approach achieves 85% precision and 81% recall in identifying ambiguous queries. Finally, we estimate that about 16% of queries in the sampled search log are ambiguous.

Copyright is held by the author/owner(s).
WWW 2007, May 8-12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. TAXONOMY OF QUERIES
By surveying the literature, we summarize the following three types of queries from being ambiguous to specific.

Type A (Ambiguous Query): *a query that has more than one meaning;*

e.g. "*giant,*" which may refer to "*Giant Company Software Inc.*" (an internet security software developer), "*Giant*" (a film produced in 1956), "*Giant Bike*" (a bicycle manufacturer), or "*San Francisco Giants*" (National League baseball team).

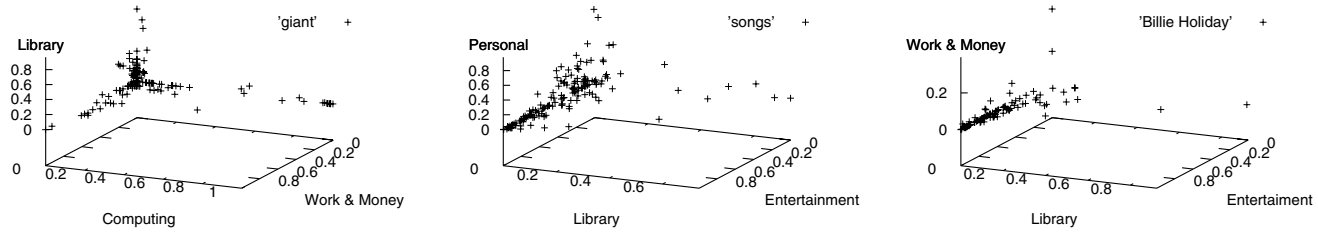
Type B (Broad Query): *a query that covers a variety of subtopics and a user might look for one of the subtopics by issuing another query.*

e.g. "*songs,*" which covers some subtopics such as "*song lyrics,*" "*love songs,*" "*party songs,*" and "*download songs.*" In practice, a user often issues such a query first, and then narrows down to a subtopic.

Type C (Clear Query): *a query that has a specific meaning and covers a narrow topic.*

e.g. "*University of Chicago*" and "*Billie Holiday.*" A clear query usually means a successful search in which a user can find several results with a high degree of quality in the first results page.

3. USER STUDY
The purpose of user study is to answer whether it is ever possible to associate a query with a certain type by looking at Web search results. Since it is difficult to find different meanings of a query by going through all the results, we use clustered search results generated by Vivisimo [5] to facilitate understanding the query.



(a) Type A: giant **(b) Type B: songs** **(c) Type C: Billie Holiday**

Figure 1. Projection of documents represented in categories for three example queries

Queries used in our user study are sampled from 12-day Live Search [4] query logs in August 2006. We use a total of 60 queries and involve five human subjects. Each participant is asked to judge whether a query is ambiguous (Type A) or not. If the query is not ambiguous, the participant would answer an additional question: "Is it necessary to add some words to the query in order to let it be clearer?" The question aims to clarify whether the query is broad (Type B) or clear (Type C).

The user study results indicate that participants are in general agreement, i.e. 90%, in judging whether a query is ambiguous or not. However, it is difficult to distinguish Type B from Type C as the agreement is only 50%.

4. LEARNING A QUERY AMBIGUITY MODEL

In this paper, we utilize a query q and a set of top n search results D with respect to the query in modeling query ambiguity. We formulate the problem of identifying ambiguous queries as a classification problem:

$$f(q, D) \mapsto (A \mid \overline{A})$$

Based on the findings in the user study, we aim to classify a query as A (ambiguous queries) or $\overline{A}$ (broad or clear queries). Support Vector Machines (SVM) developed by Vapnik [3] with RBF kernel is used as our classifier.

A text classifier similar to that used in [2] is applied to classify each Web document in D into predefined categories in KDDCUP 2005. We represent a document by a vector of categories, in which each dimension corresponds to the confidence that the document belongs to a category.

Our main idea of identifying an ambiguous query is that relevant documents with different interpretations probably belong to several different categories. To illustrate this assumption, we project documents into a three-dimensional (3D) space and show three example queries in Figure 1. The coordinates correspond to three categories that a query most likely belongs to. "*Giant*", as an ambiguous query, may refer to "*giant squid*" in Library category, "*Giant Company Inc.*" in Computing category, and "*Gaint Food supermarket*" in Work&Money category. Figure 1(a) shows scattered distribution among these three categories. "*Billie Holiday*" is a clear query and Figure 1(c) shows almost all the documents are gathered in the category of Entertainment. "*Songs*" is a broad query. A pattern of documents between being scattered and gathered is observed in Figure 1(b).

12 features are derived to quantify the distribution of D, such as the maximum Euclidean distance between a document vector and the centroid document vector in D.

5. EXPERIMENTS

We conduct the experiments of learning a query ambiguity model on 253 labeled queries. Five-fold cross validation is performed. The best classifier in our experiments achieves precision of 85.4%, recall of 80.9%, and accuracy of 87.4%. Such performance verifies that ambiguous queries can be identified automatically.

We try to estimate what percentage of queries is ambiguous in a query set sampled from Live Search logs. The set consists of 989 queries. To achieve the goal, our newly learned query ambiguity model is used to do prediction on the query set. When we increase the size of query set for estimation from 1/10 to 10/10, the percentage first vibrates between 15% and 18% and finally stabilizes at around 16%. Therefore, we estimate that about 16% of all the queries are ambiguous.

6. CONCLUSION

In this paper, we find people are in general agreement on whether a query is ambiguous or not. Thus we propose a machine learning model based on search results to identify ambiguous queries. The best classifier achieves high accuracy as 87%. By applying the classifier, we estimate that about 16% queries are ambiguous in the sampled logs.

7. REFERENCES

[1] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In Proceedings of the 25th ACM Conference on Research in Information Retrieval (SIGIR), pages 299-306, 2002

[2] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Q2c@ust: our winning solution to query classification in KDDCUP 2005. SIGKDD Explorations, 7(2):100–110, 2005

[3] V. Vapnik. Principles of risk minimization for learning theory. In D. S. Lippman, J. E. Moody, and D. S. Touretzky, editors, Advances in neural information processing systems 3, pages 831-838. Morgan Kaufmann, 1992

[4] Live Search. http://www.live.com/

[5] Vivisimo search engine. http://www.vivisimo.com