Determining the User Intent of Web Search Engine Queries

Bernard J. Jansen, Danielle L. Booth
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16801, USA
jjansen@acm.org, dlb5000@psu.edu

Amanda Spink
Faculty of Information Technology
Queensland University of Technology
Gardens Point Campus, 2 George St, GPO Box 2434
Brisbane QLD 4001 Australia
ah.spink@qut.edu.au

ABSTRACT
Determining the user intent of Web searches is a difficult problem due to the sparse data available concerning the searcher. In this paper, we examine a method to determine the user intent underlying Web search engine queries. We qualitatively analyze samples of queries from seven transaction logs from three different Web search engines containing more than five million queries. From this analysis, we identified characteristics of user queries based on three broad classifications of user intent. The classifications of informational, navigational, and transactional represent the type of content destination the searcher desired as expressed by their query. We implemented our classification algorithm and automatically classified a separate Web search engine transaction log of over a million queries submitted by several hundred thousand users. Our findings show that more than 80% of Web queries are informational in nature, with about 10% each being navigational and transactional. In order to validate the accuracy of our algorithm, we manually coded 400 queries and compared the classification to the results from our algorithm. This comparison showed that our automatic classification has an accuracy of 74%. Of the remaining 25% of the queries, the user intent is generally vague or multi-faceted, pointing to the need to for probabilistic classification. We illustrate how knowledge of searcher intent might be used to enhance future Web search engines.

Categories and Subject Descriptors
H.3.3 [1] Information Search and Retrieval – *Search process*

General Terms
Measurement, Experimentation, Human Factors

Keywords
User intent, Web queries, Web searching, search engines

1. INTRODUCTION
The Web has become an indispensable aspect in the lives of many people, and search engines are the main portal to the Web. Search engines are "the tool" for accessing the information, Internet sites, and services on the Web that many people use on a daily basis. Beyond their popularity, how are people using these Web search engines? How can we determine what these people are seeking? What task, goal, need, or intent are they trying to address with their Web searching?

Web search engines can help people find the resources they are looking for by more clearly identifying the searcher's intent behind the query. In this paper, we classify user searcher based on intent in terms of the type of content specified and operationalize these classifications with defining characteristics. We implement this operationalized classification in an application that automatically classifies queries from a search engine transaction log. We discuss how this model can be used to improve Web search engines.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. RELATED STUDIES
Discovering the intent of Web searchers is a growing research area. Some of the most initial work is from Broder [2] and Rose and Levinson [7]. Lee, Liu, and Cho [6] attempted automated classification, comparing only informational and navigational in order to simplify the problem. Baeza-Yates, Benavides, and Gonz´alez-Caro [1] use supervised and unsupervised learning to classify 6,042 Web queries as either *informational*, *not information*al, or *ambiguous*.

From a review of existing literature, efforts at classification of Web queries have usually involved small quantities of queries manually classified. There has been little effort on automated classification of queries for user intent. It is these issues that motivate our research. A comprehensive evaluation of a substantial set of Web searching queries will significantly enhance understanding user intent in Web searching.

3. RESEARCH OBJECTIVES
The following are our research objectives: (1) isolate characteristics of *informational*, *navigational*, and *transactional* for Web searching queries by identifying characteristics of each query type that will lead to real world classification. (2) Validate the taxonomy by automatically classifying a large set of queries from a Web search engine.

4. RESEARCH DESIGN
For research question one, we qualitatively analyzed samples of queries from seven Web search engine transaction logs [3, 5]. in order to identify characteristics for each query category. For the analysis, we selected random samples of queries and manually classified them in one of three categories (*information*, *navigational*, and *transactional*) as define in [2]. We then derived characteristics for each category that would serve to define the queries in that category. This was an iterative process with multiple rounds of "query selection – classification – characteristics refinement".

To address research question two, we implemented our characteristics in an algorithm (i.e., program), executed this program on a Web transaction log. The transaction log we used was from Dogpile.com (http://www.Dogpile.com/).[1] A complete statistical analysis of the Dogpile transaction log is presented in [4].

[1] We will make this log file available to the research community upon expiration of the NDA. Other search log files are available at http://ist.psu.edu/faculty_pages/jjansen/academic/transaction_logs.html.

5. RESULTS

For research question one, we derived the following characteristics for each category.

Navigational Searching

- queries containing company/business/organization/people names
- queries containing domains suffixes
- queries with "web" as the source
- queries length (i.e., number of terms in query) less than 3
- searcher viewing the first search engine results page

Transactional Searching

- queries containing terms related to movies, songs, lyrics, recipes, images, humor, and porn
- queries with "obtaining" terms (e.g., lyrics, recipes, etc.)
- queries with "download" terms (e.g., download, software, etc.)
- queries relating to image, audio, or video collections
- queries with "audio", "images", or "video" as the source
- queries with "entertainment" terms (pictures, games, etc.)
- queries with "interact" terms (e.g., buy, chat, etc.)
- queries with movies, songs, lyrics, images, and multimedia or compression file extensions (jpeg, zip, etc.)

Informational Searching

- uses question words (i.e., "ways to," "how to," "what is", etc.)
- queries with natural language terms
- queries containing informational terms (e.g., list, playlist, etc.)
- queries that were beyond the first query submitted
- queries where the searcher viewed multiple results pages
- queries length (i.e., number of terms in a query) greater than 2
- queries that do not meet criteria for navigational or transactional

Some navigational queries were quite easy to identify, especially those queries containing portions of URLs or even complete URLs. We also classified company and organizational names as navigation queries, assuming that the user intended to go to the Website of that company or organization. We also noted that most navigation queries were short in length and occurred at the beginning of the user session. Identification of transactional queries was primarily via term and content analysis, with identification of key terms related to transactional domains such as entertainment and ecommerce. With the relatively clear characteristics of navigational and transactional queries, information queries became the catch-all by default.

For research question two, we implemented our characteristics in a program. We then executed the program on the Dogpile search engine transaction log, with Table 1 presenting the results.

Table 1. Results from Automatic Classification of Queries

Classification	Occurrences	%
Informational	1,228,427	80.6%
Navigational	155,628	10.2%
Transactional	139,738	9.2%
	1,523,793	100.0%

Table 1 shows that more than 80% of Web queries were as informational in intent, with navigational and transactional queries each representing about 10% of Web queries. These results indicate a higher level of informational queries than reported in prior work. Broder [2] used a random of queries separate from the session, and Rose and Levinson [7] used only the first query in each session. These differences in data sampling may be responsible for the discrepancies in percentages with our work, which uses all queries from the user sessions.

6. CONCLUSION

In order for Web search engines to continue to improve, they must leverage an increased knowledge of user behavior, especially efforts to understand the underlying intent of the searchers. The results of this research demonstrate the ability to implement of an approach for automatically classifying queries. Our approach does not depend on external content and can be implemented in real time. This makes it a viable solution for Web search engines to classify user intent based on the type of content desired. Additionally, the larger data set provides more accurate percentages of user intent classification than smaller mostly manual studies. The higher percentage of information queries indicates that users view search engines primarily as information retrieval tools rather than instruments of navigation or commerce.

A limitation of our study is that we assigned each query to one and only one category. We are aware that a query may have multiple intents. However, from result of our research to verify the accuracy of our approach, it appears that approximately 75 percent of queries can be classified into a single category of intent (i.e., *informational*, *navigational*, or *transactional*) based on a manual coding of 400 queries. We are planning to investigate probability approaches such as naïve Bayes to arrive at a probability of classifying a query into one or more categories. Future work involves an both queries and sessions in order to identify more granular classifications of user intent (i.e. sub-categorizations of *informational*, *navigations*, and *transactional*). More targeted Web results to the underlying user content need will increase performance of future Web search engines.

ACKNOWLEDGMENT

We would like to thank Infospace.com for providing the data for this analysis. The AFOSR and the NSF funded portions of this research.

7. Reference

[1] Baeza-Yates, R., Calder´on-Benavides, L. and Gonz´alez-Caro, C. 2006. The Intention Behind Web Queries. In *Proceedings of STRING PROCESSING AND INFORMATION RETRIEVAL (SPIRE 2006).* Glasgow, Scotland, 98-109.

[2] Broder, A. 2002. A Taxonomy of Web Search. *SIGIR Forum. 36,* 2, 3-10.

[3] Jansen, B. J. and Spink, A. 2005. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management. 42,* 1, 248-263.

[4] Jansen, B. J., Spink, A., Blakely, C. and Koshman, S. forthcoming. Web Searcher Interaction with the Dogpile.com Meta-Search Engine. *Journal of the American Society for Information Science and Technology.*

[5] Jansen, B. J., Spink, A. and Saracevic, T. 2000. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing & Management. 36,* 2, 207-227.

[6] Lee, U., Liu, Z. and Cho, J. 2005. Automatic Identification of User Goals in Web Search. In *Proceedings of The World Wide Web Conference.* Chiba, Japan, 391-401.

[7] Rose, D. E. and Levinson, D. 2004. Understanding User Goals in Web Search. In *Proceedings of the World Wide Web Conference (WWW 2004).* New York, NY, USA, 13-19.