Classifying Web Queries by Topic and User Intent

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
jjansen@acm.org

Danielle Booth
College of Information Sciences and Technology
The Pennsylvania State University
nephari@gmail.com

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract
In this research, we investigate a methodology to classify automatically Web queries by topic and user intent. Taking a 20,000 plus Web query data set sectioned by topic, we manually classified each query using a three-level hierarchy of user intent. We note that significant differences in user intent across topics. Results show that user intent (*informational*, *navigational*, and *transactional*) varies by topic (15 to 24 percent depending on the category). We then use this manually classified data set to classify searches in a Web search engine query stream automatically, using an exact match followed by n-gram approach. These approaches have the advantage of being implementable in real time for query classification of Web searches. The implications are that a search engine can improve retrieval performance by more effectively identifying the intent underlying user queries.

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

ACM Classification Keywords
H.3.3 [1] Information Search and Retrieval – *Search process*

General Terms
Human Factors, Measurement

Query classification is difficult in the laboratory and is especially challenging on the Web due to:

- the highly dynamic nature of Web content, with nearly everyone with an Internet connection being able to publish
- the shifting population of user interests, that change from day to day and year to year
- the long tail of user interests and queries, with some very popular queries and some very specific ones
- the scale of users, queries, and content, with millions of users, billions of queries per week, and even more billions of pages of content
- the shortness of Web queries, typically about two or three terms, which makes categorization of queries difficult

Introduction

Understanding what the user is searching for and providing this content is at the heart of designing successful Web search applications. Therefore, a search engine that can successfully map incoming search queries to specific content can improve both the efficiency and the effectiveness Web searching.

There are many examples where such mapping can lead to significant improvements in retrieval performance. For example, successful identification of topic can help alleviate synonym issues (e.g., *cobra* the snake versus *cobra* the anime). Successful identification of user intent [3, 6] can provide the specific type of content in the proper format that the searcher desires.

Within this demanding operational environment, any query classification approach has to be achievable in real time in order to aid in Web searching. Therefore, many query classification techniques that rely on analysis of Web pages and other retrieved content are workable for post-classification analysis. However, they are not effective for real time implementation in Web searching. Many text classification approaches focus on documents where there is amble content to satisfactorily train classification algorithms. With Web queries being relatively short compared to documents, query classification is more difficult because there are very few inherent attributes.

In this research, we manually classify a 20,000-plus query set, already categorized by topic [2], with a user intent classification scheme. We then demonstrate the feasibility of using such a data set to automatically classify Web queries from a search engine transaction log.

Prior Research

Determining the topic of Web queries is an on-going area of study, with one avenue of research relying on retrieved Webpages to assist in topical classification [7]. Obviously, one can not use such an approach to classify topics prior to search engine retrieval, so researchers have explored methods that rely solely on the query [2].

A related area is determining user intent, defined as what type of content the user is seeking. A general consensus has formed around the paradigm proposed by Broder [3] of three broad classifications (*informational*, *navigational*, and *transactional*). Rose and Levinson [6] enhanced these classification with subcategories for informational and transactional queries. Jansen, Booth and Spink [4] present a comprehensive and integrated view of the different query intent taxonomies proposed in the literature.

Research Objectives and Approach

Our research objective is to: *Investigate classifying queries by topic and user intent into to automatically classify other queries.*

Prior work has examined user intent generically for all Web queries. Web queries vary by topical area, so our belief is that user intent would also probably vary by topic. We also wanted to not rely on Webpage or other external data that would not be available during the actual submission of the query, as we desired a method that would be implementable in real-time. Our approach was to code a set of queries with attributes

The 20K AOL dataset is a random sample of more than 20,000 from a search stream on the AOL search engine, sampled over one week. An AOL team of human assessors, manually classified these queries into a set of categories.

Topic	Query Count
Auto	691
Business	1,213
Computing	1,076
Entertainment	2,520
Games	475
Health	1,197
Holiday	325
Home	763
Misspellings	1,306
News	1,170
Organization	891
Other	3,313
Places	1,241
Porn	1,437
Research	1,354
Shopping	2,041
Sports	659
Travel	618
URL	1,356
Total	23,646

table 1. AOL Dataset.

This dataset has been used in prior works [1, 2] for topic classification.

and then leverage this enriched data set to classify other Web queries. In this paper, we primarily report the results from the manual classification.

Research Design

For our manually labeled data set, we used the 20,000 plus queries classified into general topical categories from AOL [2] (see side bar). With this topical label data set, we then manually labeled each query with user intent. In this research, we leveraged prior work reported in [4] (see table 2). We define user intent *as the expression of an affective, cognitive, or situational goal in an interaction with a Web search engine*. Rather than the goal itself, user intent is concerned with how the goal is expressed because the expression determines what type of resource the user desires in order to address their underlying need. Pirolli [5, p. 65] also makes a similar delineation between task (i.e., something external) and need (i.e., the concept that drives the information foraging behavior).

Levels	Examples of Queries
Level One	
(I) Informational: looking to obtain data or information	child labor law
(N) Navigational: looking for a specific URL.	capitalone
(T) Transactional: queries looking for resources that require another step to be useful	purchase Super Bowl tickets
Level Two	
(I, D) Directed: specific question	registering domain name
(I, U) Undirected: everything about a topic	singers in the 1980s
(I, L) List: list of candidates	things to do in hollywood ca
(I, F) Find: locate where some real world service or product can be obtained	pvc suit for overweight men
(I, A) Advice: advice, ideas, suggestions, instructions	what to serve with roast pork tenderloin
(N, T) Navigation to transactional : the URL desired is transactional	match.com
(N, I) Navigation to informational: the URL desired is informational	yahoo.com
(T, O) Obtain: obtain a specific resource or object	music lyrics
(T, D)Download: find a file to download	mp3 downloads
(T, R) Results Page: obtain a resource that one can find on search engine results page	(The 'answer' will be on the search engine results.)
(T, I) Interact: interact with program/resource	Purchase from ebay
Level Three	
(I, D, C) Closed: addresses one topic; question with one, unambiguous answer	9 supreme court justices
(I, D, O) Open: addresses two or more topics	the excretory system of arachnids
(T, O, O) Online: the resource will be obtained online	airline seat map
(T, O, F) Off-line: the resource will be obtained off	full metal alchemist wallpapers
(T, R, L) Links: the resources appears in the title, summary, or URL of one or more of the results on the search engine results page	(A user enters the title of a conference paper in order to locate the page numbers, which usually.)
(T, R, O) Other: the resources does not appear in one of the results but somewhere else on the search engine results page	(A user enters a query term to check for spelling.)

table 2. Definitions of classifications of web queries.

We classify the queries in the 20,000 data set into three levels of user intent using the coding scheme and characteristics in table 2, modified from [4]. The two researchers coded the queries, including 1,000 queries that both researchers coded. Intercoder reliability was good (Cohen's κ=0.81).

Results and Discussion

We examined user intent at three levels by topic. Results from the level one analysis are presented in table 3. A cross tab analysis clearly shows that there are differences among the topical categories (Chi-Square (136) = 9,764.1, p=0.01). The crosstabs method is a statistical test for measurements of association and agreement for nominal data.

As we can see from table 3, there are certainly topically categories that tend toward *informational*, *navigational*, or *transactional*. We have bolded the topical categories that have *navigational* or *transactional* percentages above of approximately 25 percent. As noted in prior research, most Web queries are *informational* in nature [3, 4, 6].

Business queries are clearly mostly *navigational* (51.9%), as are Holidays queries at 50.8%, and Organizational at 72.1%. For these queries, the searchers may be looking for information; however, they apparently have a preconceived Web destination in mind concerning where to look for this information. In fact, with the increase in aggregator sites on the Web (e.g., Wikipedia, YouTube, and Flicker), this push toward navigational queries may increase.

For transactional topics, we see that pornography has far and way the highest percentage of *transactional* queries (62.3%). Other topics with high percentages of *transactional* queries were Computing (27.7%), Games (24.8), and Shopping (35.0%). These four topics are all geared toward obtaining products or services and typically have a high commercial intent. Although Web queries are *informational* generally, we see that some topics are predominantly *informational*, such as Heath at 89.6%, Auto at 81.2%, and Entertainment at 79.7%.

	Info.	**Nav.**	**Trans.**
Auto	**81.2%**	15.8%	3.0%
Business	47.4%	**51.9%**	0.7%
Computing	**60.5%**	11.8%	**27.7%**
Entertainment	**79.7%**	6.1%	14.2%
Games	65.5%	9.7%	**24.8%**
Health	**89.6%**	8.9%	1.4%
Holiday	48.3%	**50.8%**	0.9%
Home	**60.9%**	21.0%	18.1%
News	**50.9%**	35.1**%**	14.0%
Organization	25.0%	**72.1%**	2.9%
Other	**55.6%**	26.1%	18.3%
Places	**62.9%**	31.1%	6.0%
Porn	11.6%	26.1%	**62.3%**
Research	**51.3%**	32.9%	15.8%
Shopping	33.4%	31.7%	**35.0%**
Sports	**51.7%**	30.2%	18.1%
Travel	47.4%	41.9%	10.7%
URL	0.1%	**99.2%**	0.7%
Average	51.3%	33.5%	15.3%
Sd Dev	22.9%	23.7%	15.5%
Max	89.6%	99.2%	62.3%
Min	0.1%	6.1%	0.7%

table 3. Results from Level One User Intent Classification. (Notable percentages **bolded**.)

Prior work in automatic classification by user intent has focused on level one or subsets of level one intent. In this work, we expand efforts into level two classification.

	Information					Navigational		Transactional			
	Directed	Undirected	List	Find	Advice	Info.	Trans.	Obtain	Download	Results	Interact
Auto	4.2%	23.0%	4.6%	**49.1%**	0.1%	2.7%	13.2%	2.3%	0.6%		0.1%
Business	2.9%	17.4%	**21.7%**	2.3%	3.4%	2.0%	**49.6%**	0.6%			0.1%
Computing	3.4%	**26.4%**	9.8%	11.3%	9.7%	0.7%	11.1%	7.3%	18.5%		1.9%
Entertainment	**54.1%**	14.3%	8.2%	1.3%	1.7%	0.2%	6.2%	0.7%	12.3%		1.0%
Games	9.7%	**53.1%**	0.6%	0.2%	3.4%	9.7%		**23.4%**			
Health	5.8%	0.8%	4.7%	1.2%	5.8%	**72.6%**	8.9%	0.3%	0.8%		
Holiday	6.2%	23.1%	**17.5%**	0.6%	0.9%	0.3%	50.5%	0.6%	0.3%		
Home	2.0%	22.0%	8.8%	11.0%	**17.2%**	17.7%	3.3%	16.1%	1.4%		0.5%
News	17.6%	18.4%	8.5%	3.8%	2.4%	29.1%	6.1%	7.2%	2.6%	0.3%	4.0%
Organization	6.4%	4.8%	8.2%	5.2%	0.6%	20.2%	**51.9%**	2.4%			0.4%
Other	16.0%	**29.1%**	6.6%	7.1%	4.8%	15.7%	10.4%	7.5%		0.6%	2.2%
Places	16.8%	**20.6%**	8.8%	15.8%	0.9%	11.1%	20.0%	0.7%	2.1%	2.2%	1.0%
Porn	5.2%		2.0%	2.4%	1.9%	25.7%	0.3%	3.3%	**57.7%**		1.4%
Research	**26.5%**	3.8%	10.9%	4.5%	5.5%	5.2%	27.7%	4.3%	5.5%	2.3%	3.7%
Shopping	5.0%	10.0%	5.0%	9.2%	4.2%	**29.5**%	1.1%	**27.9%**	3.6%	1.9%	2.6%
Sports	12.4%	**24.7%**	2.7%	8.2%	3.6%	15.2%	15.0%	2.9%	3.6%	1.4%	**10.2%**
Travel	4.4%	14.6%	12.1%	12.3%	4.0%	**31.6%**	10.5%	1.6%	0.2%	5.7%	3.1%
URL	0.3%	0.3%				**43.6%**	**48.8%**	0.9%	3.8%		2.3%
Average	11.0%	18.0%	8.3%	8.6%	4.1%	18.5%	19.7%	6.1%	8.1%	2.0%	2.3%
Sd Dev	12.7%	12.8%	5.4%	11.4%	4.1%	18.5%	18.7%	8.2%	15.2%	1.8%	2.5%
Max	54.1%	53.1%	21.7%	49.1%	17.2%	72.6%	51.9%	27.9%	57.7%	5.7%	10.2%
Min	0.3%	0.3%	0.6%	0.2%	0.1%	0.2%	0.3%	0.3%	0.2%	0.3%	0.1%

table 4. Results from Level Two User Intent Classification. (Notable percentages **bolded**.)

In table 4, we present the percentages of level 02 user intent, again with notable percentages bolded. Again, a cross tab analysis clearly shows that there are significant differences among the topical categories (Chi-Square (204) = 39,943.8, p=0.01).

Table 5 presents the percentage of topical queries at the level 03 classification. The topics with notable percentages are bolded. Cross tab analysis results are again significant differences (Chi-Square (119) = 11839.0, p=0.01) among the topics.

Automatically Labeling Web Queries

We are leveraging this manually labeled data set to automatically determine the topic and user intent of Web queries. As an initial effort, we automatically classify several thousand queries using an exact match (i.e., if the query in the stream exactly match one in the label set, then that query was matched with the same topic and intent).

However, give the long tail of Web queries (i.e., very infrequent queries), we also use an n-gram labeling approach, with each term as a gram in a sequence.

A probabilistic modeling approach, n-grams are used for predicting the next item in some sequence and are (n-1) order Markov models, where n is the n of the gram (i.e., sub-sequence or pattern) from the complete sequence or pattern. An n-gram model predicts state x_i using states x_{i-1}, x_{i-2}, x_{i-3}, …x_{i-n}. Therefore, the probabilistic model is: $P(x_i \mid x_{i-1}, x_{i-2}, x_{i-3}, \ldots x_{i-n})$, given the assumption that the next state only depends on the last *n - 1* states, which is, yet again, a (n-1) order Markov model. In our case, we use n-gram to predict user intent.

This mixed approach was noted by [2] in attempting to classify queries by topic.

	Informational: Directed		Transactional: Obtain		Transactional; Results	
	Closed	Open	On line	Off line	Links	Other
Auto	**62.1%**	37.9%				
Business	39.5%	**52.6%**	5.3%		2.6%	
Computing	**73.0%**	27.0%				
Entertain.	**99.6%**	0.4%				
Games	**100%**					
Health	**99.6%**	0.4%				
Holiday	22.7%	**77.3%**				
Home	0.7%	**60.9%**		**37.7%**	0.7%	
News	10.2%	**86.1%**		2.7%	1.0%	
Org.	6.5%	**85.7%**		7.8%		
Personal Finance	6.8%	**69.9%**		20.5%	2.8%	
Places	6.8%	**69.9%**		20.5%	2.8%	
Porn	3.3%	**76.0%**		20.7%		
Research	17.3%	**72.9%**		2.9%	0.7%	6.3%
Shopping	2.4%	**91.4%**		1.0%	0.1%	5.1%
Sports	30.9%	**55.5%**		5.5%	4.5%	3.6%
Travel	11.3%	**35.2%**		4.2%	2.8%	46.5%
URL						
Average	34.9%	56.2%	5.3%	12.3%	2.0%	15.4%
Sd Dev	37.2%	28.7%		12.0%	1.4%	20.8%
Max	100.0%	91.4%	5.3%	37.7%	4.5%	46.5%
Min	0.7%	0.4%	5.3%	1.0%	0.1%	3.6%

table 5. Results from Level Three User Intent Classification.

Conclusion

There are several important implications of this research. First, it shows that the general classifications scheme of *informational*, *navigational*, and *transactional* queries vary across topics. This insight can be leveraged to better classify underlying user intent and improve the performance of Web search engines. Second, it shows that the user of a labeled dataset has the potential to classify portions of the query stream of Web search engines in real-time and using no external sources, such as Web directories or retrieved documents. Therefore, this approach can work in the operational environment that Web search engines operate. Our next step is to evaluate the effectiveness of the approach in a working system.

References

[1] Beitzel, S. M., Jensen, E. C., Chowdhury, A. and Frieder, O. Varying approaches to topical web query classification. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (Amsterdam, The Netherlands, 23 - 27 July, 2007), 783 - 784.

[2] Beitzel, S. M., Jensen, E. C., Lewis, D. D., Chowdhury, A. and Frieder, O. Automatic classification of Web queries using very large unlabeled query logs *ACM Transactions on Information Systems*, 25, 2 (2007), Article No. 9.

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

[4] Jansen, B. J., Booth, D. and Spink, A. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management*, 44, 3 (2008), 1251-1266.

[5] Pirolli, P. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, Oxford, 2007.

[6] Rose, D. E. and Levinson, D. Understanding User Goals in Web Search. In *Proceedings of the World Wide Web Conference (WWW 2004)* (New York, NY, USA, 17–22 May, 2004), 13-19.

[7] Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J. and Yang, Q. Query enrichment for web-query classification *Transactions on Information Systems*, 24, 3 (2006), 320 - 352.