Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 1251–1266

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Determining the informational, navigational, and transactional intent of Web queries

Bernard J. Jansen [a,*], Danielle L. Booth [a], Amanda Spink [b]

[a] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA
[b] Faculty of Information Technology, Queensland University of Technology, Gardens Point Campus, 2 George St., GPO Box 2434, Brisbane, QLD 4001, Australia

Received 22 May 2007; received in revised form 30 July 2007; accepted 31 July 2007
Available online 11 September 2007

Abstract

In this paper, we define and present a comprehensive classification of user intent for Web searching. The classification consists of three hierarchical levels of informational, navigational, and transactional intent. After deriving attributes of each, we then developed a software application that automatically classified queries using a Web search engine log of over a million and a half 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 results from this manual classification to the results determined by the automated method. This comparison showed that the automatic classification has an accuracy of 74%. Of the remaining 25% of the queries, the user intent is vague or multi-faceted, pointing to the need for probabilistic classification. We discuss how search engines can use knowledge of user intent to provide more targeted and relevant results in Web searching.

Keywords: User intent; Web queries; Web searching; Search engines

1. Introduction

The World Wide Web (Web) has become an indispensable tool in the daily lives of many people, and search engines provide critical access to Web resources. With nearly 70% of Web searchers using a search engine as their point of entry, the major search engines receive millions of queries per day and present billions of results per week in response to these queries (Sullivan, 2006). Search engines are 'the tool' that many people use on a daily basis for accessing the information, Internet sites, services, and other resources on the Web. Although popular, how are people using Web search engines to accomplish their intended goal? How can we determine what it is that these people are actually seeking? What task, need, or goal are these people trying to address with their Web searching?

* Corresponding author. Tel.: +1 814 865 6459.
E-mail addresses: jjansen@acm.org (B.J. Jansen), dlb5000@psu.edu (D.L. Booth), ah.spink@qut.edu.au (A. Spink).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2007.07.015

Belkin (1993) states that one can classify searching episodes in terms of (1) goal of the interaction, (2) method of interaction, (3) mode of retrieval and (4) type of resource interacted with during the search. Web searching certainly possesses these aspects, so Web searching has continuity with earlier searching interactions, such as library systems. However, Web searching differs in three respects (i.e., context, scale, and variety), making it a unique domain of study. The first difference is that the direct availability of content accessible on the Web is nearly ubiquitous. Web search engines provide access to textual and multimedia content in a wide variety of settings including both home and work, as well as in mobile situations. Second, there is the number of searchers attempting to access this content via Web search engines. The scale of topics submitted by these users is surely unparalleled in pre-Web end user searching. Third, the variety of content, users, and systems is certainly unique. This combined diversity on the Web in both content and users is extreme.

In response to this diversity, Web search engines service a variety of purposes for users. In addition to satisfying information problems, modern Web search engines are navigational tools to take users to specific uniform resource locators (URLs) or to aid in browsing. People use search engines as applications to conduct e-commerce transactions, such as with sponsored search or Google's payment system. Search engines provide access to content collections of images, songs, and videos rather than directly addressing an information need with a specific object. Search engines provide access to transactional services such as maps, online auctions, driving directions, or even other search engines. Search engines perform social networking functions, as with Yahoo! Answers. Web search engines are spell checkers, thesauruses, and dictionaries. They are games, such as Google Whacking or vanity searching. Modern Web search engines are adding an increasing diverse range of features. Providers are placing more and highly varied content and services on the Web. In response, people are employing search engines in new, novel, and increasing diverse ways.

It is this cornucopia of alternatives where Web search engines differ most from classic information search and pre-Web retrieval systems. Referring back to facets outlined by Belkin, the method of interaction has remained the same (i.e., enter query, retrieve results, scan results, view results, refine query as needed). The mode of retrieval is similar, albeit within a hypermedia environment (Marchionini, 1995). In terms of goals and type of resources, however, the changes are dramatic. In fact, the facets of goals and range of resources are classic examples of the long tail effect of the Web. Namely, the Web has extended significantly both the range of search goals for people and the range of resources available (Anderson, 2006), and these resources need not be informational. We refer to the type of resource desired in the user's expression to the system as user intent. Within this great diversity, Web search engines can better assist people in finding the resources they are looking for by more clearly identifying the intent behind the query.

In this research, we developed a methodology to classify user intent in Web searching. We categorized user searches based on intent in terms of the type of content specified by the query and other user expressions, and we operationalized these classifications with defining characteristics. We implemented these catagories in a program that automatically classified Web search engine queries. We discuss how one can use this approach to improve Web search engine performance by provide more results in line with searchers' underlying intent.

The next section presents related research concerning modeling Web queries.

2. Related studies

Research aimed at discovering the intent of Web searchers is a growing field of Web focus. Determining the underlying intent of user searches has the potential to drastically improve system performance of Web search engine (Gisbergen, Most, & Aelen, 2007), with impact in the areas of information retrieval, data mining, and e-commerce. User intent research falls into three sub-areas, which are: (1) empirical studies and surveys of search engine use, (2) manual analysis of search engine transaction logs, and (3) automatic classification of Web searches. We discuss each in the following sub-sections.

2.1. User studies examining user intent on the Web

Several researchers have examined elements of user intent on the Web using a variety of controlled studies, surveys, and direct observation. Given the hypermedia environment of the Web, browsing has received a lot of attention. Carmel, Crawford, and Chen (1992) distinguished three types of browsing: (1) *search-oriented browsing* which is the process finding information relevant to a fixed task; (2) *review browsing* which is the process of scanning to finding interesting information, and (3) *scan browsing* which is the process of scanning to find information with no reviewing or integration involved. Marchionini (1995) articulated similar browsing patterns as directed browsing, semi-directed browsing, and undirected browsing.

Others have looked at the how users approach searching and how they implement it. O'Day and Jeffries (1993) outlined three broad search strategies, which are monitoring, following a plan, and exploring. Navarro-Prieto, Scaife, and Rogers (1999) categorized searching tasks as fact finding and exploratory. Byrne, John, Wehrle, and Crow (1999) developed a 'taskonomy' of Web tasks. Choo, Betlor, and Turnbull (1998) developed a behavior model of Web searching defining tasks as formal search, informal search, monitoring, and undirected viewing. Morrison, Pirolli, and Card (2001) classified searching into the categories of find, explore, monitoring, and collect. Even in this early work, we see a growing list of labels for very similar approaches to searching.

From a focus on tactics, research moved to classifying user goals. Rozanski, Bollman, and Lipman (2001) developed the categories of single mission, do it again, quickies, information please, loitering, just the facts, and surfing. Chi, Pirolli, Chen, and Pitkow (2001) examine computational methods for relating user needs to actions using information scent. Sellen, Murphy, and Shaw (2002) classified information seeking as finding, information gathering, browsing, and transacting. Bodoff (2004) did a classification of Web searching user goals.

In a return to some of the earlier browsing research, Teevan, Alvarado, Ackerman, and Karger (2004) discuss teleporting queries, defined as when a person attempts to go directly to an information target. The researchers viewed users engaged in teleporting as wanting to get to the 'vicinity' of the information in question and then searching locally to find the particular desired content. The researchers report that the study participants utilized keyword search in 39% of their searches, despite usually knowing their information need up front.

Recently, researchers have begun to quantify how often certain types of user searching occur. For example, Kellar, Watters, and Shepherd (2007) conducted a field study of 21 participants in which they recorded logs of the Web usage of the participants. In the area of information seeking, the researchers identified the tasks of fact finding, in both active and passive manner, information gathering, browsing, and transactions. Fact finding tasks accounted for 18%. Information gathering tasks accounted for 13% of Web usage.

2.2. Analysis of search logs

Rather than relying on empirical lab or panel studies, other researchers have used search logs from actual Web search engines or survey results from actual Web search engine users engaged in real Web searching contexts.

Broder (2002) proposed three broad user intent classifications of navigational, informational, and transactional for Web queries. Using survey results, Broder reported that approximately 73% of queries were informational, nearly 26% were navigational, and an estimated 36% were transactional. The researcher placed some queries into multiple categories. Based solely on the log analysis, Border reports that 48% of the queries were informational, 20% navigational and 30% transactional. We assume the remaining 2% were unclassifiable or the result of rounding.

Spink and Jansen (2004) report that e-commerce-related queries varied from approximately 12% to 24% using various Web search engine transaction logs. Jansen, Spink, and Pedersen (2005) stated that there appeared to be a significant use of search engines as a navigation appliance. The researchers report that the top 15 queries from a 2002 AltaVista search log (i.e., google, yahoo, ebay, yahoo.com, hotmail, hotmail.com, thumbzilla, www.yahoo.com, babelfish, mapquest, nfl.com, nfl, weather, www.hotmail.com, and google.com) were all likely expressions of a navigational intent. It is apparent that the hypermedia environment of the Web provides a unique capability of using searching a specialized form of browsing.

Rose and Levinson (2004) classified search queries using the categories of informational, navigational, and resource, with hierarchical sub-categories of each. The researchers investigated using just the searcher's query, the results the searcher clicked on, and subsequent queries in determining the user intent classification. Rose and Levinson (2004) reported that approximately 62% of the queries were informational, 13% navigational, and 24 percent resource. The researchers report only small differences in results when using the additional information beyond the query.

2.3. Automatic query classification

The analyses of search logs mentioned above were all performed manually, but some researchers have attempted automatic classification of user intent. Lee, Liu, and Cho (2005) automatically classified informational and navigational queries using 50 queries collected from computer science students at a US university. Their success rate for all 50 queries was 54%. Kang and Kim (2003) attempted to classify queries as either topic or homepage. After several iterations of classification, the researchers reported a classification rate of 91 percent finding using selected TREC topics (50 topic and 150 homepage finding) and portions of the WT10g test collection. However, query classification using retrieved Web documents has been shown to be an impractical approach when dealing with millions of queries (Beitzel, Jensen, Lewis, Chowdhury, & Frieder, 2007).

Dai et al. (2006) examined classifying whether or not a Web query has a commercial intent, noting that 38% of search queries have commercial intention. Baeza-Yates, Caldeŕon-Benavides, and Gonźalez (2006) used supervised and unsupervised learning to classify 6,042 Web queries as either informational, not informational, or ambiguous, achieving precision of classification of about 50%. Nettleton, Calderon, and Baeza-Yates (2006) used 65,282 queries along with click stream data and clustered these queries based on various parameters. Based on expected parameters, the researchers then label these clusters as informational, navigational, or transactional.

2.4. Synthesis of prior work

From a review of existing literature, we identified several trends. First, there have been a bewildering number of classifications of intent for similar or related Web searching. Second, the majority of the work has been lab studies with little use of actual Web transaction logs. Third, efforts at classification of Web queries have usually involved small quantities of queries manually classified. Fourth, there has been little effort on automatically classifying large numbers of Web queries for user intent. Finally, there has been little discussion of what is actually meant by user intent or what the theoretical underpinnings of the concept are.

In order to compare results across studies and move the field forward, a set of common identifiers for various types of user intent must be utilized. In fact, there must be some agreement on what intent actually is. To complement the various lab and panel studies, there must be an increase in the use of search log data where researches can validate classes of intent identified in the lab. Finally, although manual classification has been beneficial, we must explore automated methods in order to have direct impact on system design.

These issues motivate our research. A comprehensive review of prior work and an evaluation of a substantial set of Web searching queries will significantly enhance the understanding of user intent in Web searching. Deriving the underlying user intentions during Web search is critical for the further advancement of Web systems.

In the next section, we present our research objectives. We follow with a description of our research design and data analysis. We then present our results, along with a discussion of these results. We conclude with directions for future research and implications for the design of Web searching systems.

3. Research objectives

The research objectives are described below:

1. *Develop a comprehensive classification of Web searching user intent.*
   For research objective one, we analysed prior work in the area along with an analysis of numerous actual Web searching transaction logs in order to develop a detail categorization of Web searching based on user

intent. Given the plethora of categories and classifications, it is difficult to compare results across studies and research experiments. Such a comparison is vitally needed in order to place new research within prior work and to provide a foundation for future studies.

2. *Operationalize the taxonomy of informational, navigational, and transactional for Web searching queries by identifying characteristics of each query type that will lead to real world classification.*
For research objective two, we isolated characteristics of queries in each category (i.e., of informational, navigational, and transactional) that can serve as identifiers for these types of queries in operational search engines using various search logs. Although these classification have been isolated manually (c.f. Broder, 2002; Rose & Levinson, 2004), the criteria for determining each as not been articulated. In order for the classifications to be meaningful, one must isolate defining characteristics that one can operationalize to inform the design of future searching systems.
3. *Implement the informational, navigational, and transactional taxonomy by automatically classifying a large set of queries from a Web search engine and measure the effectiveness of the classification.*
For this research objective, we encoded the characteristics of informational, navigational, and transactional that we identified from research objective two to develop an automatic classifier. We executed the program on a transaction log from a Web search engine containing approximately one and half million queries from several hundred thousand users.

In order to measure the effectiveness, we manually classified a sub-set of queries as informational, navigational, and transactional, and we compared the results to those obtained via the automated method presented in research objective three. This provided a measure of the accuracy of the automatic classifier.

In the next section, we describe our research process in detail.

4. Research design

4.1. Classification of Web searching

For research objective one, we performed a comprehensive review of prior work in the area of user intent in Web searching. We cross correlated reported results from these studies to align user intent classes that were similar but variously labeled. We also supplemented this literature review by using results from our own data analysis. From this review and analysis, we derived a comprehensive categorization of Web searching intent and correlated this categorization with prior published works.

For the purpose of this research, we define user intent as the affective, cognitive, or situational goal as expressed in an interaction with a Web search engine. Referring to Belkin's states of a searching episode (1993), intent is akin to goal, and expression akin to method of interaction. Unlike goal, however, 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 overall goal. Pirolli (2007, p. 65) makes a similar delineation between task (i.e., something external) and need (i.e., the concept that drives the information foraging behavior). Saracevic's stratified model (1996, 1997) proposes that user expressions to an information searching system are based on affective, cognitive, or situational strata.

Certainly, the query is a key component of this expression of intent. The importance of the query is obvious by the considerable amount of research examining various aspect of query formulation, reformulation and processing (Belkin, Cool, Croft, & Callan, 1993; Belkin et al., 2003; Cronen-Townsend, Zhou, & Croft, 2002; Efthimiadis, 2000). Pirolli (2007, p. 65) refers to the query also as external representation of the need. We note that the query is many times an inexact representation of the underlying intent (Belkin, 1980; Croft & Thompson, 1987; Ingwersen, 1996; Taylor, 1968).

However, the query is not the only expression possible or that one can use to determine intent. Therefore, in this research, we examine other aspects of the interaction including number of query reformulations, selection of vertical, use of system feedback, and result page viewed as expressions of intent. This approach has much in common with research on implicit feedback (Jansen, 2005, 2006; Jansen & McNeese, 2005; Kelly & Belkin, 2001, 2004; Kelly & Teevan, 2003; Oard & Kim, 2001), where one attempts to use other expressions of the user as forms of relevance judgments.

4.2. Characteristics of Web queries

For research objective two, we qualitatively analysed samples of queries from seven Web search engine transaction logs from three Web search engines in order to identify characteristics for various user intent categories. Aggregate statistics on these logs are report in Jansen and Spink (2005b) and Jansen et al. (2000). The Web transaction logs used in this research are shown in Table 1.

For this process, we selected random samples of records containing not only the query but also other attributes such as the order of the query in the session, query length, result page, and vertical. These fields provided attributes beyond the query terms in order to assist in the classification. For the analysis, we manually classified the queries in one of three categories (informational, navigational, and transactional). Derived from work in Rose and Levinson (2004), we define the intent within each category as:

- *Informational searching:* The intent of informational searching is to locate content concerning a particular topic in order to address an information need of the searcher. The content can be in a variety of forms, including data, text, documents, and multimedia. The need can be along a spectrum from very precise to very vague.
- *Navigational searching:* The intent of navigational searching is to locate a particular Website. The Website can be that of a person or organization. It can be a particular Web page, site or a hub site. The searcher may have a particular Website in mind, or the searcher may just 'think' a particular Website exists.
- *Transactional searching:* The intent of transactional searching is to locate a Website with the goal to obtain some other product, which may require executing some Web service on that Website. Examples include purchase of a product, execution of an online application, or downloading multimedia.

We then derived characteristics for each informational, navigational, and transactional 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'. We then classified sub-classification for of these major categories. These sub-classifications were derived using both prior work and a priori using open coding technique which takes a grounded theory approach (Strauss & Corbin, 1990) to deriving categories. By utilizing seven transactions logs from three Web search engines, we believe that we obtained results that are generalizable across multiple search engines and user demographic populations.

4.3. Automatic classification of Web queries

To address research objective three, we used the characteristics from research objective two to develop an automatic classifier, and we then executed this program on a Web transaction log.

The transaction log we used for this research objective was from Dogpile.com (http://www.dogpile.com/). A complete statistical analysis of the Dogpile transaction log is presented in Jansen, Spink, Blakely, and Koshman (2006). The results indicate the user searching characteristics are consistent with those observed on other Web search engines, such as those reported in Jansen and Spink (2005b), Park, Bae, and Lee (2005) Silverstein,

Table 1
Web search engine transaction logs used

Web search engine	Year of data collection	Unique user identities	Queries
Excite	1997	18,113	51,473
Excite	1997	211,063	1,025,908
Excite	1999	325,711	1,025,910
Excite	2001	262,025	1,025,910
AlltheWeb	2001	153,297	451,551
AlltheWeb	2002	345,093	957,303
AltaVista	2002	369,350	1,073,388
		1,684,652	5,611,443

Henzinger, Marais, and Moricz (1999). Therefore, we expect the classifications to be also similar to other Web search engines.

For data collection, we logged searches executed on Dogpile.com on 6 May 2005. The original search log contained 4,056,374 records, representing a portion of the searches executed on that date.[1] Each record contained several fields, including:

- *User identification*: A user code automatically assigned by the Web server to identify a particular computer.
- *Cookie*: An anonymous cookie automatically assigned by the Dogpile.com server to identify unique users on a particular computer.
- *Time of day*: Measured in hours, minutes, and seconds as recorded by the Dogpile.com server.
- *Query terms*: Terms exactly as entered by the given user.
- *Source*: The content collection that the user selects to search (e.g. Web, Images, Audio, or Video) with Web being the default.

We imported the original flat ASCII transaction log file of 4,056,374 records into a relational database. We then generated a unique identifier for each record. We then used the fields of *Time of day*, *User identification*, *Cookie*, and *Query* to locate the initial query and recreate the chronological series of actions in a session.

Since we were interested only in queries submitted by humans and the transaction log also contained queries from agents, we removed all the agent submissions that we could identify using an upper cut-off similar to that used in prior work (c.f. Silverstein et al., 1999). We used an interaction cut-off to be consistent with the approach taken in previous Web searching studies (Jansen & Spink, 2005a; Jansen et al., 2005; Spink & Jansen, 2004) that was substantially greater than the mean search session (Jansen, Spink, & Saracevic, 2000) for human Web searchers. This approach certainly introduced some agent or common user terminal sessions; however, it also ensured that we had included most of the queries submitted primarily by human searchers.

Web search engine logging systems of Web search engines usually record result pages viewing as separate records with an identical user identification and query, but with a new time stamp (i.e., the time of the second visit). This permits the calculation of results page viewings, but it also introduces duplicate query records that skew the query calculations. To account for this, we collapsed the search log using user identification, cookie, and query. We calculated the number of identical queries by user, storing in a separate field within the transaction log. This collapsed transaction log provided us the data by user for analysing user queries without skewing by result list viewing. We also removed all records with null queries. After processing the transaction log, the database contained 1,523,793 queries from 534,507 users (identified by unique IP address and cookie) containing 4,250,656 total terms.

We then used the program we create to classify each query according to the characteristics developed in research question two. The algorithm for the classification was:

Algorithm: *Web Query Classification based on User Intent*

Assumptions:

1. Transaction log is sorted by IP address, cookie, and time (ascending order by time).
2. Search engine result page requested are removed.
3. Null queries are removed.
4. Queries are primarily English terms.

Input:

Record R_i with IP address (IP_i), cookies (K_i), query Q_i, source S_i, and query length QL_i;

Record R_{i+1} with IP address (IP_{i+1}), cookies (K_{i+1}), query Q_{i+1}, source S_{i+1}, and query length QL_{i+1}.

I: conditions of information query characteristics

[1] We expect to make this search engine transaction log available to the research community once the current non-disclosure agreement expires and upon successful negotiation with Infospace.

N: conditions of navigational query characteristics
T: conditions of transactional query characteristics
Variable: B: Boolean // (if query matches conditions, 'yes' else 'no')
Output: Classification of User Intent, C

begin
Move to R_i (this module establishes the initial boundary condition)
Store values for IP_i, K_i, Q_i, Fi, and QL_i
Compare (IP_i, K_i, Q_i, Fi, and QL_i) to N
  If B then $C = N$
  Elseif Compare (IP_i, K_i, Q_i, Fi, and QL_i) to T
  If B then $C = T$
  Elseif Compare (IP_i, K_i, Q_i,Fi,and QL_i) to I
  If B then $C = I$
While not end of file
  Move to R_{i+1}
  Compare (IP_i, K_i, Q_i, Fi, and QL_i) to N
  If B then $C = N$
  Elseif Compare (IP_i, K_i, Q_i, Fi, and QL_i) to T
  If B then $C = T$
  Elseif Compare (IP_i, K_i, Q_i, Fi, and QL_i) to I
  If B then $C = I$
(R_{i+1} now becomes R_i)
Store values for R_{i+1}as IP_i, K_i, Q_i, S_i, and QL_i
end loop

To address the effectiveness of classification, we selected a random sample of 400 queries from the Dogpile transaction log and manually classified these queries. We use a Delphi approach, where each evaluator independently rated each query. The three evaluators met to come to an aggregate classification. Once all evaluators had agreed to a common classification for all queries, we then compared our manual classification results to the classification results from our program in order to evaluate the effectiveness of our algorithm.

5. Results

5.1. Research objective one

For research objective one (Develop a comprehensive classification of Web searching user intent), we present in Table 2 a three-level hierarchical taxonomy, with the top most level being informational, navigational, and transactional. Each of these level one categories has multiple level two classifications. Some classifications also can involve a third level classification.

Below this developed taxonomy, Table 2 presents user intent studies and their best-fit classification across studies. The blank spaces indicate gaps in prior work where the particular study did not address a specific type of intent. In other cases, the studies findings were not as specific as presented in Table 2. In these cases, the particular study classification crosses multiple categories.

Table 3 presents definitions of each of the classifications in the user intent taxonomy.

All query examples in Table 3 are from the Dogpile transaction log used in this research for automatic classification. These high level classifications are the same as presented by Broder (2002) and are similar to those reported by Rose and Levinson (2004). Prior work has dealt mostly with informational and navigation searching, with few works focusing on transactional searching. In our analysis, we have noted that informational searching has five subcomponents (directed, undirected, find, list, and advice), for which we used labels proposed by Rose and Levinson (2004).

Navigational searching appears to exhibit itself in two sub-categories, (navigation to a transactional site or navigation to an information site). From a Web search engines perspective, the goal is to the get the user to the

Table 2
Hierarchical classification of user intent as expressed by Web queries

Level	User intent classification													
Level 01	Informational						Navigational		Transactional					
Level 02	Directed		Undirected	Find	List	Advice	Navigation to trans-actional	Navigation to inform-ational	Obtain		Download		Search engine results page	Interact
Level 03	Closed	Open						Online	Off-line	Free	Not free	Links	Other	
Prior studies							*Corresponding labels*							
Carmel et al. (1992)							Browsing (search-oriented, review, scan)							
Navarro-Prieto, Scaife, Rogers (1999)	Fact finding			Exploratory										
Choo and Turnbull (2000)	Formal search	Informal search	Monitoring				Undirected viewing							
Morrison et al. (2001)	Find		Explore monitoring	Collect										
Rozanski et al. (2001)	Single mission do it again quickies		Information please loitering	Just the facts		Quickies	Surfing							
Sellen et al. (2002)	Finding			Information gathering			Browsing		Transacting					
Broder (2002)	Informational						Navigational		Transactional					
Bodoff (2004)							Browsing (navigating, current awareness, undirected, scanning)							
Rose and Levinson (2004)	Informational directed closed	Inform-ational directed open	Inform-ational undirected	Inform-ational locate	Inform-ational list	Inform-ational advice	Navigational		Resource obtain		Resource download		Resource interact	
Teevan et al. (2004)	Orienteering						Teleporting							
Kellar et al. (2007)	Fact finding (looking for specific information) Fact finding (monitoring)			Information gathering			Browsing		Transactions					

Table 3
Definitions of classifications of Web queries

Levels	Examples of queries
Level one	
• (I) Informational: queries meant to obtain data or information in order to address an information need, desire, or curiosity • (N) Navigational: queries looking for a specific URL • (T) Transactional: queries looking for resources that require another step to be useful	• Child labor law • Capitalone • Buy table clocks
Level two	
• (I, D) Directed: specific question • (I, U) Undirected: tell me everything about a topic • (I, L) List: list of candidates • (I, F) Find: locate where some real world service or product can be obtained • (I, A) Advice: advice, ideas, suggestions, instructions • (N, T) Navigation to transactional: the URL the user wants is a transactional site • (N, I) Navigation to informational: the URL the user wants is an informational site • (T, O) Obtain: obtain a specific resource or object • (T, D) Download: find a file to download • (T, R) Results page: obtain a resource that one can printed, save, or read from the search engine results page • (T, I) Interact: interact with program/resource on another Website	• Registering domain name • Singers in the 1980s • Things to do in hollywood ca • PVC suit for overweight men • What to serve with roast pork tenderloin • match.com • yahoo.com • Music lyrics • mp3 downloads • (The user enters a query with the expectation that 'answer' will be on the search engine results page and not require browsing to another Website) • Buy table clock
Level three	
• (I,D, C) Closed: deals with one topic; question with one, unambiguous answer • (I,D, O) Open: deals with two or more topics • (T, O, O) Online: the resource will be obtained online • (T, O, F) Off-line: the resource will be obtained off-line and may require additional actions by the user • (T, D, F) Free: the downloadable file is free • (T, D, N) Not free: the downloadable file is not necessarily free • (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 • (T, R, O) Other: the resources does not appear one of the results but somewhere else on the search engine results page	• Nine supreme court justices • The excretory system of arachnids • Airline seat map • Full metal alchemist wallpapers • Free online games • Family guy episode download • (As an example, a user enters the title of a conference paper in order to locate the page numbers, which usually appear in one or more of the results) • (As an example, a user enters a query term to check for spelling with no interest in the results listing)

appropriate Website. Naturally, from a user perspective, there may be follow-on goals once the user arrives at a particular destination. So, one can view navigational searching as an expression of an intermediate intent aimed at satisfying some larger searching goal.

Interestingly, transactional searching is extremely nuanced with four sub-categories (obtain, download, interact, and search engine results page). This last sub-category is fascinating because it shows the capabilities offered by modern Web search engines. This classification represents those searches for which the Web search engine results page is the final destination. For this type of the searching, the 'answer' appears directly on the search engine results page, such as suggestions for correct spelling or terms in the results title, URL, or snippet.

5.2. Research objective two

For research objective two (Operationalize the taxonomy of informational, navigational, and transactional for Web searching queries by identifying characteristics of each query type that will lead to real world classification.), we derived the following characteristics for each category.

5.2.1. 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; and
- searcher viewing the first search engine results page.

5.2.2. 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.); and
- queries with movies, songs, lyrics, images, and multimedia or compression file extensions (jpeg, zip, etc.).

5.2.3. 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; and
- 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. Although it may seem counter intuitive to some, it has been noted in prior work that many Web searchers type in portions of URLs into search boxes as a shortcut to typing the complete URL in the address box of a browser (Jansen et al., 2005). 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. Naturally, there may be other reasons for a user entering a URL or proper name. 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 e-commerce.

With the relatively clear characteristics of navigational and transactional queries, informational queries became the catchall by default. However, we did note characteristics that indicated informational searching. The most pronounced was the use of natural language phrases. Informational queries were also more likely to be lengthier and, sessions of informational searching were longer in terms of the number of queries submitted.

For each of these classifications, we developed databases of key terms relating to each. We employed this database of key terms in our automatic classifier. For conditional characteristics such as query length and session length, we used program variables.

5.3. Research objective three

For research objective three (Implement the informational, navigational, and transactional taxonomy by automatically classifying a large set of queries from a Web search engine and measure the effectiveness of the classification.), we implemented the attributes we derived in research question two in a program. We then executed the program on the Dogpile search engine transaction log, with Table 4 presenting the results.

Table 4 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. We find this a surprising high percentage of informational queries. Prior work has reported that navigational intent was significantly represented in Web searching (Broder, 2002; Jansen et al., 2005). For example, Broder (2002) reports navigational queries of 24% based on approximately 3,100 survey responses and 20% based on an analysis of 400 Web queries.

The low percentage of transactional queries is also surprising. Broder (2002) reports transactional queries of 36% based on survey responses and 30% based on the analysis of Web query. Jansen and Spink (2005b) report that e-commerce-related queries ranged from 12% to 24% based on analysis of approximately 2,500 queries from multiple transaction logs.

The variation in reported percentage of navigational and transactional queries may be related to the size of the samples used in prior studies (which were much smaller than we used in this research) and the power log distribution of Web queries. Jansen et al. (2005) reported on the most frequently occurring queries, so navigational queries may be more prevalent in the more frequently occurring queries than the entire distribution, especially those in the long tail. A similar effect may be happening with transactional queries. Rose and Levinson (2004) classified only the initial query in the session. These approaches may have led to the increased percentage of navigational and transactional queries.

For measuring the effectiveness of automatic query classification, we randomly selected 400 queries and manually classified them and compared the results to those obtained via automatic classification. The results are shown in Table 5.

Table 5 shows that approximately 26% of the 400 queries were misclassified by the automated method. Primarily, the algorithm under classified transactional and navigational queries and over classified informational queries. Assuming that these percentages hold throughout the dataset, informational queries would occur approximately 65%, navigational queries approximately 15 percent, and transactional queries about 20%. However, these percentages are based on an assumption that the manual classifications are correct, namely that a particular query, as an expression of a user need, has one and only one intent. Naturally, multiple users may use the same query as an expression of different underlying intent. This relates to our comment earlier concerning possible multiple intents with entering a URL or company name.

From our analysis and review of the datasets, about 70–80% of the queries can be classified into one category will a high degree of confidence. The remaining queries are more problematic and may represent multiple intents. This is where most of the misclassifications occurred. For example, we manually classified the query ‘oreo’ as a navigational query (assuming that the searcher wanted to go to the Oreo cookie Website).

Table 4
Results from automatic classification of Web queries

Level 01 classification	Occurrences	%
Informational	1,228,427	80.6
Navigational	155,628	10.2
Transactional	139,738	9.2
	1,523,793	100.0

Table 5
Error checking of automatic classification

Classification (manual)	Classification (automatic)	Occurrences	% of differences in classification	% of total sample
Transactional	Informational	47	45.6	11.8
Navigational	Informational	38	36.9	9.5
Informational	Navigational	15	14.6	3.8
Informational	Transactional	2	1.9	0.5
Transactional	Navigational	1	1.0	0.3
		103	100.0	25.8

However, one could also, with a lower probability, classify it as an informational query. Other examples include ‘zelda sheet music’, ‘italy government’, and ‘mothers day poem’. Each of these queries could have multiple underlying intents. This points to the need for a probabilistic classification for that least a sub-set of queries.

However, based on our analysis, it appears that this is a relatively small sub-set of Web searcher, approximately 25%. With an accuracy of nearly 75%, this research shows that automatic classification of user intent is achievable using data that is currently available to most Web search engines.

6. Discussion and implications

In this study, we employed a three-level classification of Web searching that is useful in identifying the intent of the searcher. This model is based on our own analysis and on prior published work, most notably that of Broder (2002) and Rose and Levinson (2004). However, Broder (2002) did not present a description of the process and metrics used to classify the queries. Similarly, Rose and Levinson (2004) also did not elaborate on the details of their classifications. In our work, we have operationalized each category. Therefore, the classifications are meaningful for use by Web searching systems and for other studies.

Additionally, this research demonstrates the ability to implement our approach for automatically classifying queries. Our automated approach achieved a 74% successfully classification rate. Comparing this with other attempts at automatic classifications, we see that this success rate is quite good. Lee et al. (2005) had a 54% success rate with 50 queries. Kang and Kim (2003) had a 91% success rate but used documents from a TREC test collection. Baeza-Yates et al. (2006) achieved an approximately 50% success rate after clustering queries. These prior works used much smaller data sets, had higher error rates, and did not classify informational, navigational, and transactional queries. Not only does our approach have a success rate better than that reported in prior work, it uses a much larger data set of queries, does not depend on external content, and can be implemented in real time. This makes it a viable solution for Web search engines as they attempt to provide relevant content to users.

In analysing our results, we are aware of certain limitations that may restrict the ability to generalize our conclusions. One issue is that the Dogpile user population may not be representative of Web search engine users in general. Therefore, their queries would not be representative of the general Web population. We would certainly like to apply our classification methods on data from other major search engines. This may also involve a qualitative analysis of newer transaction logs than the ones we used in this study. Perhaps such logs would provide increased clarity on characteristics of various user intents. However, Jansen and Spink (2005b) report that query characteristics across search engines are fairly consistent. Additionally, we derived our initial characteristics from seven other transaction logs from three other search engines. Therefore, we would expect similar results from other datasets.

Another limitation is that we assigned each query to one and only one category. We are aware that a query may have multiple possible intents. In fact, instead of a decision tree approach that arrives at a binary answer, further research will focus on investigating approaches such as naïve Bayes or data mining to arrive at a probability of classifying a query into one or more categories. However, from results of this research, it appears that approximately 75% of queries can be classified into a single category of intent (i.e., informational, navigational, or transactional) with a high degree of certainty.

Our findings are also limited by the inherent shortcoming of relying solely on data from transaction logs. Transaction logs are excellent for collecting large amounts of data from a large number of users engaged in real searching tasks. However, we do not have access to these users, so we can only infer their intent from the data available. It would be an exciting area of future research to conduct a laboratory study to gain further insight into the underlying intent of Web searchers. Such a laboratory study would be a good supplement to the transaction log research presented here.

The strengths of this study are the variety and quantity of the datasets employed. Broder (2002) and Rose and Levinson (2004) both used a very small number of queries and classified the queries manually, with no presentation of the metrics used. Lee et al. (2005) used 50 queries, and Kang and Kim (2003) used 200 queries. Baeza-Yates et al. (2006) used approximately 65,000 queries but clustered them before categorizing them. Our dataset had over one and half million queries. Therefore, our results are robust.

In terms of implications, the approach used in this research can be implemented for real time classification by search engines since it uses just the characteristics of the current user interaction and query. By identifying the user intent of Web queries in real time, Web search engines can provide more relevant results to searchers and more precisely targeted sponsored links. This is especially fruitful in the area of transactional queries. Assuming that transactional queries carry a higher commercial inclination, these would be the queries that online advertising would be most interested. For these users, Web search engines could more heavily weight results with commercial content or sponsored links, for example. Similarly, targeted actions could be taken for navigational and informational queries.

There are several areas for future research. As mentioned, a laboratory study would be a good complement to this log analysis. Such a laboratory study might be able to shed further light in how searchers express their underlying intent. Additionally, a detailed qualitative analysis on a search log from a major search engine might lead to more granular attributes of user intent. We would like to develop algorithmic approaches for utilizing this knowledge of user intent in order to provide searchers with more targeted results. Finally, we are aiming to expand our automated classification methods to include the more granular categories at level two and three.

7. Conclusion and further research

In order for Web search engines to continue to improve, they must leverage an increased knowledge of user behavior in order to identify the underlying intent of searchers. In this research, we highlighted characteristics of Web queries based on user intent. These characteristics were derived from an examination of Web queries from multiple search engine transaction logs. We have also demonstrated an automated method that can successfully classify Web queries based on user intent. Web search engines can use this knowledge for more precisely associating user goals with queries and thereby providing more targeted content. If Web search engines can determine search goals based on queries and other interactions, designers can leverage this knowledge by implementing algorithms and interfaces to help users achieve their searching goals.

Acknowledgements

We would like to thank Excite, AlltheWeb.com, AltaVista, and especially Infospace.com for providing the data for this analysis, without which we could not have conducted this research. We encourage other search engine companies to engage members of academic community in Web searching research. The Air Force Office of Scientific Research (AFOSR) and the National Science Foundation (NSF) funded portions of this research.

References

Anderson, C. (2006). *The long tail: Why the future of business is selling more of less*. New York: Hyperion.

Baeza-Yates, R., Calderón-Benavides, L., & González, C. (2006). In *The intention behind Web queries* (pp. 98–109). Paper presented at the string processing and information retrieval (SPIRE 2006), 11–13 October, Glasgow, Scotland.

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

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5*, 133–143.

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In *Information retrieval '93, Von der Modellierung zur Anwendung* (pp. 55–66). Konstanz, Germany: Universitaetsverlag Konstanz.

Belkin, N., Cool, C., Croft, W. B., & Callan, J. (1993). In *The effect of multiple query representations on information retrieval systems* (pp. 339–346). Paper presented at the 16th annual international ACM SIGIR conference on research and development in information retrieval.

Belkin, N., Cool, C., Kelly, D., Lee, H.-J., Muresan, G., Tang, M.-C., et al. (2003). In *Query length in interactive information retrieval* (pp. 205–212). Paper presented at the 26th annual international ACM conference on research and development in information retrieval, 28 July–1 August, Toronto, Canada.

Bodoff, D. (2004). Relevance for browsing, relevance for searching. *Journal of the American Society of Information Science and Technology, 57*(1), 69–86.

Broder, A. (2002). A taxonomy of Web search. *SIGIR Forum, 36*(2), 3–10.

Byrne, M., John, B., Wehrle, N., & Crow, D. (1999). In *The tangled Web we wove: A taskonomy of WWW use* (pp. 544–551). Paper presented at the human factors in computing systems: CHI 99, May 15–20, Pittsburgh, PA.

Carmel, E., Crawford, S., & Chen, H. (1992). In *Browsing in hypertext: A cognitive study* (pp. 865–884). Paper presented at the IEEE transactions on systems, man and cybernetics, 5–10 October, Chicago IL.

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). In *Using information scent to model user information needs and actions on the Web* (pp. 490–497). Paper presented at the ACM CHI 2001 conference on human factors in computing systems, 31 March–5 April, Seattle, WA.

Choo, C., & Turnbull, D. (2000). Information seeking on the web: An integrated model of browsing and searching. *First Monday*, 5(2). Available from <http://firstmonday.org/issues/issue5_2/choo/index.html>.

Choo, C., Betlor, B., & Turnbull, D. (1998). In *A behavioral model of information seeking on the Web: Preliminary results of a study of how managers and IT specialists use the Web* (pp. 290–302). Paper presented at the 61st annual meeting of the American society for information science, Pittsburgh, PA, ASIS.

Croft, W. B., & Thompson, R. H. (1987). I3: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science, 38*(6), 389–404.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). In *Predicting query performance* (pp. 299–306). Paper presented at the 25th annual international ACM SIGIR conference on research and development in information retrieval, 11–15 August, Tampere, Finland.

Dai, H. K., Nie, Z., Wang, L., Zhao, L., Wen, J. -R., & Li, Y. (2006). In *Detecting online commercial intention (OCI)* (pp. 829–837). Paper presented at the World Wide Web conference (WWW2006), 23–26 May, Edinburgh, Scotland.

Efthimiadis, E. N. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society of Information Science and Technology, 51*(11), 989–1003.

Gisbergen, M. S. V., Most, J. V. D., & Aelen, P. (2007). *Visual attention to online search engine results*. Market Research Agency De Vos & Jansen.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3–50.

Jansen, B. J. (2005). Seeking and implementing automated assistance during the search process. *Information Processing & Management, 41*(4), 909–928.

Jansen, B. J. (2006). Using temporal patterns of interactions to design effective automated searching assistance systems. *Communications of the ACM, 49*(4), 72–74.

Jansen, B. J., & McNeese, M. D. (2005). Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *Journal of the American Society for Information Science and Technology, 56*(14), 1480–1503.

Jansen, B. J., & Spink, A. (2005a). An analysis of Web searching by European Alltheweb.com users. *Information Processing & Management, 41*(2), 361–381.

Jansen, B. J., & Spink, A. (2005b). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 248–263.

Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2006). Web searcher interactions with the Dogpile.com meta-search engine. *Journal of the American Society for Information Science and Technology, 58*(4), 1875–1887.

Jansen, B. J., Spink, A., & Pedersen, J. (2005). Trend analysis of AltaVista Web searching. *Journal of the American Society for Information Science and Technology, 56*(6), 559–570.

Jansen, B. J., Spink, A., & 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.

Kang, I., & Kim, G. (2003). In *Query type classification for Web document retrieval* (pp. 64–71). Paper presented at the 26th annual international ACM SIGIR conference on research and development in information retrieval, 28 July–1 August, Toronto, Canada.

Kellar, M., Watters, C., & Shepherd, M. (2007). A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology, 58*(7), 999–1018.

Kelly, D., & Belkin, N. J. (2001). In *Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback* (pp. 408–409). Paper presented at the 24th annual international ACM SIGIR conference on research and development in information retrieval, New Orleans, Louisiana, United States.

Kelly, D., & Belkin, N. J. (2004). In *Display time as implicit feedback: Understanding task effects* (pp. 377–384). Paper presented at the 27th annual international conference on research and development in information retrieval, 25–29 July, Sheffield, United Kingdom.

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum, 37*(2), 18–28.

Lee, U., Liu, Z., & Cho, J. (2005). In *Automatic identification of user goals in Web search* (pp. 391–401). Paper presented at the World Wide Web conference, 10–14 May, Chiba, Japan.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge University Press.

Morrison, J. B., Pirolli, P., & Card, S. K. (2001). In *A taxonomic analysis of what world wide Web activities significantly impact people's decisions and actions* (pp. 163–164). Paper presented at the conference on human factors in computing systems (CHI '01), 31 March–05 April, Seattle, Washington.

Navarro-Prieto, R., Scaife, M., & Rogers, Y. (1999, July). *Cognitive strategies in Web searching*. Paper presented at the the 5th Conference on human factors and the web, Gaithersburg, Maryland.

Nettleton, D. F., Calderon, L., & Baeza-Yates, R. (2006). Analysis of Web search engine query and click data from two perspectives: Query session and document. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2006), Philadelphia, Pennsylvania.*

Oard, D., & Kim, J. (2001). In *Modeling information content using observable behavior* (pp. 38–45). Paper presented at the 64th annual meeting of the American society for information science and technology, 31 October–4 November, Washington, DC, USA.

O'Day, V., & Jeffries, R. (1993). In *Orienteering in an information landscape: How information seekers get from here to there* (pp. 438–445). Paper presented at the ACM InterCHI '93, Amsterdam, The Netherlands.

Park, S., Bae, H., & Lee, J. (2005). End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research, 27*(2), 203–221.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford: Oxford University Press.

Rose, D. E., & Levinson, D. (2004). In *Understanding user goals in Web search* (pp. 13–19). Paper presented at the World Wide Web conference (WWW 2004), 17–22 May, New York, NY, USA.

Rozanski, H. D., Bollman, G., & Lipman, M. (2001). Seize the occasion! The seven-segment system for online marketing. Retrieved 3 August 2006, Available from http://faculty.msb.edu/homak/HomaHelpSite/WebHelp/Online_Segmentation_S+B_Q4_2001.htm.

Saracevic, T. (1996). In *Modeling interaction in information retrieval (IR): A review and proposal: Vol. 33* (pp. 3–9). Paper presented at the 59th American society for information science annual meeting, 19–24 October, Baltimore, MD.

Saracevic, T. (1997). In *Extension and application of the stratified model of information retrieval interaction: Vol. 34* (pp. 313–327). Paper presented at the annual meeting of the American society for information science, 1–6 November, Washington, DC.

Sellen, A. J., Murphy, R., & Shaw, K. L. (2002). In *How knowledge workers use the Web* (pp. 227–234). Paper presented at the conference on human factors in computing systems (CHI '02), 20–25 April, Minneapolis, Minnesota, USA.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum, 33*(1), 6–12.

Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the Web*. New York: Kluwer.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage Publications.

Sullivan, D. (2006). Nielsen/NetRatings search engine ratings. Retrieved 1 June 2006, Available from http://www.searchenginewatch.com/reports/netratings.html (February 23).

Taylor, R. S. (1968). Question negotiation and information seeking in libraries. *College & Research Libraries, 28*, 178–194.

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). In *The perfect search engine is not enough: A study of orienteering behavior in directed search* (pp. 415–422). Paper presented at the CHI 2004, 24–29 April, Vienna, Austria.