Investigating customer click through behaviour with integrated sponsored and nonsponsored results

Bernard J. Jansen*

College of Information Science and Technology
University Park, Pennsylvania 16802, USA
Fax: +01–814–865–6426
E-mail: jjansen@acm.org
*Corresponding author

Amanda Spink

Faculty of Information Technology
Queensland University of Technology
2 George St., GPO Box 2434
Brisbane QLD 4001, Australia
E-mail: ah.spink@qut.edu.au

Abstract: This paper reports findings from a study investigating the effect of integrating sponsored and nonsponsored search engine links into a single web listing. The premise underlying this research is that web searchers are chiefly interested in relevant results. Given the reported negative bias that web searchers have concerning sponsored links, separate listings may be a disservice to web searchers as it might not direct them to relevant websites. Some web meta-search engines integrate sponsored and nonsponsored links into a single listing. Using a web search engine log of over 7 million interactions from hundreds of thousands of users from a major web meta-search engine, we analysed the click-through patterns for both sponsored and nonsponsored links. We also classified web queries as *informational*, *navigational* and *transactional* based on the expected type of content and analysed the click-through patterns of each classification. The findings show that for more than 35% of queries, there are no clicks on any result. More than 80% of web queries are *informational* in nature and approximately 10% are *transactional*, and 10% *navigational*. Sponsored links account for approximately 15% of all clicks. Integrating sponsored and nonsponsored links does not appear to increase the clicks on sponsored listings. We discuss how these research results could enhance future sponsored search platforms.

Keywords: sponsored search; web search engines; sponsored results; sponsored links; internet advertising; online auctions; paid search; organic results; internet marketing and advertising; customer click through behaviour.

Reference to this paper should be made as follows: Jansen, B.J. and Spink, A. (2009) 'Investigating customer click through behaviour with integrated sponsored and nonsponsored results', *Int. J. Internet Marketing and Advertising*, Vol. 5, Nos. 1/2, pp.74–94.

Biographical notes: Dr. Bernard J. Jansen (Jim Jansen) is an Assistant Professor in the College of Information Sciences and Technology at The Pennsylvania State University, USA. Jansen has more than 150 publications in the area of information technology and systems, with articles appearing in a multi-disciplinary range of journals and conferences. His specific areas of expertise are web searching, sponsored search, and personalisation for information searching. He is coauthor of the book, *Web Search: Public Searching of the Web* and coeditor of the book *Handbook of Weblog Analysis*. Jansen is a member of the editorial boards of six international journals. He has received several awards and honours, including an ACM Research Award and six application development awards, along with other writing, publishing, research, and leadership honours. Several agencies and corporations have supported his research. He is actively involved in teaching both undergraduate and graduate level courses, as well as mentoring students in a variety of research and educational efforts. He also has successfully conducted numerous consulting projects.

Amanda Spink is a Professor in the Faculty of Information Technology at the Queensland University of Technology and Co-Leader of the Information Science Cluster. Her primary research includes: basic, applied, industry and interdisciplinary studies in information science, information behaviour, cognitive information retrieval; web retrieval, including relevance, feedback and multitasking models. Professor Spink has published over 300 journal articles, refereed conference papers and book chapters, and five books. She is a member of numerous journal editorial boards including: *Information Processing and Management*, *Journal of Documentation*, *Journal of Information Systems Education and Webology*.

1 Introduction

From a user point of view, web search engines need to provide relevant and useful results in response to some user input, typically a query. Web search engines use dozens of factors in determining how to score relevance and to rank the retrieved results. Typically, the user has no idea what factors lead to a particular result being retrieved and ranked.

One exception is sponsored links. Many web search engines present at least two types of search listings on the web Search Engine Results Page (SERP). One type of links is the nonsponsored list, which the web search engine presents using proprietary algorithms. The other type of link is the sponsored list, which appears due to a bid on the phrase that the searcher entered as the query. Many major web search engines present sponsored and nonsponsored listings in separate areas on SERP. For this one (and usually only one category), searchers are told in broad terms how the result was retrieved. Why are other results not highlighted (*e.g.*, anchor text, page rank)? Certainly, the underlying motivation is to emphasise the economic incentive that caused the sponsored listing to appear. Research shows that web searchers have a bias against sponsored links (Greenspan, 2004). However, assuming that web searchers desire relevant and useful results in response to their queries and that sponsored links are as relevant as nonsponsored results, are web search engines doing searchers a disservice by highlighting sponsored links so prominently? What would be the effect of combining the sponsored and nonsponsored links in a single listing? This is the major research question that motivates this study.

This paper presents findings from a study analysing a web search engine transaction log of over 7 million records representing click through data for both sponsored and nonsponsored links. On the SERP of this web search engine the sponsored and nonsponsored links are integrated and shown as a single list. In this study, we first address the overall search characteristics of these web searchers. We then categorise the web queries based on user intent and then examine click through patterns of sponsored and nonsponsored listing from various perspectives. Finally, we discuss the implications for sponsored web search platforms and sponsored link presentation.

In the next section, we present the related research to our study.

2 Related studies

2.1 Sponsored search studies

Major web search engines such as Yahoo!, MSN, Google, and Ask have significantly altered online commerce. The unique characteristics of the web for sharing information, distributing content, engaging in ecommerce, and providing access to web-based retailing are fundamentally transforming the way in which people and commercial entities interact. Studies show that consumer attitudes towards attitudes towards web advertising can affect attitude towards specific brands (Lee and Miller, 2006).

Certainly, one of the most transforming innovations in web services since page rank (Brin and Page, 1998) is sponsored search. Sponsored search is the process where advertisers pay web search engines for traffic from the search engine to particular websites. The web search engines display these sponsored links in response to user queries alongside the nonsponsored links. There are various versions of the sponsored search model where there might be payment for just displaying the sponsored links or only when the user takes a specified action on the advertiser's website. The specific characteristics of the web for ecommerce and web-based retailing are alternating how vendor target consumers. Pachauri (2002) reviewed the streams of research in the ecommerce area and highlights future research questions for this ongoing transformation.

Focusing specifically on sponsored search, Battelle (2005) provides an overview of Google Adwords and Yahoo! Search Marketing and the factors that have led to the development of these sponsored search platforms. Fain and Pedersen (2006) discussed a history of sponsored search. Feng (2002) examined paid placement strategies for web search engines. Feng *et al*. (2007) presented the search engine mechanisms for implementing sponsored search. Liu and Chen (2006) analysed sponsored search as weighted unit-price-contract auctions. Feng (2002) discussed the role of gatekeepers in sponsored web search. The study reported in this paper examines the searcher interaction area of sponsored web search, specifically the effect of combining sponsored and nonsponsored links on web search engine SERPs. Our study represents a shift from prior research, which focused on efforts to separate more clearly the two categories of web search engine links. We instead examine if integration is appropriate.

An investigation sponsored by the Federal Trade Commission recommended that web search-engine companies clearly identify sponsored links on their websites (Hansen, 2002). The study reports that phrases such as '*Recommended Sites*', '*Featured Listings*', '*Premier Listings*', *or* '*Search Partners*' do not adequately inform people of the nature of the links. The report does pointed out even more ambiguous terms were *Products and*

Services, *News*, *Resources*, *Featured Listings*, or *Spotlight*. The report did not state why these were inadequate. The study concluded that when users suspect that web search engines are intentionally disguising the presence of sponsored links, these searchers might be less likely to consider them.

Interestingly though, empirical studies have shown that the typical web searcher has limited understanding of how web search engines retrieve, rank or prioritise any links on the results page (Marable, 2003). This includes both sponsored and nonsponsored links. Using data obtained during a user study, Marable (2003) reported that searchers did not realise that 41% of links on the SERP were sponsored listings. When informed of the nature of the sponsored listings, participants reported negative emotional reactions. The Pew Internet and American Life Project (Fallows, 2005) reported that 38% of searchers reported that they were aware of the distinction between sponsored links and nonsponsored links. Less than 17% of survey respondents reported that they could always distinguish between sponsored and nonsponsored links. This finding has not held up in lab studies, where participants seem capable of distinguishing between the two types of links (Hotchkiss, 2006).

Examining user perceptions, Hotchkiss (2004) used an enhanced focus group format to observe the web search behaviours of 24 participants and interviewed them for their reactions to what they saw on the SERP. The researcher reports that as the web search process becomes focused the likelihood that searchers will consider clicking on sponsored listings increases. Hotchkiss (2004) also reported four distinct types of searchers, and these search patterns affected the portion of SERP seen and the likelihood of conversion (*i.e.*, the searcher buying something). Novice searchers did have particular trouble identifying sponsored links, however. Half the study participants were suspicious that payments influence even the nonsponsored links. However, many web searchers visually ignored or did not see the sponsored listings, partly due to their screen location on the right side of the SERP.

In a follow-on study, Hotchkiss *et al*. (2004) surveyed 425 respondents who overwhelmingly preferred links that they viewed as offering sources of trusted unbiased information. More than 77% of participants also favoured nonsponsored links more than sponsored links. Even in an ecommerce-like scenario, survey respondents still choose nonsponsored over sponsored links. Similarly, Greenspan (2004) found that users preferred nonsponsored listings to sponsored links. The study also raised ethical issues regarding how web search engines present sponsored listings. Greenspan (2004) reported that web searchers are more likely to select sponsored listings with web search engines that do not clearly identify them as such, suggesting that they might not have selected them had they known these links were sponsored.

Examining the effect of sponsored link ranking, Brooks (2004a) established that the likelihood of a web searcher selecting a sponsored listing is a curvilinear function of its placement on the page (*i.e.*, its rank). The higher the link's placement in the results listing, the more likely a web searcher is to select that link. The study reported similar results with nonsponsored listings. Generally, the difference between the first position and the tenth position is a 20%–30% drop in click through (*i.e.*, customers visiting website by clicking on a link on the SERP) for the listing. In a related study, Brooks (2004b) reported that the conversion rate (*i.e.*, customers that actually bought something) dropped nearly 90% between the first and tenth position. There appears to be an intrinsic trust value associated with the rank of a listing as presented by the web search engine.

Dobrow (2004) reported that study participants are significantly more likely to recall the name of the company from a web search listing compared to a banner ad or tile ad. Saaksjarvi and Pol (2007) shows that consumers do not wish to click on any advertising in certain contexts. Therefore, even if study participants do not select the link, there may be some marketing benefit of the sponsored listing. Investigating loyalty and interaction with web search engines, iProspect Inc. (2004) surveyed 1649 web searchers. Of the respondents, 60% of Google users reported nonsponsored links to be more relevant than sponsored. This was even higher for predominantly Google users (70%). Frequent users of the web (four or more years of internet use) found nonsponsored listings to be more relevant than sponsored listings (65% to 56%). More women (43%) than men (34%) found sponsored listings to be generally relevant.

The study reported in this paper focuses on the effect of melding sponsored and nonsponsored links on the SERP, rather than highlighting their differences. This focus is important because prior research has established a potential disconnect between the perception of sponsored listings by businesses and searchers. Web searchers appear to be suspicious of sponsored links and may see these links as less relevant than nonsponsored links. Thus, they are less likely to select them. Yet, businesses see sponsored search as a lucrative marketing avenue. Commercial and other organisations spend billions on sponsored search, and this amount is expected to grow for the foreseeable future (Jarboe, 2005). However, sponsored links are currently primarily transactional nature. Businesses are trying to gauge the intent of the searcher in order to get qualified customers to their websites. Qualified customers are those that are interested in transactions, either now or sometime in the future.

2.2 Automatic query classification

A critical aspect of any online advertising campaign, including sponsored search, is an understanding of the underlying user need, goal, or intent. Not surprisingly, understanding the intent of web searchers is a growing research area. Broder (2002) surveyed 3190 users of and analysed 400 queries from AltaVista. Broder proposed three broad web queries classifications of *navigational*, *informational*, and *transactional*. Using survey results, Broder reported that approximately 73% of queries were informational, about 26% were navigational, and an estimated 36% were transactional, with some queries placed in multiple categories. From the log analysis, Broder (2002) reported that 48% of the queries were informational, 20% navigational and 30% transactional. It is unclear what the remaining 2% were. Rose and Levinson (2004) manually analysed approximately 1500 queries from AltaVista, classifying queries as informational, navigational, and resource with hierarchical sub-categories. Rose and Levinson (2004) reported that approximately 62% of the queries were informational, 13% navigational, and 24% resource.

While Broder (2002) and Rose and Levinson (2004) relied on manual classification, Lee *et al*. (2005) attempted automated classification of queries as informational and navigational. They used 30 queries, after eliminating 20 queries (40%) that they deemed 'unclassified'. Counting all 50, their success rate was 54%. Kang and Kim (2003) endeavour to classify queries as either topic (*i.e.*, informational) or homepage finding (*i.e.*, navigational) using selected Text REtrieval Conference (TREC) topics

(50 information and 150 navigational) and portions of the WT10g test collection. They report a classification rate of 91% overall, tailoring the algorithm for the particular set of queries and utilising documents within the test collection.

Dai *et al*. (2006) investigated classifying web queries in terms of their commercial intent, finding that the more frequent queries have the most commercial intent. Baeza-Yates *et al*. (2006) used supervised and unsupervised learning to classify 6042 web queries as either informational, not informational, or ambiguous, achieving precision classification of more than 50%. In a related study, Nettleton *et al*. (2006) used 65 282 queries and click stream data. The researchers then label these clusters as *information*, *navigational*, or *transactional*. The researchers conducted no verification of the classification. Finally, Jansen *et al*. (2008) used a binary tree algorithm to classify user intent, reporting that web users typically (80% of queries) have an informational intent.

2.3 Synthesis of prior work

Based on prior work, we note a small but increasing interest in sponsored search literature from a user and marketing perspective. Most of this research has focused on user perceptions of sponsored results or investigations into how relevant are sponsored links. This prior work has noted that web searchers have a bias against sponsored links and that sponsored links are at least as relevant to user queries as nonsponsored links. In terms of query classification, efforts at classification of web queries have usually involved small quantities of queries manually classified, but with a growing interest and effort in automatically classifying queries for user intent. Table 1 presents a synthesis of prior work.

Table 1 Synthesis of prior work

Prior work	*Major finding(s)*
Broder (2002)	Proposal of three broad areas of user intent (*e.g.*, informational, navigational, and transactional)
Rose and Levinson (2004)	Classified queries as informational, navigational, and resource with hierarchical sub-categories
Lee *et al*. (2005)	Automatically classified queries as informational and navigational
Dai *et al*. (2006)	Automatically classified queries as containing commercial intent
Nettleton *et al*. (2006)	Automatically classified queries as informational, navigational, and transactional
Kang and Kim (2003)	Automatically classified queries as informational and navigational
Jansen (2007)	Sponsored and nonsponsored search engine results are equivalent in terms of relevance
Jansen *et al*. (2008)	Automatically classified queries as informational, navigational, and transactional
Hotchkiss (2004)	Distinct types of searchers (or searching interaction) that affect which search engine results interacted with
Hotchkiss *et al*. (2004)	Searcher prefer trusted and information that they view as unbiased
Greenspan (2004)	Searchers prefer nonsponsored over sponsored results
Brooks (2004a–b)	Probability of a web searcher selecting a sponsored listing is a curvilinear function of its rank

However, limited prior studies have investigated the effects of separate listings for sponsored and nonsponsored links on user behaviour. None of the prior work focused on identifying user intent from the aspect of click through on sponsored links. Understanding the underlying user intent concerning sponsored links is critical for the further advancement of web advertising and commercial systems.

This synthesis of prior studies defines and motivates our research questions, which we present in the following section. Given the implications of sponsored web search as the predominant business model for web search engines, the results of our study could have substantial impact on the future development of and use of sponsored links. In the following section, we present our research questions. We follow with a description of our data set and data analysis methods. We then present our findings, along with discussion of these findings. We conclude with directions for future research and implications for the design of web searching systems.

3 Research objectives

The following are our research objectives:

- Investigate the click through rates of searchers when the combined sponsored and nonsponsored links are presented in a single listing on the SERP.

  We analysed a transaction log file from Dogpile, a web meta-search that combines both sponsored and nonsponsored links from multiple search engines into a single listing on the SERP.

- Compare differences in click though rates of informational, navigational, and transactional web queries when integrated sponsored and nonsponsored links are presented as a single listing on the SERP.

  We isolated characteristics of queries within each category that serve as identifiers for these types of queries in real world web searching. We then classify queries from the Dogpile transaction log and compare click though rates of sponsored and nonsponsored links among the three types of queries.

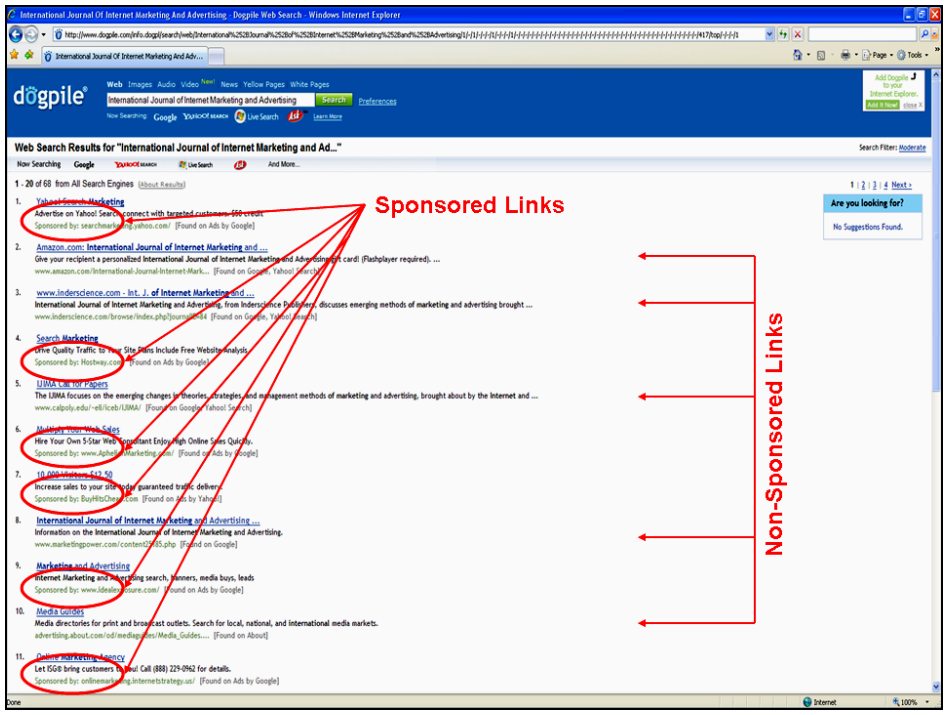
4 Research objectives

Dogpile.com[1] is a meta-search engine owned by Infospace, Inc. Nielsen/NetRatings[2] reports that Dogpile was the eighth most popular web search engine in 2006 as measured by number of searches executed. When a searcher submits a query, Dogpile concurrently submits the query to multiple web search engines, collects the results from each, removes duplicate results, and aggregates the remaining results into a combined listing using a proprietary algorithm.

Dogpile integrates the results of the four leading web search indices (*i.e.*, Ask, Google, MSN, and Yahoo!) along with other search engines into its search results listing. Meta-search engines provide a service by presenting the alternate results provided by the various search engines, which have a low rate of overlap (Ding and Marchionini, 1996). Dogpile has indexes for searching the *Web*, *Images*, *Audio*, and *Video* content, which searchers can access via tabs off the search engine interface. Dogpile also offers query

reformulation assistance with alternate query suggestions listed in an *Are you looking for?* area of the interface. Figure 1 shows the Dogpile interface with query box, tabbed indexes, and 'Are you looking for? features'.

Figure 1 Dogpile SERP with sponsored and nonsponsored results displayed (see online version for colours)



Referring to Figure 1, the integrated sponsored and nonsponsored listings are combined on the SERP. Infospace clearly states this integration to searchers. From Dogpile's Webpage concerning the sponsored listings:[3]

> The search results you get are a combination of the sponsored and non-sponsored results from the top engines on the Web. The specific ranking and mixture of sponsored and non-sponsored results will generally depend on the nature of your Web search.

4.1 Methodology for identifying characteristics of web queries

For research question one, we logged web searches executed on Dogpile on 15 May 2006. The original general transaction log contained 7 142 874 records, representing a portion of the searches executed on Dogpile that date.[4] 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.
- *Vertical* – the content collection that the user selects to search (*e.g.*, *Web*, *Images*, *Audio*, or *Video*) with *Web* being the default.
- *Sponsored* – whether or not the user click was on a sponsored link.
- *Organic* – whether or not the user click was on a nonsponsored link.
- *Rank* – the position in the results listing of the clicked link.

We imported the original flat ASCII transaction log file of 7 142 874 records into a relational database. We generated a unique identifier for each record. From the original transaction log, we removed records with NULL queries (*i.e.*, no search was executed) and records with corrupted data. We used four fields (*Time of day*, *User identification*, *Cookie*, and *Query*) to locate the initial query and then recreated the chronological series of actions by a user.

We define our searching terminology similar to that used in other web transaction log studies (Spink and Jansen, 2004).

- *Term* – a series of characters separated by white space or other separator.
- *Query* – string of terms submitted by a searcher in a given instance.
- *Query length* – the number of terms in the query. (Note: this includes traditional stop words.)
- *Session* – series of queries submitted by a user during one interaction with the web search engine.
- *Session length* – the number of queries submitted by a searcher during a defined period of interaction with the web search engine.

We were only interested in queries submitted by humans and the transaction log contained web queries from both human users and agents. Therefore, we removed all the agent submissions that we could identify using an upper cut-off similar to that used in prior work (Montgomery and Faloutsos, 2001; Silverstein *et al*., 1999). We used an interaction cut-off by separating all sessions with 100 or fewer queries into an individual transaction log to be consistent with the approach taken in previous web searching studies. This cut-off is substantially greater than the mean web search session (Park *et al*., 2005; Wang *et al*., 2003) for human web searchers. This increased the probability that we were not excluding any human searches. This cut-off probably introduced some agent or common user terminal sessions; however, we were satisfied that we had included most of the queries submitted primarily by human searchers.

Transaction log applications 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. It also introduces duplicate records that skew the queries' calculations. To correct for these duplicate queries, we collapsed the transaction log upon user identification, cookie, and query. We calculated the number of identical queries by user, storing in a separate field within the web transaction log. This collapsed transaction log provided us records by user for analysing user queries without skewing caused by the result list viewing.

After processing the transaction log, the database contained 1 874 397 queries from 666 599 users (identified by unique IP address and cookie) containing 5 455 449 total terms with 4 201 071 total interactions.

4.2 Methodology for automatic classification of web queries

To address research question two, we derived characteristics of *information*, *navigational*, and *transactional* queries, implementing our characteristics in an algorithm (*i.e.*, program), and executed this program on the web transaction log. The complete approach is presented in (Jansen *et al*., 2008), with an abbreviated version presented here.

To determine characteristics of web queries, we selected random samples of queries from seven transactions logs of three web search engines and manually classified them in one of three categories (*information*, *navigational*, and *transactional*). We then identified characteristics for each category that would serve to define the web queries in that category. This was an iterative process with multiple rounds of 'query selection – classification – characteristics refinement'. These high level classifications are the same as presented by Broder (2002) and are similar to those reported by Rose and Levinson (2004). We define the intent within each category as:

- *Informational web searching* – The intent of information web 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.
- *Navigational web searching* – The intent of web *navigational* searching is to locate a particular website. The website can be that of a person or organisation. It can be a particular web page, site or a hub website. The searcher may have a particular website in mind, or the searcher may just 'think' a particular website exists.
- *Transactional web searching* – The intent of web *transactional* searching is to locate a website with the goal to execute a service in order to obtain some other product. Examples include purchase of a product, execute a web service, or download multimedia.

By utilising seven transactions logs from three web search engines, the characteristics are generalisable across multiple web search engines and user demographic populations. The defining characteristics derived are:

1 Navigational web searching
    - queries containing company/business/organisation/people names
    - queries containing domains suffixes
    - queries length (*i.e.*, number of terms in query) less than three

- queries with 'web' as the source
- searcher viewing the first search engine results page

2 Transactional web searching

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

3 Informational searching

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

Some of the *navigational* queries were quite easy to identify, especially those web queries containing portions of Uniform Resource Locators (URLs) or even complete URLs. Although it may seem counter intuitive to some, 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. We also classified company and organisational names as navigation queries, assuming that the user intended to go to the website of that company or organisation. Naturally, there may be other reasons for a user entering a URL or proper. We also noted that most navigation queries were short in length and occurred at the beginning of the user session.

The 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, *informational* 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 more likely to be longer, and sessions of information searching were longer in terms of number of queries submitted. For each of these classifications, we developed databases of key terms relating to

characteristics of each classification. We employed these databases of key terms in our program to classify automatically the web queries. For conditional characteristics such as query length and session length, we used program variables. We then used the program we create to classify each query according to the characteristics developed in research objective one. The algorithm for the classification is:

Algorithm: *Web query classification based on user intent*

Assumptions:

1 Transaction log was sorted by IP address, cookie, and time (ascending order by time within each session).

2 Web search engine result pages requested are removed.

3 Null queries 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

N: conditions of information query characteristics

T: conditions of information query characteristics

Variable: B: Boolean // (if query matches conditions, 'yes' else 'no')

Output: Classification of user intent, C

begin

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

In the next section we present results, with preliminary results reported in (Jansen and Spink, 2007).

5 Results

We examined web search behaviour as recorded by the Dogpile log. Table 2 presents an aggregate statistical analysis of the data.

Table 2 Dogpile transaction log (15 May 2006) aggregate statistics

Users	666 599	
Queries	1 874 397	
Total interactions (Queries, Page Views, and Click Throughs)	4 201 071	
Terms		
Unique	360 174	6.6%
Total	5 455 449	
Mean terms per query	2.83	
Terms per query		
1 term	352 285	52.8%
2 terms	114 391	17.2%
3+ terms	199 923	30.0%
	666 599	100.0%
Users modifying queries	314 314	47.15%
Repeat queries (queries submitted more than once by two or more searchers)	152 771	11.6%
Unique queries (queries submitted only once in the entire data set)	1 159 764	88.4%
	1 312 535	100.0%
Session size		
1 query	352 285	52.8%
2 queries	114 391	17.2%
3+ queries	199 923	30.0%
	666 599	100.0%
Boolean queries	42 138	2.2%
Other query syntax	95 232	5.1%
Terms not repeated in data set (208 804 terms, which is 58% of the unique terms)	208 804	3.8%
Use of 100 most frequently occurring terms (100 terms which is 0.0003 of the unique terms)	1 011 729	18.5%
Use of other 126 208 terms (151 370 terms which is 42% of the unique terms)	5 246 645	96.2%
Unique term pairs (occurrences of terms pairs within queries from the entire data set)	2 753 468	

Comparing these web searching statistics to those reported in prior works (Park *et al*., 2005; Silverstein *et al*., 1999), the results indicate the user searching characteristics are consistent with those observed on other web search engines. The general characteristics present in the Dogpile and other web logs are short queries of three terms or less, short sessions of one or two queries, and a power law distribution of terms, with a small number of terms used quite often and a long tail of terms used very infrequently. The use of Boolean and other advanced query operators is minimal. Therefore, we would expect similar results concerning click through rates on other search engines as well.

5.1 Research objective 01

Table 3 presents results for research objective 01 (Investigate the click through rates of searchers when the combined sponsored and nonsponsored links are presented in a single listing on the SERP.), including general click through patterns.

Table 3 Proportion of clicks on sponsored and nonsponsored links

Interaction type	*Occurrences*	*% (including no clicks)*	*% (excluding no clicks)*
Sponsored	430 068	10.2	15.8
Organic	2 290 804	54.5	84.2
No click	1 480 199	35.2	
Total	4 201 071	100.0	
Total (discounting no clicks)	2 720 872		100.0

We examined clicks on the SERP from three perspectives, *clicks on sponsored links*, *clicks on nonsponsored links*, and *no clicks*. These are listed in Column 1 of Table 3. Column 2 (*Occurrences*) shows the raw number of incidents for each click type. Column 3 *(percentage including no click*) shows the percentages of each. Column 4 *(% excluding no clicks*) shows the percentage of clicks on sponsored and nonsponsored links if the no clicks are excluded from the total.

Column 4 of Table 3 shows that approximately 16% of the clicks where executed on the sponsored links. About 84% of the clicks were on the nonsponsored links. These percentages do not include the *no clicks* (*i.e.*, times where a user submitted a query but did not click on any result). Column 2 shows that there were *no clicks* approximately 35% of the time. Including *no clicks*, searchers clicked on a sponsored link approximately 10% of the query submissions and on a nonsponsored link about 54% of the time. Although 35% of visits to a web search engine not resulting in a click may seem high, according to comScore,[5] Dogpile had one of the highest click-through rates of any major web search engine in 2004. Typically, web search engines experience nonclick thorough rates of approximately 45%, based on private correspondence with search engine researchers.

We could locate no prior work on what overall percentage of queries resulted in a click on a sponsored link. As far as we could determine, the study presented here is the first published research accounting of actual click through rates of sponsored links from a major web search engine. As such, there is limited benchmarking that one can do. However, popular press reports that 25 to 30% of clicks are on sponsored links.[6] Results from user studies seem to confirm that approximately 30% of web searchers will click on sponsored listings over a series of searches (Jansen and Resnick, 2006). Naturally, there is a ranking bias of web search engine results (*i.e.*, the link at the top most rank is clicked more than the link at the second position on a typical SERP).

Using the reported figure of approximately 30%, it appears, perhaps counter intuitively, that integrating sponsored and nonsponsored links in the same listing does not raise overall sponsored link click through. However, it could also be that the popular press reports of click through rates are inflated. This seems highly likely. Based on calculations of click through rates versus impressions for many key terms on the Google and Yahoo! sponsored search platforms, the percentage of clicks on sponsored links appears lower than the 25% or 30% than has been reported. One of the overall implications of this study, supported with other collaborating factors, is that actual sponsored link click through rate is probably in the neighbourhood of 15% or lower, rather than the 25% to 30% reported elsewhere. Certainly, there may be certain key phrases that do achieve 30% and possible higher sponsored links click through rates.

5.2 Research objective 02

For research objective 02 (Compare differences in click though rates of informational, navigational, and transactional web queries when integrated sponsored and nonsponsored links are presented as a single listing on the SERP.) we implemented our characteristics in a program. We then executed the program on the Dogpile web search engine transaction log. Table 4 shows the results.

Table 4 Results from automatic classification of queries

Classification	*Occurrences*	*%*
Informational	3 502 013	83.4
Navigational	321 611	7.7
Transactional	377 447	9.0
	4 201 071	100.0

Table 4 shows that nearly 84% of web queries classified as *informational* in intent, with *transactional* queries representing about 9% of web queries, and *navigational* queries representing approximately 8% of the queries in the log. Our analysis found a surprising high percentage of *informational* queries. Naturally, these occurrences are statistically different (chi square (2) = 107.133, p = 0.01).

Results presented in prior work indicated that *navigational* intent was a significantly portion of web searching (Broder, 2002). As an example, Broder (2002) reported *navigational* queries of 24% based on approximately 3100 survey responses and 20%

based on an analysis of 400 web queries. The low occurrence of *transactional* queries in our results is also somewhat surprising. Broder (2002) reported *transactional* queries of 36% based on survey responses and 30% based on the analysis of web query.

The variation in reported percentage of *navigational* and *transactional* queries may be related to the size of the samples used in prior studies and the power log distribution of web queries (Anderson, 2006). Therefore, *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. Moreover, Broder (2002) used a random but small sample of queries. Rose and Levinson (2004) classified only the initial query in the user's session, which may have skewed results. Additionally, there may be ecommerce queries that are not transactional in nature, but may represent efforts to contain information about products. Finally, the prior studies (Broder, 2002; Rose and Levinson, 2004) were all executed on data from AltaVista, which may be non-representative of the overall web user population.

Table 5 shows that *navigational* queries had the highest occurrence of click through on sponsored links (more than 19%), and *transactional* queries had the lowest sponsored link click through (slightly more than 14%).

Table 5 Click through of informational, navigational, and transactional queries

User intent	*Count of sponsored*	*%*	*Count of organic*	*%*	*Total*
Informational	357 127	15.70	1 917 402	84.30	2 274 529
Navigational	37 333	19.31	156 030	80.69	193 363
Transactional	35 608	14.08	217 372	85.92	252 980
	430 068	15.81	2 290 804	84.19	2 720 872

It appears that sponsored links can provide relevant results for searchers to navigate to websites. In addition, the sponsored link click through rates for information queries was also quite high, indicating that these links can possibly provide relevant results to web searchers. The differences among the three types of user intent were not significant for either sponsored or nonsponsored links.

Table 6 shows the click through rates of each category by rank of the link as shown on the Dogpile SERP. The default listing on Dogpile is 20 results, so we can assume that most SERP contained this number of links. From Table 6, we see that nearly 30% of clicks for *navigational* queries were on the first result, which was statistically significant (chi square (2) = 14.941, p = 0.01). This noticeable different user behaviour, relative to *informational* and *transactional* queries, (where a substantial percentage of the clicks occurred on links 11 and higher) may assist in identifying user intent from implicit feedback measures. Web search engines are effective at ranking results for navigational queries but may be less effective at ranking for informational and transactional. This is understandable as the user intent behind these classes of queries is much more complex and dimensional.

Table 6 Click through of informational, navigational, and transactional queries by rank

Rank	*Informational*	*%*	*Navigational*	*%*	*Transactional*	*%*
1	381 286	10.9	95 318	**29.6**	38 436	10.2
2	264 836	7.6	26 124	8.1	27 597	7.3
3	191 379	5.5	16 148	5.0	21 044	5.6
4	159 283	4.5	10 893	3.4	16 708	4.4
5	126 553	3.6	7428	2.3	13 766	3.6
6	111 636	3.2	5832	1.8	12 054	3.2
7	88 974	2.5	4307	1.3	10 359	2.7
8	83 196	2.4	3658	1.1	9410	2.5
9	66 912	1.9	2951	0.9	7876	2.1
10	66 614	1.9	2485	0.8	7311	1.9
>10	733 860	**21.0**	18 219	5.7	88 419	**23.4**
No click	1 227 484	35.1	128 248	39.9	124 467	33.0
	3 502 013	100.0	321 611	100.0	377 447	100.0

Note: Highest percentage rank is bolded.

6 Discussion

Returning to our first research question, we investigated the effect of integrating sponsored and nonsponsored links within the same SERP listing. We analysed a transaction log from a web meta-search engine that displayed an integrated search listing. We examined overall user statistics from the transaction log, which were similar to user web searching statistics reported elsewhere (Park *et al*., 2005; Wang *et al*., 2003; Wolfram, 1999). Therefore, we believe that the findings from this research are generalisable to the user populations from other web search engines.

For research question two, we also employed a three-category classification (*i.e.*, *informational*, *navigational*, and *transactional*) of web searching that is useful in identifying the intent of the searcher. This classification is based on our own analysis and on previously published articles, most notably that of (Broder, 2002; Rose and Levinson, 2004). Additionally, our study uses a much larger data set of queries than prior work (Baeza-Yates *et al*., 2006; Kang and Kim, 2003; Lee *et al*., 2005), does not depend on external content and can be implemented in real time. This makes it a viable solution for web search engines to identify user intent of searchers and improve system performance.

We investigated the effect of integrating sponsored and nonsponsored links within the same SERP listing. We analysed a transaction log from a web meta-search engine that displayed integrated search listings of both sponsored and nonsponsored links. We noted that slightly less than 16% of clicks were on sponsored links. This is substantially lower than the sponsored link click through rate of approximately 30% reported in popular press comments. We cannot answer this discrepancy by transaction log analysis alone. This indicates the need for future research involving lab studies or panel data.

With the sponsored links not so prominently labelled on Dogpile, one would expect that integrated listings of sponsored and nonsponsored links would cause higher sponsored link click through rates. The fact that this did not occur may indicate that the popular press comments concerning click through rates of sponsored results might not be correct. However, assuming the published reports are correct or even in the ballpark, these research results certainly indicate that integrated listings do not cause a substantial increase of clicks on sponsored links.

There may be substantial benefits in an integrated SERP. First, given the reported negative bias users appear to have for sponsored links (Jansen and Resnick, 2006) and that research has shown sponsored links are relevant for user queries (Jansen, 2007), the prominent highlighting of presenting separate listings may be denying users relevant results. The second benefit concerns screen real estate (*i.e.*, the amount of pixel space for displaying information to the user) on the SERP, which is a limited resource. By presenting separate sponsored and nonsponsored listings in response to queries, there are many times duplicate links on the SERP (*i.e.*, links that appear both in the sponsored and nonsponsored listings). Valuable real estate taken up by duplicate links pushes other relevant links further down the results listing. This reduces the probability that users view these links. Integrating sponsored and nonsponsored listings addresses this issue and may be one method in helping overcoming the ranking bias of search engines (Introna and Nissenbaum, 2000) and improve the management of screen real estate.

In classification of web queries, a high proportion was *informational*, indicating that searchers continue to view web search engines primarily as traditional information retrieval systems. *Transactional* queries were approximately 9%, with *navigational* queries at slightly less than 8%. *Navigational* queries had the highest sponsored link click through rate of more than 19% (*informational* at nearly 16% and *transactional* at 14%). Although one might expect *transactional* queries to have the highest percentage of sponsored link clicks, major commercial entities buy sponsored links for branding purposes, which many times may be relevant results for *navigational* queries (*e.g.*, walmart, dell, ibm, target).

We examined click through rates at various ranks by query classifications. Based on this analysis, users submitting *navigational* queries clicked on the number one result nearly 30% of the time, compared to approximately 10% for *informational* and *transactional* queries. This points to the need for further analysis to identify other implicit feedback (Oard and Kim, 2001) metrics to assist in identifying user intent. *Informational* and *transactional* queries had higher click through rates at ranks greater than ten (21% and 23% respective) relative to *navigational* queries (approximately 6%).

In analysing our results, we are aware of certain limitations that may restrict the ability to generalise our conclusions. One issue is that the Dogpile user population may not be representative of web search engine users in general. However, Spink and Jansen (2004) report that characteristics among queries across search engines is fairly consistent. We also examined overall user statistics from the transaction log, which were similar to user search statistics reported elsewhere (Spink and Jansen, 2004). Therefore, we believe that the findings from this research are generalisable to the users of other search engines.

Nonetheless, we would certainly like to extend our research to other major search engines (*i.e.*, Ask, Google, MSN, and Yahoo!) if one could obtain this data. Another limitation is that we do not know the exact breakdown of sponsored and nonsponsored links actually displayed in respond to this set of queries at the time users submitted them.

Perhaps, the click through on sponsored links might just represent the actual percentage of displayed sponsored links. From a sampling of the SERP to a portion of these queries, the percentage of sponsored links appears much higher than 15%.

7 Conclusion

Sponsored web search is a critically important business model that finances the 'free' search that millions of users of web search engines have come to depend upon. Web search engine are continually exploring new and novel methods of both providing relevant results and increasing financial returns. In our study, we explored the effects of integrating sponsored and nonsponsored links into one listing on the SERP. Studies have shown that web searchers do not understand how web search engines rank results. Web search engines do not generally disclose how nonsponsored links are listed and ranked. Why do so with sponsored links? Certainly, exploring this line of research is an area worth pursuing. To improve, we search engines must leverage an increased knowledge of user behaviour, especially efforts for understanding the underlying intent of searchers and how this intent relates to sponsored links. 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 better help searchers achieve their goals.

Acknowledgement

We would like to thank especially Infospace for providing the data for this analysis, without which we could not have conducted this research. We encourage other web search engine companies to explore ways to collaborate with the academic community.

References

Anderson, C. (2006) *The Long Tail: Why the Future of Business Is Selling More of Less*, New York: Hyperion.

Baeza-Yates, R., Caldeŕon-Benavides, L. and Gonźalez, C. (2006) 'The intention behind web queries', *Paper Presented at the String Processing and Information Retrieval (SPIRE 2006)*, Glasgow, Scotland, 11–13 October.

Battelle, J. (2005) *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed our Culture*, New York: Penguin Group.

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual web search engine', *Computer Networks and ISDN Systems*, Vol. 30, No. 1, pp.107–117.

Broder, A. (2002) 'A taxonomy of web search', *SIGIR Forum*, Vol. 36, No. 2, pp.3–10.

Brooks, N. (2004a) 'The atlas rank report i: how search engine rank impacts traffic', July, http://www.atlasdmt.com/media/pdfs/insights/RankReport.pdf (retrieved 1 August 2004).

Brooks, N. (2004b) 'The atlas rank report ii: how search engine rank impacts conversions', October, http://www.atlasonepoint.com/pdf/AtlasRankReportPart2.pdf (retrieved 15 January 2005).

Dai, H.K., Nie, Z., Wang, L., Zhao, L., Wen, J-R. and Li, Y. (2006) 'Detecting Online Commercial Intention (OCI)', *Paper Presented at the World Wide Web Conference (WWW2006)*, Edinburgh, Scotland, 23–26 May.

Ding, W. and Marchionini, G. (1996) 'A comparative study of web search service performance', *Paper Presented at the 59th Annual Meeting of the American Society for Information Science*, Baltimore, Maryland.

Dobrow, L. (2004) 'Internet 2.0 the next level of the ultimate customer-based tool', *1to1 Magazine*, http://www.responsetek.com/resource/1to1Magazine.pdf (retrieved 31 March 2005).

Fain, D.C. and Pedersen, J.O. (2006) 'Sponsored search: a brief history', *Electronic Journal, Bulletin of the American Society for Information Science and Technology*, Vol. 32, No. 2, pp.12–13.

Fallows, D. (2005) 'Search engine users', 23 January, http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf (retrieved 28 March 2005).

Feng, J. (2002) 'Paid placement strategy for internet search engines', *Paper Presented at the 11th International World Wide Web Conference (WWW 02)*, Honolulu, Hawaii, 7–11 May.

Feng, J., Bhargava, H.K. and Pennock, D.M. (2007) 'Implementing sponsored search in web search engines: computational evaluation of alternative mechanisms', *Informs Journal on Computing*, Vol. 19, No. 1, pp.137–148.

Greenspan, R. (2004) 'Searching for balance', www.clickz.com/stats/big_picture/applications/article.php/3348071 (retrieved 4 May 2004).

Hansen, E. (2002) 'FTC wants paid search to shape up', http://news.com.com/2100-1023-940598.html (retrieved 14 January 2005).

Hotchkiss, G. (2004) 'Inside the mind of the searcher', http://www.enquiro.com/research.asp (retrieved 15 March 2005).

Hotchkiss, G. (2006) 'Enquiro eye tracking report ii: Google, MSN and Yahoo! Compared', *Journal*, November, http://www.enquiro.com/research.asp.

Hotchkiss, G., Garrison, M. and Jensen, S. (2004) 'Search engine usage in North America', http://www.enquiro.com/research.asp (retrieved 1 February 2005).

Introna, L. and Nissenbaum, H. (2000) 'Defining the web: the politics of search engines', *IEEE Computer*, Vol. 33, January, pp.54–62.

iProspect Inc. (2004) 'Search engine user attitudes', April–May, http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf (retrieved 28 March 2005).

Jansen, B.J. (2007) 'The comparative effectiveness of sponsored and non-sponsored results for web ecommerce queries', *ACM Transactions on the Web*, Vol. 1, No. 1, Article 3.

Jansen, B.J., Booth, D. and Spink, A. (2008) 'Determining the informational, navigational, and transactional intent of web queries', *Information Processing & Management*, Vol. 44, No. 3, pp.1251–1266.

Jansen, B.J. and Resnick, M. (2006) 'An examination of searchers' perceptions of non-sponsored and sponsored links during ecommerce web searching', *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 14, pp.1949–1961.

Jansen, B.J. and Spink, A. (2007) 'Sponsored search: is money a motivator for providing relevant results?', *IEEE Computer*, Vol. 40, No. 8, pp.50–55.

Jarboe, G. (2005) 'Online ads & search: looking back, looking forward', 17 January, http://searchenginewatch.com/searchday/article.php/3459961 (retrieved 18 January 2005).

Kang, I. and Kim, G. (2003) 'Query type classification for web document retrieval', *Paper Presented at the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 28 July–1 August.

Lee, K. and Miller, K.E. (2006) 'Internet users' attitude and behavioural intention on ebranding', *Int. J. Internet Marketing and Advertising*, Vol. 3, No. 4, pp.335–354.

Lee, U., Liu, Z. and Cho, J. (2005) 'Automatic identification of user goals in web search', *Paper Presented at the World Wide Web Conference*, Chiba, Japan, 10–14 May.

Liu, D. and Chen, J. (2006) 'Designing online auctions with past performance information', *Decision Support Systems*, Vol. 42, No. 3, pp.1307–1320.

Marable, L. (2003) 'False oracles: consumer reaction to learning the truth about how search engines work, results of an ethnographic study', Unpublished Research Report, Consumer WebWatch.

Montgomery, A. and Faloutsos, C. (2001) 'Identifying web browsing trends and patterns', *IEEE Computer*, Vol. 34, No. 7, pp.94–95.

Nettleton, D.F., Calderon, L. and Baeza-Yates, R. (2006) 'Analysis of web search engine query and click data from two perspectives: query session and document', *Paper Presented at the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, http://webmining.spd.louisville.edu/webkdd06/papers/paper-8-Analysis%20of%20Web%20Search%20Engine%20Query%20Sessions-WM_1058[1].pdf.

Oard, D. and Kim, J. (2001) 'Modelling information content using observable behaviour', *Paper Presented at the 64th Annual Meeting of the American Society for Information Science and Technology*, Washington, DC, USA, 31 October–4 November.

Pachauri, M. (2002) 'Researching online consumer behaviour: current positions and future perspectives', *Journal of Customer Behaviour*, Vol. 1, No. 2, pp.269–300.

Park, S., Bae, H. and Lee, J. (2005) 'End user searching: a web log analysis of NAVER, a Korean web search engine', *Library & Information Science Research*, Vol. 27, No. 2, pp.203–221.

Rose, D.E. and Levinson, D. (2004) 'Understanding user goals in web search', *Paper Presented at the World Wide Web Conference (WWW 2004)*, New York, NY, USA, 17–22 May.

Saaksjarvi, M. and Pol, G. (2007) 'Consumer mind-sets and attitudes towards online marketing messages', *Int. J. Internet Marketing and Advertising*, Vol. 4, No. 1, pp.26–36.

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999) 'Analysis of a very large web search engine query log', *SIGIR Forum*, Vol. 33, No. 1, pp.6–12.

Spink, A. and Jansen, B.J. (2004) *Web Search: Public Searching of the Web*, New York: Kluwer.

Wang, P., Berry, M. and Yang, Y. (2003) 'Mining longitudinal web queries: trends and patterns', *Journal of the American Society for Information Science and Technology*, Vol. 54, No. 8, pp.743–758.

Wolfram, D. (1999) 'Term co-occurrence in internet search engine queries: an analysis of the excite data set', *Canadian Journal of Information and Library Science*, Vol. 24, Nos. 2–3, pp.12–33.

Notes

1 http://www.dogpile.com/

2 http://www.clickz.com/showPage.html?page=3624821

3 http://www.dogpile.com/info.dogpl/search/help/paidlistings.htm

4 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. Other search engine logs are available at: http://ist.psu.edu/faculty_pages/jjansen/.

5 http://www.comscore.com/press/release.asp?press=325

6 http://www.internetnews.com/xSP/article.php/3502611