Analyzing URL queries

Wei Meng Lee, Mark Sanderson
Department of Information Studies
University of Sheffield, UK
[Wei_Meng_LEE@nlb.gov.sg, m.sanderson@shef.ac.uk]

Abstract

This study investigated a relatively unexamined query type, queries composed of URLs. The extent, variation and user click-through behavior was examined to determine the intent behind URL queries. The study made use of a search log from which URL queries were identified and selected for both qualitative and quantitative analyses. It was found that URL queries accounted for approximately 17% of the sample. There were statistically significant differences between URL queries and non-URL queries in the following attributes: mean query length; mean number of tokens per query; and mean number of clicks per query. Users issuing such queries clicked on fewer result list items higher up the ranking compared to non-URL queries. Classification indicated that nearly 86% of queries were navigational in intent with informational and transactional queries representing about 7% of URL queries each. This is in contrast to past research that suggested that URL queries were 100% navigational. The conclusions of this study are that URL queries are relatively common and that simply returning the page that matches a user's URL is not an optimal strategy.

1 Introduction

Search engines aim to effectively retrieve and order relevant documents with respect to a query supplied by a user. For Web searching, Broder (2002) proposed three types of queries based on user goal: *informational*, *navigational* and *transactional*. Informational queries were intended to find information about a certain topic. Navigational queries were used to discover a specific website that the user had in mind. Transactional queries were meant to reach a website where further interactions by users were expected. Rose and Levinson (2004) assigned queries to similar categories, though viewing transactional queries as resource queries, where the user sought to obtain things that were not information. Based on a large review of past work, Jansen et al. (2008) developed a three-level hierarchical classification with informational, navigational, and transactional goals at the top most level.

Numerous research studies related to the classification of user intent in Web search derived their findings from the results of search log analysis (Broder, 2002; Jansen et al., 2008; Rose and Levinson, 2004). By manually inspecting a random set of 400 queries from an AltaVista log, Broder (2002) found that 48% of the queries were informational, 30% transactional and 20% navigational. Rose and Levinson (2004)

randomly selected three sets of approximately 500 queries each from an AltaVista query log. Their results showed that there was a greater proportion of informational queries, and smaller proportions of navigational and transaction queries than Broder's study. Jansen et al. (2008) identified characteristics of informational, navigational, and transactional queries by qualitatively analyzing samples of seven logs taken from three Web search engines. Due to the size of the study, Jansen et al. (2008) argued that their findings generalized "*across multiple search engines and user demographic populations*". There were also examinations of specific query genres, such as geographic queries (Sanderson and Kohler, 2004), adult searching (Spink et al., 2006), and religious queries (Jansen et al., 2009).

Findings from an analysis of a search log for the MSN search engine showed that many of its popular queries were specifying an almost full URL (Zhang and Moffat, 2006). For example, the query 'yahoo.com' was the fourth most popular query. Their findings were supported by Jansen et al. (2008) who also found Web queries that contained partial or complete URLs. Jansen et al stated that some Web users submitted "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). In that study, URL queries were assumed to be navigational in nature, although that assumption did not appear to be tested. Similarly, Teevan et al (2007), while describing the existence of URL queries in their work on re-finding, assumed that such queries were navigational only.

Few studies appear to have sought to understand why users submit URL queries. None of the prior work focused on investigating the click through patterns of Web queries containing complete or partial URLs. These issues form the motivations for this research. The extent and variation of URL queries were investigated; click through behavior was also examined. A methodology to classify URL queries was developed. The methodology was then used to categorize the intent for Web searching when users submit URL queries to search engines. The study starts with a review of research related to URL queries.

2 Research studies relating to URL queries

Research studies related to URL queries can be categorized into three sub-areas: (1) URL type classification and tokenization; (2) relations between query terms and URLs; and (3) use of URLs for classifying user queries. Each area is now described.

2.1 URL type classification and tokenization

A Uniform Resource Locator (URL) refers to "*the syntax and semantics of formalized information for location and access of resources via the Internet*" (Berners-Lee et al., 1994:1). Using the depth of URL as the base metric, Westerveld et al. (2001) classified URLs into four different types: root, subroot, path and file, which were defined as follows

- "***root**: a domain name, optionally followed by 'index.html' (e.g. http://trec.nist.gov)*

- ***subroot****: a domain name, followed by a single directory, optionally followed by 'index.html' name (e.g. http://trec.nist.gov/pubs/)*
- ***path****: a domain name, followed by an arbitrarily deep path, but not ending in a file name other than 'index.html'(e.g. http://trec.nist.gov/pubs/trec9/papers/)*
- ***file****: anything ending in a filename other than 'index.html' (e.g. http://trec.nist.gov/pubs/trec9/t9-proceedings.html)*"

In order for URL queries to be examined further, they are often tokenized in two steps.

- The URLs can be split into a string of tokens bounded by the '/', '.', and ',' characters. This approach was used by Collins-Thompson et al. (2002), Ogilvie & Callan (2003); further work (Deepak and Khemani, 2006; Hinne et al., 2009; Kan, 2004; Kan and Thi, 2005) appeared to agree on this approach.
- Next, if any segment contains multiple concatenated words, further segmentation can take place. Kan, 2004 suggested transitions between digits, upper and lowercase as word boundaries. Hinne et al., 2009; Kan and Thi, 2005 took a similar approach.

2.2 Use of URLs for classifying user queries

According to Downey et al. (2008), the information goals of Web users can be inferred by examining the last URL visited in a session, as this is a reasonable proxy for the user's goal. Such a definition is suitable for large-scale analysis of search logs. Nevertheless, inferring the user's goal from the last URL in a session is a simplification because it assumes that URL satisfies the user's informational goal.

In their proposal of a query type classification method, Kang (2005) contended that navigational queries were associated to URLs with either 'root' or 'path' as the URL type. In a similar vein, Web documents containing URLs of the URL type 'root' were classified as navigational pages, destinations for navigational queries (He et al., 2007; Kang and Kim, 2003). He et al. (2007), however, chose to use URL type 'subroot' instead of URL type 'path' as another criteria to classify Web pages as navigational. When the URL only contained a domain name or a domain name followed by 'index.html', 'home.html' and 'index.jsp', the Web page was also considered as a navigational page (Zhu et al., 2007). For transactional queries used in Web searching, Kang (2005) proposed a formula that accounted for the existence of URLs that ended with music, picture, text, application or service file.

Based on the TREC 2003 data, Song et al. (2006) measured the distributions of URL depth types for homepage finding (navigational) queries and topic distillation (informational) queries. Most of the relevant documents for navigational queries have 'root' type URLs. For informational queries, 'file' type URLs accounted for more than half of the relevant documents. In a field study that aimed to characterize Web-based information-seeking behavior, Kellar et al. (2007) described certain URLs such as

'www.mail.yahoo.com/logout' as transactional URLs. Although Kellar et al. (2007) did not justify their definition the nature of transactional URLs was likely based on the nature of Web activity implied by the URL.

2.3 Synthesis of prior work

Two trends can be identified. First, there was little discussion of the user goals for URL queries. Despite the high popularity of these queries, few studies have sought to understand why users submit Web queries containing portions of URL or complete URLs. Second, none of the prior work has focused on investigating the click through patterns of URL queries. These issues form the motivations for this research. The results of this study will enhance the understanding of user goals when searching for URL queries.

3 Research objectives

The aim of this study is to understand the intent for Web searching when users submit URL queries to search engines. The research objectives are described below:

1. *Investigate the extent and variation of URL queries*
   For research objective one, URL queries were first identified and then selected for quantitative analysis.
2. *Examine the click through behavior for URL queries*
   For research objective two, the study analyzed the click through behavior for URL queries.
3. *Determine the informational, navigational, and transactional intent of URL queries*
   For research objective three, URL queries were qualitatively analyzed and manually classified in one of the three categories (informational, navigational, and transactional).

The study also analyzed the non-URL queries in the same sample. This helped to provide a research context for comparison of results obtained from the analyses of URL queries and non-URL queries.

4 Research design

The research adopted an inductive approach and was both quantitative and qualitative in nature. This study employed a search log analysis (SLA) methodology as outlined in (Jansen, 2008). There are three main stages of SLA: data collection; data preparation; and data analysis.

4.1 Data collection

For data collection, this study made use of a search log collected by the MSN search engine (http://search.msn.com). The search log (hereafter called query log) consisted of approximately fifteen million queries which were collected over a one-month period from 01 May 2006 to 31 May 2006. The log was released in the spring of 2006 as part of the "Microsoft Live Labs: Accelerating Search in Academic

Research" incentive. For each query the following details are available: a query identifier, the query itself, a user session identifier, a timestamp and the number of results returned. As part of efforts on the part of Microsoft to anonymize the log data, the session identifier only covered the searches conducted by a user within a few minutes of each other. Click-through data was provided in a separate file (hereafter called click log), with the two linked by the query identifier. Each click through record contained a query identifier, the query itself, a timestamp, the URL accessed, and information about the rank of that result in the results page.

4.2 Data preparation

The MSN search log consisted of 14,921,285 queries and 1,239,598 clicks (provided in a separate file). With such a large data set, two forms of data analysis were used. First an examination of the logs took place using searches of the entire data set. Then a random sample was obtained in order for a more detailed manual exploration to take place. To calculate the size of the sample, an assumption was made that standard sampling techniques could be used to estimate margins of error and confidence levels. These were chosen at 2% and 95% respectively, which required a sample of at least 2,401 queries (in fact 2,500 were chosen). The queries are randomly sampled from the log using a pseudo random function seeded with the time of day. The sampling used assumed that properties of the log were distributed normally, however, many log properties are not so distributed. Therefore, comparisons between the sample and searches on the full log set were made to determine the accuracy of the sample.

4.3 Data analysis

The terminology used in the study was derived from that used in earlier Web transaction log studies (Jansen and Pooch, 2001; Jansen, 2008; Jansen and Spink, 2006). Below are definitions for the common searching terminology used in the study:

- Term: a string of characters separated by white space
- Query: a sequence of terms submitted to a search engine
- Query length: the number of terms in the query
- Session: a sequence of queries from one user submitted within a pre-specified time period
- Session length: the number of queries submitted in a session
- Link: the URL of a document retrieved in response to a query
- Rank: the link's position in the result list

Other specific terminology that the study used is defined as follows.

- Token: a delimited segment of text (i.e., a series of alphanumeric characters) in the query, with every non-alphanumeric character treated as a delimiter
- Segment: a delimited segment of text in the URL, with forward slash or a punctuation mark treated as a delimiter

- URL query: a string of terms submitted by a searcher in a given instance, but contains a 'www' sub-domain prefix, a top-level domain or portions of URLs with at least 2 tokens (e.g. 'groups.yahoo', 'news.bbc.co', etc.)

The 2,500 queries in the sample were randomly selected from the query log. As it was possible that some queries in the log could be from automated processes (e.g. robots), the study used an interaction cut-off of 100 queries reported in previous Web searching studies (Jansen et al., 2007; Jansen and Spink, 2009) to identify such queries.

This study carried out basic statistical analysis of its quantitative data. Unless mentioned otherwise, the default values for the various statistical terms are as follows.

- Population size: 14,921,285
- Confidence level: 95%
- Margin of error: 2%

To compare whether a statistically significant difference exists between the means of two samples, a t-test was carried out for some of the analyses. This study makes several assumptions about the t-test, and these assumptions are as follows.

- independent two-sample, two-tailed t-test;
- the two sample sizes are unequal; and
- the two distributions have different variances.

The process of data analysis is described in details according to the order of research objectives.

4.3.1 Investigating the extent and variation of URL queries

For research objective one, URL queries first were identified. A small number of queries in the logs contained URLs as part of a command to the web search engine to either return links to a particular URL or domain or to restrict a search to the pages of a particular web site. These instructions were identified by the commands, "link:", "site:", and "linkdomain:". A small number of URLs were prefixed with the string "url:". It was not determined what this prefix was intended for, but as with the URLs used in the three instructions it was decided to identify but ignore these uses of URLs in a query.

A query was considered as a URL query if it began with 'www' as the sub-domain prefix, ended with a top-level domain (TLD) as specified by IANA (Undated), or contained partial URLs with at least 2 tokens. To determine the extent and variation of URL queries, four types of analysis were performed as follows: URL type distribution; top-level domain distribution; query length; and number of tokens. Analyses of query length and number of tokens were also performed on the non-URL queries in the same sample. This helped provide a context for comparison of results obtained from the analyses of URL queries and non-URL queries. Note in all searching of matching strings, the case of letters was normalized.

The classification of URL types from Westerveld et al. (2001), described above, was used in this study so as to determine the distribution of URL depth types. The study examined the TLD distribution amongst URL queries. Based on the list of TLDs from IANA (Undated), there are five types of TLDs: country-code, generic, generic-sponsored, infrastructure, and sponsored. The full list of TLDs can be obtained from IANA website.

4.3.2 Examining the click through behavior for URL queries

For research objective two, the study analyzed the complete click through behavior for sampled URL queries (identified in section 4.3.1). The study examined the click through rates of URL queries in the four forms of measurement as follows: number of clicks per URL query; rank of the link clicked by each URL query; and URL type of the link clicked by each URL query. For analyses of the rank and URL type of the link clicked by each URL query, only URL queries with at least one click were considered. All four forms of analysis were also performed on the non-URL queries in the same sample.

4.3.3 Determining the user intent of URL queries

For research objective three, URL queries (identified in section 4.3.1) were qualitatively analyzed and manually classified to be either informational, navigational, or transactional queries. Broder (2002) defined the intent of these queries as.

- **Navigational searching**: to locate a particular Website. The user may have the site in mind, or they may assume that such a Website exists.
- **Informational searching**: to find information on the web presented in a static form. No further interaction is predicted, except reading.
- **Transactional searching**: to reach a site where further interaction will happen. The main categories for such queries are purchasing, accessing an application, database or web-mediated service, or downloading multimedia. Often, username and password are associated with the transaction. Examples of such transactions include Web-based e-mail, banking, and posting to a message board.

Based on these definitions of intent, the following classification scheme was determined.

4.4 URL classification method

The query classification approach used by this study is based on the characteristics of URL query and URL click through data. The classification is based on Jansen et al's (2008) methodology, though in their work they did not have access to click data and so their method has been adjusted accordingly. To classifying the URL queries into the categories, the study defined the following characteristics.

4.4.1 Navigational searching

A URL query is considered as a navigational query if the following scenario occurs.

- Number of clicks per URL query is 1;
- URL type of the link is either root or subroot; and
- URL tokens in URL query must match those in link.

Based on the click through characteristics for navigational queries reported in various Web searching studies (Broder, 2002; Liu et al., 2007; Nettleton et al., 2006), the number of clicks per URL query was chosen to be 1. Findings from studies conducted by Lee et al. (2005) and Liu et al. (2006) demonstrated that a large number of navigational queries could be separated from other queries (informational and transactional) when the number of user clicks per query was set at less than 2. Lee et al. (2005) reported an accuracy of 80% in that same experiment, though the sample used by them was relatively small (~50).

A large number of URL queries did not have any associated clicks (43.1%), for the purposes of type classification, these were left unclassified.

For the choice of URL type 'root' and 'subroot' as a defining characteristic for navigational queries, several research studies (He et al., 2007; Kang, 2005; Kang and Kim, 2003) used the same approach. As the intent of a navigational query was to access a particular Website, the study made the assumption that the searcher may have the complete or partial form of the Website's URL in mind. Thus, the URL tokens in URL query must match those in the link.

4.4.2 Transactional searching

A URL query is considered as a transactional query if either of the following two scenarios occurs.

- Scenario 1
  - Number of clicks per URL query is 1;
  - URL type of the link is neither root nor subroot; and
  - Link leads to a Website where further interaction will happen.
- Scenario 2
  - Number of clicks per URL query is more than 1;
  - At least two of the clicked links need to be different from one another; and
  - Last link visited in the session leads to a Website where further interaction will happen.

Downey et al. (2008) asserted that the goals of human searchers can be inferred by examining the last link visited in the session. The intent of transactional searching is to reach a site for the purpose of further interaction. According to Broder (2002), the interaction represents the transaction defining transactional queries. Further interaction can be verified through either of the two methods as follows.

- **Method 1**: Term analysis of URL tokens within the clicked link
  - Link contains terms related to films, music, recipes, images, humor, and adult content; and
  - Link contains terms related to either 'obtaining', 'download', 'entertainment' or 'interact' such as lyrics, recipes, download, software, pictures, games, buy and chat.
- **Method 2**: Content analysis of the Website. The Website allows users to perform further interaction. Examples of such transactions include:
  - Purchasing;
  - Running some form of online application;
  - Finding various web-mediated services;
  - Downloading multimedia;
  - Accessing certain databases; and
  - Using username and password to log in to an account on the Website.

4.4.3 Informational searching

A URL query is considered as an informational query if any of the following three scenarios occurs:

- Scenario 1
  - Number of clicks per URL query is 1;
  - URL type of the link is neither root nor subroot;
  - Link leads to a page that does not appear to have been created in response to the fetching of the page; and
  - The only interaction with the page is expected to be reading.
- Scenario 2
  - Number of clicks per URL query is more than 1;
  - At least two clicked links need to be different from one another;
  - Last link visited in the session leads to a page that does not appear to have been created in response to the URL query; and
  - Last link visited in the session leads to a page where only reading of its content is reasonably expected.
- Scenario 3
  - Number of clicks per URL query is at least 1; and
  - The URL query is neither transactional nor navigational.

The level of interaction on a web page can be verified using the same approach for classifying URL query as a transactional query. The last scenario was a catchall.

4.5 Determining web query intent

To provide a research context for comparison of results obtained from the classification of URL queries, the study qualitatively analyzed the non-URL queries in the same sample. This facilitated the comparison of results obtained from the classification of URL queries and non-URL queries. The approach, however, was different. The complete approach used by this study to analyze non-URL queries was presented in (Jansen et al., 2008). We re-produce the bulleted version of their approach from Section 4.2 of their paper.

- "*Navigational web searching*
  - *queries containing company, business, organization or people names;*
  - *queries containing domains suffixes;*
  - *queries length (i.e., number of terms in query) less than three;*
  - *queries with 'web' as the source; and*
  - *searcher viewing the first search engine results page.*
- *Transactional web searching*
  - *queries containing terms related to movies, songs, lyrics, recipes, images, humor, and adult content;*
  - *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.); and*
  - *queries with movies, songs, lyrics, images, and multimedia or compression file extensions (jpeg, zip, etc.).*
- *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; and*
  - *uses question words (i.e., 'ways to', 'how to', 'what is', etc.)*"

Note, because URL queries are a novel form of query to be analyzed, the analysis was concentrated only on queries that had an associated click in the logs so as to ensure that query type was accurately determined. For URL queries without a click, it was decided to leave them unclassified. For non-URL queries, as the

means of determining type is more established, all non-URL queries in the sample were typed using the method presented in Jansen et al. (2008).

5 Results

A summary of the results is presented according to the three research objectives as follows. Full details of the results can be obtained from the dissertation work of one of the authors (Lee, 2009).

5.1 Research objective one: Determining the extent and variation of URL queries

Here both the large scale searching of the full logs was conducted as was the manual examination of the sample.

5.1.1 Searching the full log

In searching the ~15 million queries, first the URLs used in search engine commands were located and removed; Table 1 shows that this use of URLs was relatively rare (1.78%). A small number of queries were found where multiple URL commands were used

URL command type	Number	Percent
link:	210,887	1.41%
linkdomain:	33,212	0.22%
site:	49,172	0.33%
url:	3,660	0.02%
Union of types	265,946	1.78%
Total queries	**14,921,286**	**100.00%**

Table 1: Number of URL-based search commands used in query log expressed as a % of total number of queries

The queries with these commands were removed from the overall query set and further analyzed. A series of case insensitive search patterns were used to locate URLs in the logs. The strings 'http:' and 'https:' were searched for, as was 'www' and the character '.' followed by any one of the generic IANA Top Level Domains (TLD). In addition, the pattern of letter '.' letter was found to accurately locate URLs and it was also used. Finally all five search patterns were combined to collectively find the union of matching URLs. The figures in Table 2 show the frequency of occurrence of these patterns and their relative percentage compared to the total number of queries in the log. The total number of URLs found in the logs was 16.6%. As can be seen the letter '.' letter pattern found almost all URLs in the logs.

URL search string	Number	Percent
https:	6,232	0.04%
http:	33,652	0.23%
www	887,055	6.05%
dot TLD	2,248,533	15.34%

letter dot letter	2,374,187	16.20%
Union of searches	2,423,964	16.54%
Total queries	**14,655,340**	**100.00%**

Table 2: Number of URL queries found

As shown in Table 3, the TLDs used across the query log were counted; the commonest TLD by far was '.com', though matches on all twenty TLDs were found, though not shown in the table.

TLD	Number	Percent
.edu	27,257	1.2%
.gov	34,306	1.5%
.net	85,115	3.8%
.org	93,385	4.1%
.com	2,005,381	89.0%
Total TLD queries	**14,655,340**	**100.00%**

Table 3: Frequency of the five commonest TLDs found in URL queries

5.1.2 Examining the log sample

As the sampling made certain assumptions about the way that data was distributed in the logs, particular attention was paid to the degree of similarity between comparable results from the full log search and the examination of the sample. In a random sample of 2,500 Web queries, there were 432 URL queries which accounted for slightly more than 17% of the Web queries. From the full data set 16.6% of queries were found to be URLs. Within the 432, 37 queries were found to be search commands; 1.5% of the sample, similar to the 2% found in the full log search (see Table 1).

More than 92% of the URL queries from the sample were of URL type 'root'. Top-level domains were present in nearly 96% of the URL queries, with '.com', '.org' and '.net' as the top three percentages of the top-level domains (80.8%, 6.3% and 3.7% respectively). The same three TLDs were found to be the commonest in the same order in the full log search, though with slightly different proportions. The mean query length was 1.21 terms with the maximum query length of 7 terms. The number of one-term queries alone accounted for approximately 89% of the queries. Examples of URL queries that have more than 3 terms are "www.tabas, freedman at law in miami florida" (7 terms), "3 grade math and reading.com" (5 terms) and "www.frends of the gorge.oreg.com" (4 terms). In terms of the number of tokens per query, the mean was 3.1 tokens and the maximum number of tokens per URL query was 32. URL queries with two and three tokens represented the two highest percentages of the URL queries (48.4% and 32.4% respectively). The differences in the mean query length and mean number of tokens per query between URL queries and non-URL queries were statistically significant. Results from comparisons of mean query lengths and mean number of tokens in URL and non-URL queries are shown in Table 4 and Table 5 respectively.

Type of	Queries	Mean (query	Std.	Std.	95% Confidence interval for

query		length)	dev.	error	mean	
					Lower bound	Upper bound
URL	432	1.21	0.74	0.04	1.13	1.28
non-URL	2068	2.62	1.58	0.03	2.55	2.69

Table 4: Comparison of query lengths in URL and non-URL queries

Type of query	Queries	Mean (number of tokens)	Std. dev.	Std. error	95% Confidence interval for mean	
					Lower bound	Upper bound
URL	432	3.10	2.33	0.11	2.88	3.32
non-URL	2068	2.68	1.62	0.04	2.60	2.75

Table 5: Comparison of number of tokens in URL and non-URL queries

In both studies, URL queries were found to be relatively common. The properties of the sample of queries were found to be very similar to the properties found in the log as a whole.

5.2 Research objective two: Investigating the click through behavior for URL queries

As with research objective one, a combination of automated full log analysis and manual inspection of a sample was conducted.

5.2.1 Full log analysis

The 14,655,340 queries reported in Table 2 were split into the 2,423,964 URL queries; and the remaining 12,231,376 non-URL queries. Examining clicks on search results, it was found that at least one link in the results was clicked 59.5% times for URL queries and 60.4% for non-URL queries. The average rank of all result clicks for URL queries was 1.60 and 3.04 for non-URL queries. Using the short term session information in the log, for those queries for which there was a result click, the average number of results clicked per query was 1.14 and 1.44 respectively. Across all queries including those for which no result was clicked, the number of clicks per query was 0.68 and 0.87. A graph of the numbers of clicks per query was plotted in Figure 1. Of those URL queries for which there was a click on a result, 76.8% resulted in a single click; for URL queries this was 91.7%.

While the number of queries for which a result was clicked on was no different between URL and non-URL queries, the number of clicks per query and the rank of those clicks was different between the two types. The search engine appeared to be more successful in returning the results users wanted for URL queries than non-URL queries. However, there was still a notable number of URL queries (just over 8%) for which users clicked on more than one result. Queries of this sort were examined in more detail in Section 5.3.

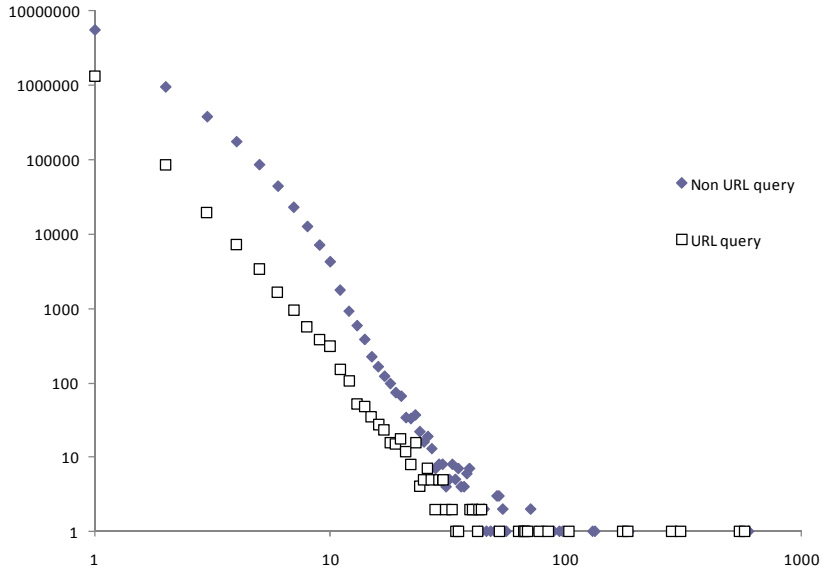

Figure 1: Log scale graph of the count (y-axis) of result clicks per query (x-axis).

5.2.2 Sample

From a total of 432 URL queries in the sample, 246 URL queries (56.9%) had one or more clicks. The mean number of clicks per query was 0.61 clicks with the maximum number of clicks of only 4 clicks. These numbers were similar to those gathered from the full log, providing more confidence in the construction of the sample. URL queries with only one click accounted for nearly 54% of the URL queries in the sample. For non-URL queries, the maximum number of clicks per query was 12. The percentage of URL queries with zero and one click per query (43.1% and 53.9%) were both slightly higher than the percentages of zero-click and one-click in non-URL queries (39.1% and 46.7%). The differences in the mean number of clicks per query in URL queries and non-URL queries were statistically significant. Table 6 shows the results from a comparison of clicks per query in URL and non-URL queries.

Type of query	Queries	Mean (clicks per query)	Std. dev.	Std. error	95% Confidence interval for mean	
					Lower bound	Upper bound
URL	432	0.61	0.58	0.03	0.55	0.67
non-URL	2068	0.89	1.09	0.02	0.84	0.93

Table 6: Comparison of number of clicks per query in URL and non-URL queries

In terms of the rank of the link clicked by URL queries, the mean rank was 1.52, and the highest rank number observed was 10. Nearly 80% of the clicks for URL queries were on rank 1. For non-URL queries, approximately 45% of the clicks for non-URL queries were on the first result. The first three links on the search engine results page together accounted for approximately 68% of the clicks for non-URL queries. URL queries, however, had nearly 94% of the clicks on the first three links on the results page. Non-URL queries had 31.4% of clicks on links 4 to 10; this is in contrast to 7% for URL queries. The highest rank

observed for non-URL queries in the sample was also much higher at 27. The differences in the rank of links clicked by URL queries and non-URL queries were statistically significant, Table 7 shows the comparison. Note, that as some queries had more than one click, the number of clicks, e.g. for URL queries (263), is larger than the number of queries (246).

Type of query	Clicks	Mean (rank)	Std. dev.	Std. error	95% Confidence interval for mean	
					Lower bound	Upper bound
URL	263	1.52	1.41	0.09	1.35	1.70
non-URL	1834	3.16	2.89	0.07	3.02	3.29

Table 7: Comparison of rank for click for URL and non-URL queries

Type	URL	Percent	non-URL	Percent
Root	212	80.6%	734	40.0%
Subroot	13	4.9%	110	6.0%
Path	2	0.8%	123	6.7%
File	36	13.7%	867	47.3%
Total	**263**	**100.0%**	**1834**	**100.0%**

Table 8: Distribution of URL type for links clicked by URL queries and non-URL queries

More than 80% of the result links clicked from URL queries were of 'root' URL type. The percentages of URL types 'subroot', 'path' and 'file' for the links from URL queries were lower at 4.9%, 0.8%, and 13.7% respectively. The percentage of clicked links with 'root' URL type for URL queries is two times the percentage of links with 'root' URL type for non-URL queries. URL queries, however, had a relatively low percentage (13.7%) of links with URL type 'file'. In the case of non-URL queries, links of URL type 'file' accounted for more than 47% of the total number of links clicked by non-URL queries. Table 8 shows the distribution of URL type for links clicked by URL queries and non-URL queries.

5.3 Research objective three: Determining the user intent of URL queries

Table 9 show the classification results for URL queries and non-URL queries respectively. Nearly 86% of URL queries were navigational in intent, with informational and transactional URL queries each representing about 7% of URL queries. Only 1 URL query was classified as an informational URL query under the catchall scenario (i.e. scenario 3) as stated in section 4.4.3. In the case of non-URL queries, informational queries accounted for the highest percentage of non-URL queries (58.4%). The percentages of navigational and transactional queries were 32.8% and 8.9% respectively.

Classification	URL	Percent	non-URL	Percent
Informational	18	7.3%	1207	58.4%
Navigational	228	85.8%	678	32.8%
Transactional	17	6.9%	183	8.9%
Total	**246**	**100.0%**	**2068**	**100.0%**

Table 9: Classification of URL queries with clicks (n=246) and all non-URL queries (n = 2,068).[1]

6 Discussion

The mean query length for URL queries was 1.21 terms with the maximum query length of seven. The number of one-term URL queries alone accounted for approximately 89% of the URL queries. Previous studies used other search engines run in different time periods and did not differentiate between the two types of queries in their transaction log analysis making direct comparison hard (Beitzel et al., 2004; Jansen et al., 2000; Jansen et al., 2007; Xue et al., 2004; Zhang and Moffat, 2006). The mean query length for these studies was in the range of 2.2 to 2.8 terms, about twice the mean query length for URL queries. Jansen et al. (2000) and Jansen et al. (2007) reported that percentages of one-term queries in Excite and Dogpile Web search engine logs were 31% and 18.5% respectively. The percentages of one-term queries in both their studies were less than one-third of the percentage of one-term URL queries. Hence, results from this study (perhaps unsurprisingly) differ from results reported in previous studies.

In terms of the number of tokens per query, the mean was 3.1 tokens and the maximum number of tokens per URL query was 32 tokens. The large number of tokens per URL query is due to the structure of URL query where multiple (partial) words within a URL are concatenated together (Chi et al., 1999). One reason for the presence of compound words in URLs is that most delimiters are not allowed in the URL address except for some pre-defined delimiters such as '.' and '/'.

It appears that URL queries are different from non-URL queries in terms of the characteristics of the query. URL queries tend to have a shorter query length but a higher number of tokens than non-URL queries. The differences in the mean query length and mean number of tokens per query between URL queries and non-URL queries were statistically significant.

Approximately 57% of the URL queries had at least one click per query. Non-URL queries with one or more clicks per query accounted for nearly 60% of the non-URL queries. These figures are both slightly lower than the percentage (64.8%) of Web queries with one or more clicks per query as reported by Jansen and Spink (2009). The mean number of clicks per URL query was 0.61 clicks which is lower than has been reported elsewhere (Zhang and Moffat, 2006). Zhang and Moffat (2006) analyzed 12,251,067 click throughs from a Microsoft search log and found the average number of clicks per query at 0.82. This is closer to mean number of clicks per query (0.89 clicks) for non-URL queries. The difference may be due to the fact that prior research studies (Jansen and Spink, 2009; Zhang and Moffat, 2006) had analyzed Web queries which contained both URL and non-URL queries.

[1] Results for non-URL queries do not total 100% due to rounding error.

The mean rank of the link clicked by each URL query was 1.52, and the highest rank observed was 10. Nearly 80% of clicks for URL queries were on rank 1; noticeably higher than the percentages of 49.6% and 18.9% reported respectively by Zhang and Moffat (2006) and Jansen and Spink (2009). Jansen and Spink (2009) found that approximately 31% of clicks were on rank 11 and higher. In this study, the maximum rank of the link clicked from URL queries was 10. The percentage of the clicks that occurred on links 2 and higher was only 20.2%.

It appears that URL queries are different from non-URL queries in terms of the click through behavior. URL queries tend to have fewer clicks but a higher ranking (i.e., higher percentage of clicks on the first few results of the search engine results page) than non-URL queries. The mean number of clicks and mean rank of links clicked by URL queries were 0.61 and 1.52 respectively. For non-URL queries, the mean number of clicks is higher at 0.89 clicks per query. The ranking of non-URL queries, however, is lower at 3.16. The differences in the mean number of clicks and mean rank of links clicked by URL queries and non-URL queries were statistically significant.

Nearly 86% of URL queries were navigational in intent, with informational and transactional URL queries each representing 7% of URL queries. Examples of informational and transactional URL queries include "myoce.oceusa.com", "4 KIDS PLAY .COM" and "www.ticketmaster.com". For instance, the URL query "myoce.oceusa.com" led to clicks on 3 links that end in different filenames

- http://myoce.oceusa.com/txmas/WF_txmas_index.htm
- http://myoce.oceusa.com/ resourcecenter/faq.asp
- http://myoce.oceusa.com/registration/register.asp?id=8

This finding of URL queries having all three possible user intents contradicts those of Jansen et al. (2008) and others who had classified URL queries as having only one user intent (i.e., navigational). Seeing more complex user search behavior implies that a retrieval strategy of returning the web page that exactly matches the user's query is a sub-optimal retrieval approach. Retrieving a richer set of possible matches is likely to improve search quality.

In the case of non-URL queries, informational queries accounted for the highest percentage of non-URL queries (58.4%). The percentages of navigational and transactional queries were about 32.8% and 8.9% respectively. For the entire sample of 2,500 queries (less the unclassified URL queries without a click), the percentages of informational, navigational and transactional queries were 52.9%, 38.4% and 8.6% respectively. The results are different from those reported in previous studies. Broder (2002) reported a proportion of 48% informational, 20% navigational and 30% transactional. Broder (2002) did not account for the remaining 2% of the Web queries. Other researchers (Jansen et al., 2008; Rose and Levinson, 2004), however, found a higher percentage of informational queries (approximately 64%) fewer navigational (14%) and more transactional queries (22%).

It appears that URL queries are different from non-URL queries in terms of the user intent. Nearly 86% of URL queries were navigational. On the other hand, more than half (approximately 58%) of the non-URL queries were informational.

Naturally, the study has limitations. Primarily, only a single search log was examined; other logs need to be studied to understand how much the results will generalize. However, the general Web searching statistics of the MSN search log as collected by Zhang and Moffat (2006) are consistent with those observed on other web search engines (Beitzel et al., 2004; Jansen et al., 2000; Jansen et al., 2007; Spink et al., 2000; Xue et al., 2004). Therefore, results obtained from this study are expected to generalize across multiple search engines.

Secondly the data was collected over a month and hence may not be able to fully represent the overall usage of the search engine users. Jansen and Spink (2006) found that Web query characteristics were relatively consistent across time except in the use of advanced search features and result pages viewed. Therefore, similar results are expected from the analyses of other query logs.

The results of this study are also limited by its methodology where each URL query is classified in one and only one category. Grimes et al. (2007) contend that users would often state a complex intent by using a general query. This argument was supported by Kang and Kim (2003) who highlighted that the same query could represent different information needs. However, manual classification of the URL queries and content analysis of the terms and destination gave us a high level of certainty that the web pages in this study were accurately assigned the correct user intent.

The strength of this study is the large dataset employed. Based on a low margin of error (2%) and a high confidence level of 95%, the study chose a sample size of 2,500 queries. The 2,500 queries were randomly selected from the MSN search log which consisted of 14,921,285 queries. From the results of the comparison between the sample and the full log, the random sample of Web queries is likely to be representative of the Web queries submitted by users of MSN search engine in general. Conversely, Broder (2002) and Rose and Levinson (2004) both used relatively smaller samples of 400 and 500 queries respectively.

7 Conclusion and recommendations

This study analyzed the 2006 MSN search engine query log so as to investigate the extent and variation of URL queries and to examine the click through behavior for URL queries. This study also attempted to determine the intent of URL queries.

It appears that URL queries are different from non-URL queries in the following areas: characteristics of the query; click through behavior; and user intent. URL queries tend to have a shorter query length but a higher number of tokens than non-URL queries. URL queries also seem to have fewer result list clicks but those clicks are at higher ranked results than non-URL queries. In terms of user intent, nearly 86% of URL queries were navigational, far more than non-URL queries, the majority of which (58%) were informational. Results showed that there were statistical differences between URL queries and non-URL queries in the following attributes: mean query length; mean number of tokens per query; mean number of clicks; and mean rank of links clicked.

This work makes clear that URL queries are common in web search and that although many such queries are served straightforwardly by treating them as navigational in intent, a notable number of users enter such queries with a different intent in mind. An implication for researchers of this result is that further study of how to best to serve the non-navigational URL queries will likely improve the quality of web search. The work presented here could also be potentially examined in the context of other query log studies. For example, the search and browse work of White and Drucker (2007), who examined query logs and subsequent browsing patterns, identified different types of user search behavior (i.e. navigators and explorers). An examination of URL queries could allow better classifications of such users. Also Teevan et al's (2007) who called for search systems to better support URL queries, could use the outputs of this work to understand how such queries can be managed well by search engines.

There are also several avenues for future research for the URL studies. First, Nettleton et al. (2006) used document viewing duration (and number of clicks) to classify Web queries into informational, navigational and transactional queries. Second, future research can extend the analysis of URL queries to include query reformulation and the use of various multimedia content collections on the Web search engine. Although the Web collection was the default, the rest of the content collections (images, audio, video and news) accounted for more than 28% of the queries submitted by users of Dogpile metasearch engine (Jansen et al., 2007). Finally, a laboratory study would complement the search log analysis in this study by recording the reasons for the search. This can help researchers to better understand the underlying intent of human users who submit URL queries to Web search engines.

8 Acknowledgements

We would like to thank the employees of Microsoft Research who made time to make their search log available to us, without which we could not have conducted this research.

9 References

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D. & Frieder, O. (2004). "Hourly analysis of a very large topically categorized web query log". In: *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 25-29 July 2004, Sheffield, United Kingdom*. pp. 321-328. New York: ACM.

Berners-Lee, T., Masinter, L. & McCahill, M. (1994). *Uniform Resource Locators (URL)* [Online]. California: Internet Engineering Task Force. http://www.ietf.org/rfc/rfc1738.txt [Accessed June 6, 2010]

Broder, A. (2002). "A taxonomy of web search". *SIGIR Forum*, **36** (2), 3-10.

Chi, C., Ding, C. & Lim, A. (1999). "Word segmentation and recognition for web document framework". In: *Proceedings of the 8th international Conference on Information and Knowledge Management, 02-06 November 1999, Kansas City, Missouri, United States*. pp. 458-465. New York: ACM.

Collins-Thompson, K., Ogilvie, P., Zhang, Y. & Callan, J. (2002). "Information filtering, novelty detection, and named-page finding". In: *Proceedings of the 11th Text REtrieval Conference (TREC 2001), 19–22 November 2002, Gaithersburg, Maryland, USA*. pp 338-349.

Deepak, P. & Khemani, D. (2006). "Unsupervised learning from URL corpora". In: *Proceedings of the 13th international Conference on Management of Data, 14-16 December 2006, Delhi, India*.

Downey, D., Dumais, S., Liebling, D. & Horvitz, E. (2008). "Understanding the relationship between searchers' queries and information goals". In: *Proceeding of the 17th ACM Conference on Information and Knowledge Management, 26-30 October 2008, Napa Valley, California, USA*. pp. 449-458. New York: ACM.

Grimes, C., Tang, D. & Russell, D.M. (2007). "Query logs alone are not enough". In: *Workshop on Query Log Analysis at WWW 2007: International World Wide Web Conference, 08-12 May 2007, Banff, Alberta, Canada*.

He, K., Chang, Y. & Lu, W. (2007). "Improving identification of latent user goals through search-result snippet classification". In: *Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence, 02-05 November 2007, Washington, DC*. pp. 683-686. Washington: IEEE Computer Society.

Hinne, M., Kraaij, W., Raaijmakers, S., Verberne, S., Weide, T. & Heijden, M. (2009). "Annotation of URLs: More than the sum of parts". In: *Proceedings of the 32nd Annual ACM SIGIR Conference, 19-23 July 2009, Boston, Massachusetts, USA*. New York: ACM.

IANA. (Undated). *Root Zone Database* [Online]. Washington DC: ICANN. http://www.iana.org/domains/root/db/ [Accessed June 6, 2010]

Jansen, B.J. (2008). "The methodology of search log analysis". In: Jansen, B. J., Spink, A. & Taksa, I. (eds.), *Handbook of Research on Web Log Analysis*, pp. 100-123. Hershey, PA: IGI.

Jansen, B.J. & Pooch, U. (2001). "A review of web searching studies and a framework for future research". *Journal of the American Society for Information Science and Technology*, **52** (3), 235-246.

Jansen, B.J. & Spink, A. (2006). "How are we searching the world wide web? A comparison of nine search engine transaction logs". *Information Processing and Management*, **42** (1), 248-263.

Jansen, B.J. & Spink, A. (2009). "Investigating customer click through behaviour with integrated sponsored and nonsponsored results". *International Journal of Internet Marketing and Advertising*, **5** (1/2), 74-94.

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

Jansen, B.J., Spink, A. & Koshman, S. (2007). "Web searcher interaction with the Dogpile.com metasearch engine". *Journal of the American Society for Information Science and Technology*, **58** (5), 744-755.

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 and Management*, **36** (2), 207-227.

Jansen, B.J. and Tapia, A. and Spink, A. (2009), "Searching for salvation: An analysis of US religious searching on the World Wide Web". Religion, **40**(1), 39-52

Kan, M.-Y. (2004). "Web page classification without the web page". In: *Proceedings of the 13th international World Wide Web Conference Alternate Track Papers & Posters, 17- 22 May 2004, New York City, NY, USA*. pp. 262-263. New York: ACM.

Kan, M.-Y. & Thi, H.O.N. (2005). "Fast webpage classification using URL features". In: *Proceedings of the 14th ACM international Conference on Information and Knowledge Management, 31 October - 05 November 2005, Bremen, Germany*. pp. 325-326. New York: ACM.

Kang, I.-H. (2005). "Transactional query identification in web search". In: *Proceedings Information Retrieval Technology, Second Asia Information Retrieval Symposium, 13-15 October 2005, Jeju Island, Korea*. pp. 221–232.

Kang, I. & Kim, G. (2003). "Query type classification for web document retrieval". In: *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 28 July – 01 August 2003, Toronto, Canada*. pp. 64-71. New York: ACM.

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.

Lee, U., Liu, Z. & Cho, J. (2005). "Automatic identification of user goals in Web search". In: *Proceedings of the 14th international Conference on World Wide Web, 10-14 May 2005, Chiba, Japan*. pp. 391-400. New York: ACM.

Lee, W. (2009). *Analyzing URL queries*. Masters Dissertation, University of Sheffield.

Liu, Y., Fu, Y., Zhang, M., Ma, S. & Ru, L. (2007). "Automatic search engine performance evaluation with click-through data analysis". In: *Proceedings of the 16th international Conference on World Wide Web, 08-12 May 2007, Banff, Alberta, Canada*. pp. 1133-1134. New York: ACM.

Liu, Y., Zhang, M., Ru, L. & Ma, S. (2006). "Automatic query type identification based on click through information". In: *Proceedings of the 3rd Asia Information Retrieval Symposium, 16-18 October 2006, Singapore*. pp. 593-600. Berlin Heidelberg: Springer-Verlag.

Nettleton, D., Calderon-Benavides, L. & Baeza-Yates, R. (2006). "Analysis of web search engine query sessions". In: *Proceedings of WebKDD 2006: KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD 2006), 20-23 August 2006, Philadelphia, PA*. New York: ACM.

Ogilvie, P. & Callan, J. (2003). "Combining structural information and the use of priors in mixed named-page and homepage finding". In: *Proceedings of the 12th Text Retrieval Conference (TREC 2003), 18-21 November 2003, Gaithersburg, Maryland, USA*. pp. 177-184.

Rose, D.E. & Levinson, D. (2004). "Understanding user goals in web search". In: *Proceedings of the 13th international Conference on World Wide Web, 17-22 May 2004, New York, NY, USA*. pp. 13-19. New York: ACM.

Sanderson, M. & Kohler, J. (2004). "Analyzing geographic queries". In: *Proceedings of Workshop on Geographic Information Retrieval SIGIR, Sheffield, UK*.

Song, R., Xin, G., Shi, S., Wen, J.-R. & Ma, W.-Y. (2006). "Exploring URL hit priors for web search". In: Lalmas, M., MacFarlane, A., Ruger, S.M., Tombros, A., Tsikrika, T. & Yavlinsky, A. (eds.), *ECIR, Vol. 3936 of Lecture Notes in Computer Science*, pp. 277–288. Berlin: Springer.

Spink, A., Jansen, B.J. & Ozmultu, C.H. (2000). "Use of query reformulation and relevance feedback by Excite users". *Internet Research: Electronic Networking Applications and Policy*, **10** (4), 317-328.

Spink, A. and Partridge, H. and Jansen, B.J. (2006) "Sexual and pornographic Web searching: Trends analysis". *First Monday*, **11**(4)

Teevan, J., Adar, E., Jones, R., & Potts, M. A. S. (2007). "Information re-retrieval: repeat queries in Yahoo's logs". In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* p.p. 151-158. ACM.

Westerveld, T., Kraaij, W. & Hiemstra, D. (2001). "Retrieving web pages using content, links, URLs and anchors". In: *Proceedings of the 10th Text REtrieval Conference (TREC 2001), 13-16 November 2001, Gaithersburg, Maryland, USA*. pp. 663-672.

White, R. W., & Drucker, S. M. (2007). Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web* (pp. 21-30). ACM.

Xue, G., Zeng, H., Chen, Z., Yu, Y., Ma, W., Xi, W. & Fan, W. (2004). "Optimizing web search using web click-through data". In: *Proceedings of the 13th ACM international Conference on Information and Knowledge Management, 08-13 November 2004, Washington, D.C., USA. pp. 118-126*. New York: ACM.

Zhang, Y. & Moffat, A. (2006). "Some observations on user search behavior". In: *Proceedings of the 11th Australasian Document Computing Symposium, 11 December 2006, Brisbane, Australia*. pp. 1-8.

Zhu, H., Raghavan, S., Vaithyanathan, S. & Löser, A. (2007). "Navigating the intranet with high precision". In: *Proceedings of the 16th international Conference on World Wide Web, 08-12 May 2007, Banff, Alberta, Canada*. pp. 491-500. New York: ACM.