# Some Observations on User Search Behavior

*Yuye Zhang*
NICTA Victoria Research Laboratory,
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
*zhangy@csse.unimelb.edu.au*

*Alistair Moffat*
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
*alistair@csse.unimelb.edu.au*

**Abstract** *We explore some issues that arise in the way that users interact with a web search engine, as evidenced by the records of their interaction provided by query and clickthrough log data. Our observations are derived from approximately fifteen million user queries recorded by the* `search.msn.com` *search service in May 2006.*

**Keywords** Log analysis, user behavior, search.

## 1 Introduction

Query logs from large scale search engines have always been a research commodity, providing crucial insights into the interaction between the users of the system and the system itself. The results of studies on search engine query logs can be applied to a range of fields in computing, including contributions to the fields of user interface design, to search result reranking, and to predictive caching and prefetching [Fagni et al., 2006].

Although there has been much done in analysis of logs containing queries submitted to a range of search engines [Silverstein et al., 1999, Spink et al., 2002, 2001, Lempel and Moran, 2003], research into the usefulness of clickthrough data as a model of user search behavior has received little attention due to a lack of public datasets. Because clickthroughs are indicative of a user's preference with respect to a particular query, they can be used as evidence to explore, for example, web personalization [Eirinaki and Vazirgiannis, 2003], and implicit relevance feedback [Joachims et al., 2005, Joachims, 2002, White et al., 2005].

In this paper we report findings from our analysis of a recently released log for the Microsoft MSN Search whole-of-web search engine (`http://search.msn.com`) containing approximately fifteen million queries and the corresponding clickthrough data, both of which are representative of a one month period in May 2006. We conduct our analysis by reporting key statistics of the dataset, and provide detailed insight into three major aspects of this query log: queries, sessions, and clickthroughs.

The dataset examined in our study is both large and recent, and is accompanied by clickthrough outcomes. The trends we have extracted from this dataset are thus both topical and timely, and provide evidence of a range of trends in user search behavior.

**Proceedings of the 11th Australasian Document Computing Symposium, Brisbane, Australia, December 11, 2006.**


## 2 Definitions

To ensure consistency of terminology, we make use of these straight-forward definitions:

**Query:** A string issued by an user to a search engine as a request for information.

**Term:** Individual words within a query, separated by whitespace. Terms may include alphanumeric characters, punctuation and other symbols. The number of terms in a query is the query *length*. Note that multi-word quoted phrases are considered multiple terms.

**Session:** A set of queries from a particular user, deemed (usually by a heuristic) to be part of a single interaction with a search engine. The session might include queries that relate to more than one information need, or topic. The *length* of a session is the number of queries contained in it.

**Results Page:** An ordered list of results presented to the user for a given query. The results page usually contains links to ten Results, plus a variable number of sponsored and other commercial links.

**Result:** An individual URL on the results page (plus a snippet of representative text extracted from that page), providing access to a document suggested by the search system as being an answer to the query.

**Clickthrough:** The action of the user in clicking on one of the Results listed in a Results Page, in order to access the page at the indicated URL.

Using these definitions we can see that a session contains one or more queries, each composed of one or

more terms. Each query execution generates a Results Page, and as a result of examining that Results Page, the user may generate zero, one, or multiple clickthroughs.

## 3 The MSN dataset

The MSN Search dataset was released as part of the "Microsoft Live Labs: Accelerating Search in Academic Research" incentive in 2006 (`http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx`). This dataset contains approximately fifteen million queries originating from users from the United States during May 2006, as recorded by the MSN search engine. All queries contained within the dataset are timestamped, sessionized and also anonymized to remove any personally identifiable information. Clickthrough data is provided in a separate file, with the two linked by a query identifier. Each clickthrough record contains that identifier, a timestamp, the URL accessed, and information about the rank of that result in the results page (being positions 1 to 10 on the first results page, 11 to 20 on the second, and so on). No information regarding unclicked results is retained.

In examining the logs, it quickly became apparent that not all of the queries in the query log were issued via the MSN Search frontend, and that the log included queries that originated from other external sources such as Web APIs, toolbars, third party programs, and so on. This presented a concern during preliminary analysis, as closer examination of the hundred largest sessions in the dataset revealed that around $90\%$ of them appeared to be machine-driven. In particular, the largest session in the dataset (containing over $30{,}000$ queries, issued at a consistent rate of around three queries per second) stepped in sequence through a sorted list of URLs to query their backlinks via the `linkdomain:` search option. Additionally, the second and third largest sessions (both consisting of $3{,}081$ queries) contained repeated requests for two different sets of exactly ten queries, again at a rate higher than would normally be associated with a real "user". We can only conclude that these two sessions were the result of someone exploring the search API, possible with a faulty program. Only two queries in these three sessions had any clickthroughs associated with them. Other common session variants of dubious usefulness included sorted alphabetical lists relating to some given niche such as real estate, as well as repeated non-sensical queries.

In one sense, the log thus depicts a "warts and all" approach to querying, in that it fairly reflects the workload that the system is asked to execute (whether erroneously, surreptitiously, or maliciously is unknown). On the other hand, our purpose in this investigation is to estimate the behavior of genuine users, and these sessions significantly distort the underlying trends – in reality it is extremely unlikely that a user would issue more than $100$ queries in a single session.

In order to report usage patterns of users rather than general web search engine traffic, we thus faced the issues of:

- Whenever possible, eliminating machine-generated sessions; and
- Ensuring each session represents one exchange between a single user and the search engine.

The latter is difficult to deal with, as it appears that the sessionization heuristic make use of the user's IP address, which can be identical for multiple users on a network behind, for example, a proxy server. Manual segmentation into individual sessions is not an option – the sheer size of the dataset, and the fact that a session can genuinely include multiple query threads (meaning queries can legitimately occur in very short intervals), makes this impossible.

Hence, we chose to filter the query log by removing all queries which did not have any corresponding clickthroughs. We believe that this strategy correctly removes all query requests originating from automated sources, as they are generally concerned with aggregating result page data and not exploring individual results. After applying the filtering process, out of the hundred longest sessions, only nine remained, matching our initial informal exploration.

Unfortunately, the filtering strategy has the disadvantage of also removing any queries originating from real users for which no results were clicked. In these cases non-clicking is valid information, since it suggests that the user was either not interested in any of the proposed results, or that their information need was satisfied by the snippets alone. Since these two cases are impossible to differentiate anyway, we felt that the removal of such sessions was an acceptable compromise in order to be sure that the machine generated traffic had been largely removed.

In the remainder of the paper, results are presented primarily for the filtered query log (note that the filtering was not applicable to the log of clickthrough data), but those results are contrasted with the unfiltered query log where it is appropriate to do so.

## 4 General statistics

Table 1 provides a range of statistics extracted from the query log (before and after filtering) and the clickthrough log. Over a third of the original queries were removed by the filtering process. Of those queries that had clickthroughs associated with them, around two thirds had a clickthrough to the first of the results, and the average first clickthrough position was $2.2$. In the unfiltered dataset, there was fewer than one clickthrough recorded per query, on average; removal of the queries that had no clickthroughs increased this ratio to around $1.4$.

Figure 1 shows the distribution of queries across the individual days in May 2006, beginning on Monday,

| Attribute | Original | Filtered |
|---|---|---|
| Number of queries | 14,923,285 | 8,831,275 |
| Number of unique queries | 7,095,622 | 3,875,436 |
| Number of terms | 35,824,851 | 20,641,810 |
| Number of unique terms | 2,605,699 | 1,151,998 |
| Number of sessions | 7,470,913 | 5,684,599 |
| Average query length (terms) | 2.401 | 2.337 |
| Median query length (terms) | 2 | 2 |
| Average session length (queries) | 1.997 | 1.554 |
| Median session length (queries) | 1 | 1 |
| Average time between queries in a session (mm:ss) | 4:21 | 7:28 |
| Median time between queries in a session (mm:ss) | 1:13 | 3:20 |
| Number of clickthroughs | 12,251,067 | |
| Number of clickthroughs at rank 1 | 6,074,872 | |
| Average clickthroughs per query | 0.821 | 1.387 |
| Median clickthroughs per query | 1 | 1 |
| Average rank of first clickthrough in a query | 2.161 | |
| Median rank of first clickthrough in a query | 1 | |
| Average time between clickthroughs in a query (mm:ss) | 2:03 | |
| Median time between clickthroughs in a query (mm:ss) | 0:47 | |

Table 1: Key statistics describing the query log and clickthrough log, before and after filtering to remove machine-generated queries.

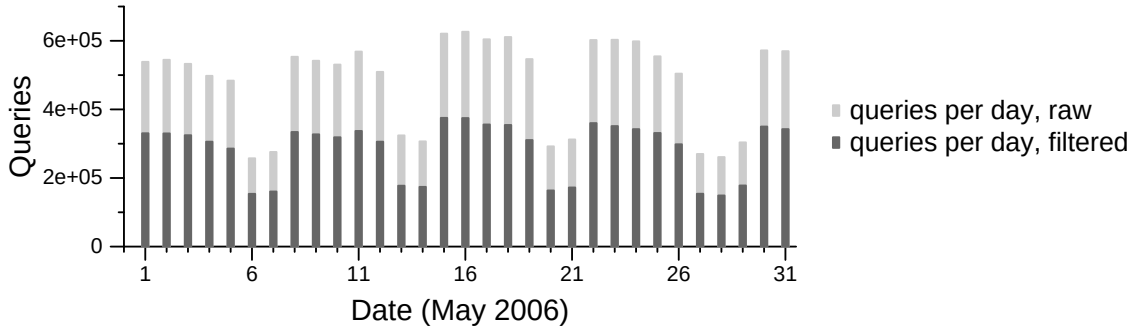


Figure 1: Daily query volumes for the collection period during May 2006, starting from Monday, May 1st. Query activity generally peaks on Monday, with a gradual drop off through until Friday, and a significant reduction through the weekend. Monday 29 May was a national holiday in the United States.

May 1st. It is clear that the volume of queries received by the search engines follows a general pattern of peaking early in the week and dropping steadily until Friday, with a sharp decrease over the weekend. This reflects quite accurately the weekly working cycle of the average white collar worker, and perhaps even captures the trend in which query volumes drops towards the end of the working week indicating a possible decrease in individual productivity. It also suggests that either search activity has become an important and integral part of the standard office routine, or that employees exploit company resources (time and connectivity) to undertake private searches.

Similarly, Figure 2 depicts an hourly breakdown for queries and clickthroughs received by the search engine, amalgamated across the entire month. Note that the logs supplied are a "representative sample" of US-originated queries during this period, but are not comprehensive. That is, the total hourly/daily/monthly search volume handled by the Microsoft engine is unknown, but the pattern of usage depicted in Figures 1 and 2 is accurate.

When broken down across the day, a pattern emerges where query volumes rise substantially from early morning (around 4am PST, at which time it is 7am in New York) peaking at around noon PST, when the whole country is at "work", and then decreasing steadily through until midnight PST. The ratio of clickthroughs to queries – both before filtering and after filtering – is relatively constant. This suggests that the machine-generated sessions that were removed by the filtering step are distributed through the day in the same pattern as are the user-generated query requests.

In both a daily sense and a hourly sense the filtering process does not appear to have affected the trends within the data.

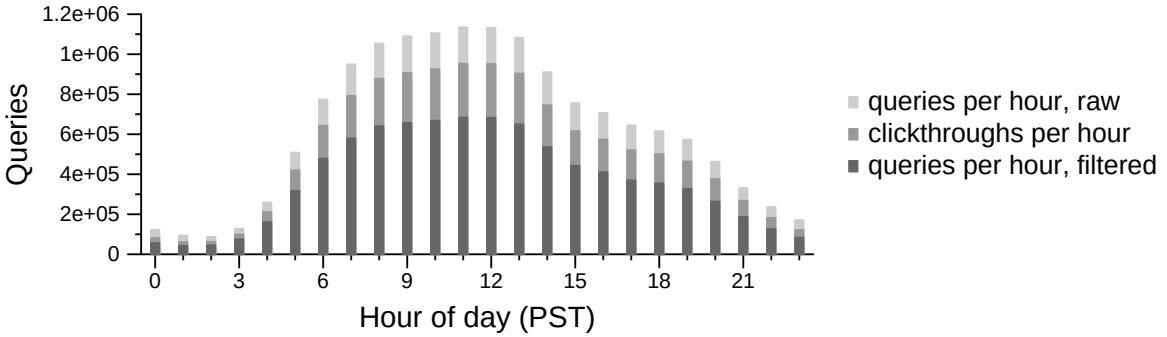


Figure 2: Hourly query and corresponding clickthrough volumes during the collection period. Search activity rises sharply during the early morning (when the US eastern states start work), peaking around noon and gradually drops off in the evening. Clickthrough volume stays consistent at around $0.8$ clickthroughs per query before filtering, and around $1.2$ clickthroughs per query after filtering.

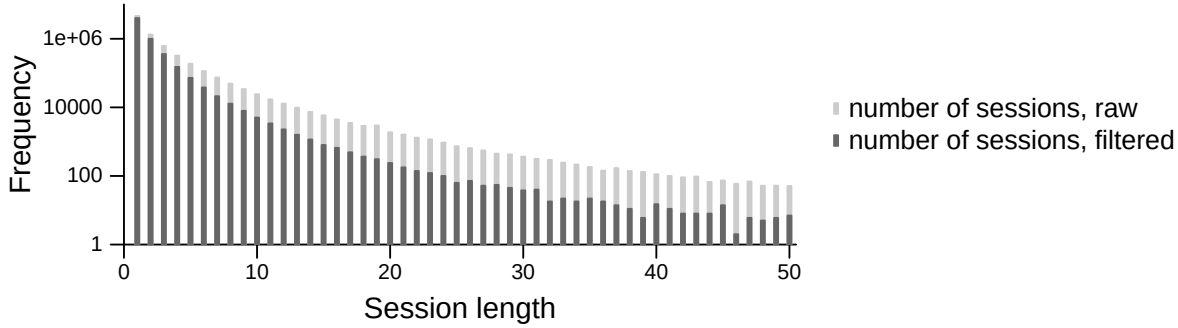


Figure 3: Frequency distribution for sessions of length $1$ to $50$, before and after the filtering process. Short sessions are more common than long sessions, a similar pattern witnessed in many other datasets. The filtering process removes queries, and thus tends to shorten sessions.

## 5 Sessions

Figure 3 portrays the distribution for sessions of different lengths, both before and after the filtering process. As is typical of other web search query sets, session are typically very short and contain just a few queries (less than two queries per session on average, for both raw and filtered query sets), and indicate relatively brief exchanges between users and the search engine. In the case of the filtered dataset, parts of the distribution gets shifted left as each session is trimmed of queries without any clickthroughs, resulting in much smaller numbers of longer sessions.

Figure 4 then shows the time difference in seconds between consecutive pairs of queries in multi-query sessions, using only the filtered query log. The majority of queries are issued within around one minute of each other, with very few queries issued at small time intervals (which, when it occurs, is a telltale sign of a session being machine driven). The largest interval recorded is $86{,}397$ seconds or just $3$ seconds under $24$ hours, which is perhaps indicative of the upper bound for the sessionization heuristic when the log data was prepared for distribution by Microsoft. Interactions over such a long period should probably not be considered as a single session.

The calculation used to determine resemblance is based on work by Broder [1997]. We define *resemblance* $R(A, B)$ between two queries $A$ and $B$ as:

$$R(A,B) = \frac{|S(A,n) \cap S(B,n)|}{|S(A,n) \cup S(B,n)|}$$

where $S(D, n)$ is the multiset of substrings of length $n$ in the string $D$, not permitting any whitespace characters. In our calculations, we used $n = 3$ to obtain character *trigrams*. For example,

$$S(\text{“}\texttt{eat at the theater}\text{”}) = \{\text{“}\texttt{eat}\text{”}, \text{“}\texttt{the}\text{”}, \text{“}\texttt{the}\text{”}, \text{“}\texttt{hea}\text{”}, \text{“}\texttt{eat}\text{”}, \text{“}\texttt{atr}\text{”}, \text{“}\texttt{ter}\text{”}\}.$$

Figure 5 shows the distribution of multiset resemblance scores between consecutive queries in multi-query sessions for the filtered query log. Most follow-on queries bear relatively little resemblance to their predecessor, except in the special case when resemblance is $1.0$. A resemblance of $1.0$ means that it is highly likely that an identical query was submitted consecutively; and this happens when the user requests the “next” results page.

## 6 Queries

One of the great fascinations with query logs is to see what it is that people are searching for. A startling dis-

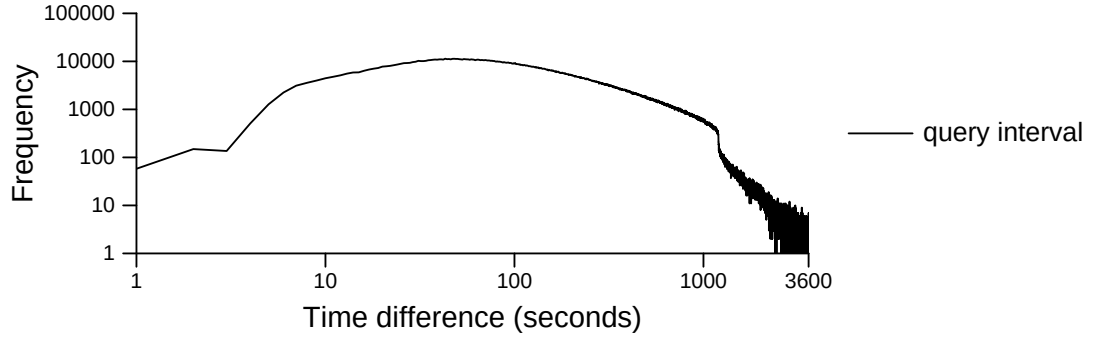


Figure 4: Interval in seconds between queries issued within a session, where sessions are as defined by Microsoft. The majority of intervals between same-session queries is less than a minute (60 seconds), although intervals of up to twenty minutes (1,200 seconds) are not uncommon. Only the filtered query log is shown in this graph. Note that the time intervals are quantized at one second values, but plotted as if they were continuous data.

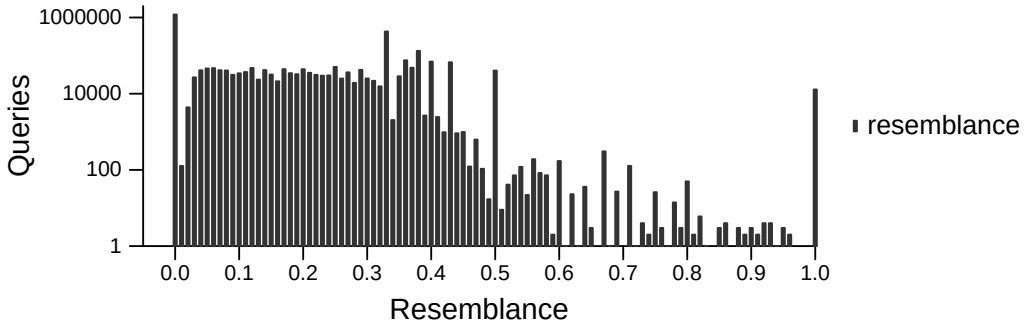


Figure 5: Trigram resemblance between consecutive queries issued within a session based on multiset overlap of trigrams of the query strings after filtering. A resemblance of 1.0 indicates that the pair of queries are identical; a resemblance of 0.0 occurs when the two queries have no trigrams in common.

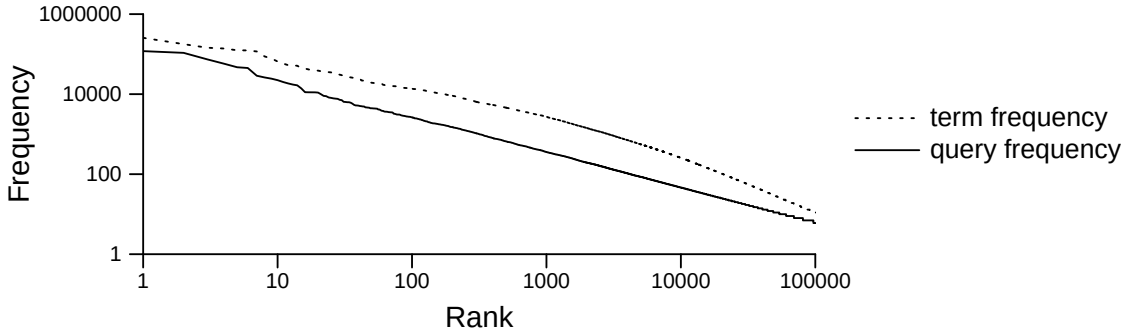


Figure 6: Rank-frequency distribution for queries and terms, in both cases after filtering. Both the query and term distributions follow the usual distribution. Note that neither queries nor terms were altered in any way when generating this figure, and we did not apply any stopping or case-folding techniques when creating the frequency distributions shown in the graph.

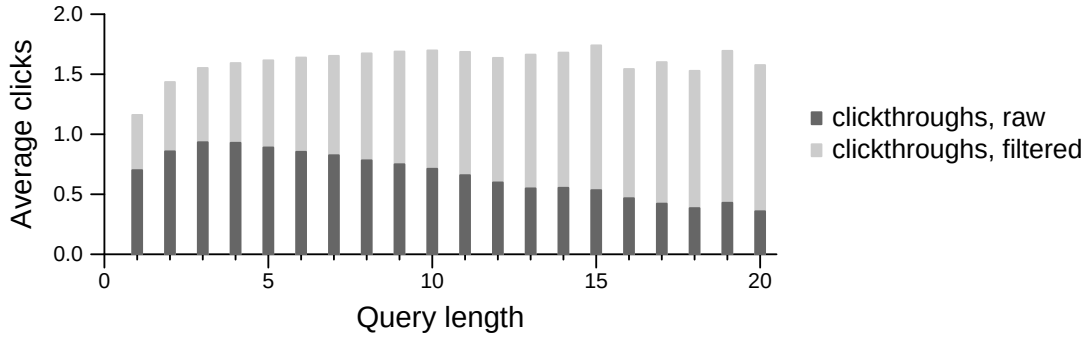


Figure 7: Query and clickthrough rates as a function of query length, measured as the average number of clickthroughs per query. Queries of more than ten terms have a reduced fraction of clickthroughs, indicating a possible lower availability of resultant data, or that (as is assumed in the filtering step) that these queries were automatically generated.

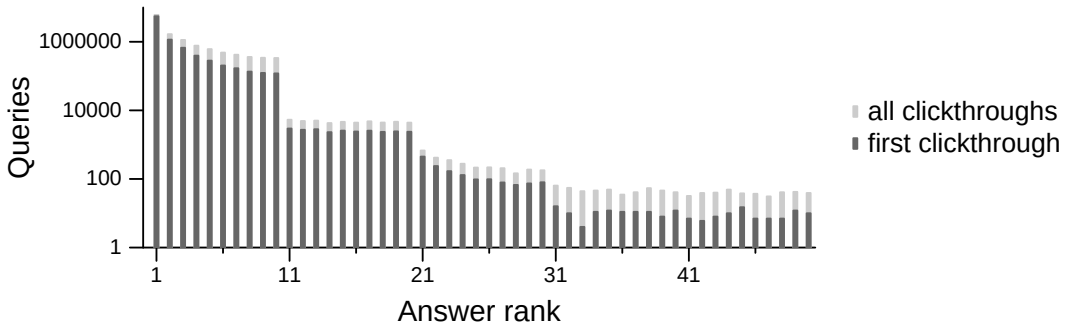


Figure 8: Positions in the results page at which clickthroughs occur, counting both the position of the first clickthrough per query, and the aggregate over all clickthroughs per query.

covery we made in this project is that when a query dialog box is available in a visible position (as it is, for example, in the MSN home page used by people to access hotmail accounts), its most common use is for navigational queries. The fifteen most popular queries in the query log supplied by Microsoft were all requests for other popular web services (including other search services), many specifying an almost full URL. For example, the query "`yahoo.com`" occurred more than $57{,}000$ times in the filtered query set, and was the 4th most common query. These top fifteen queries added up to $7.2\%$ of the filtered query log, and it would appear from this snapshot – rather dishearteningly for academic IR researchers – that canned answers are probably the best way to respond to these queries. Also somewhat disheartening is that the sixteenth most popular query, and the first non-web-service one, was "`american idol`". (Note that the Microsoft asset includes a separate log of "adult" queries, and that the log we have used in this paper is the sanitized one. Determining the extent to which the Microsoft cleaning process alters query and response characteristics is left to others.)

Figure 6 shows the distribution of both whole queries, and terms within queries, taking frequency as a function of rank in the usual manner. The most frequent individual query term was the word "`of`", with "`in`" the second most common term. In both the raw and filtered query logs the most frequent non-web-service term and non-contentless term was the word "`county`".

## 7 Clickthroughs

One of the reasons why the Microsoft data is of great interest is because of the supplied clickthrough logs. Figure 7 shows the average number of clickthroughs per query for the raw and the filtered data. In the raw data, a clickthrough after a long query is relatively unlikely, decreasing as the query gets longer. In the filtered query set, the clickthrough rate is relatively constant across the range of query lengths for queries longer than five terms. For short queries – which represent the majority – the clickthrough rate is lowest on queries of length one. Given the nature of many of the short queries, discussed in the previous section, this is plausible – the query "`mapquest`" is highly likely to generate exactly one clickthrough, for example.

Figure 8 shows the position at which clickthroughs occur. The answer in rank position one is the most likely to be clicked. There is then a gradual drop in likelihood of a clickthrough through the rest of that first page, followed by a marked drop in the probability of any document beyond rank $10$ being clicked. This pattern confirms that users are relatively unwilling to examine a second or subsequent results page via a "next"

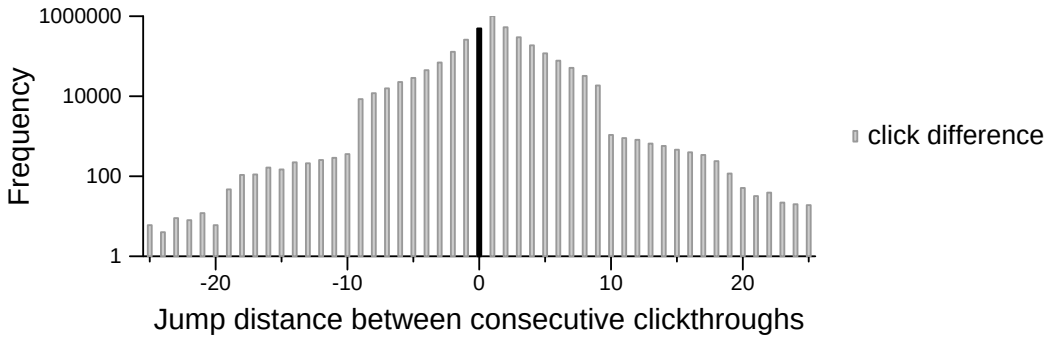


Figure 9: Jumps in clicked answer rank for queries that have two or more clickthroughs. The most common clickthrough jump is $+1$, to step from one proposed answer to the next.

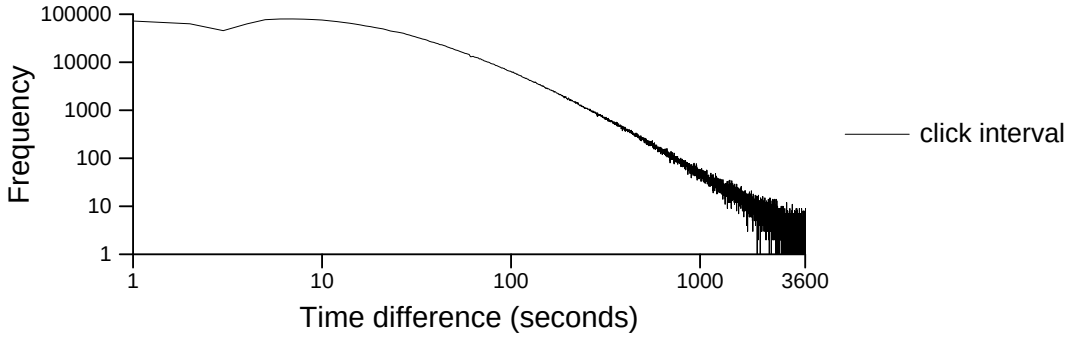


Figure 10: Interval in seconds between clickthroughs from a user for multi-clickthrough queries. Note that the time intervals are quantized at one second values, but plotted as if they were continuous data.

button, and that, within the first results page, preference is given to the answers that are presented near the top. Somewhat surprising in this graph is that there is a non-trivial minority of queries/users for which the first clickthrough does not take place until the third or even fourth results page for the query has been requested.

Figure 9 sheds further light on the manner in which users pursue paths through the presented data, by analyzing the sequence of clickthroughs on queries with more than one clickthrough. The most common clickthrough jump is $+1$, to step from one proposed answer to the next, as might be expected. But users are also nearly equally willing to backtrack through the results page, and click earlier answers, as they are to move forwards through the results pages. Users also (somewhat surprisingly) often click on the same answer document as consecutive actions, a jump of zero, perhaps caused by impatience, as they wait for a slow page to load. The sharp drops in frequency at jumps of $-10$ and $+10$ reinforces the fact that users are reluctant to examine subsequent result pages.

Our final graph, Figure 10, plots the time interval between consecutive clickthroughs for queries that generate multiple clickthroughs. Many user decisions to back out of one page, and clickthrough to another, are made within just a few seconds, and the decision time is typically less than a minute. This represents a quite different temporal distribution to the time intervals between queries in a session (Figure 4). Even unsophisticated users appear to have the ability to rapidly assess a page's relevance to them.

## 8 Related work

A recent review by Jansen and Spink [2006] provides a comprehensive overview into research activities in various fields of computer science utilizing different query logs, and compares key statistics over several significant studies prior to 2002. Several key outcomes regarding user browsing activity as well as syntax preferences were presented, although the authors noted the difficulty in drawing comparisons between studies involving different datasets.

One of the earliest log studies was conducted by Silverstein et al. [1999], who explored a query log containing approximately one billion queries from the AltaVista search engine and collected over a $43$ day period, which is generally regarded as the largest dataset of its kind to date. The authors reported key statistics regarding query and session distributions, as well as significant query-term correlations. Silverstein et al. also note that their dataset was not filtered to remove queries from automated sources.

In a similar experiment, Spink et al. [2001] examined a query log from Excite, comprised of over one million queries. The study found that typical queries are quite short and users generally only look at a few answer pages. Additionally, the authors provided a snapshot of query distribution in terms of topics, and discov-

ered that content within queries does not reflect the content available on the web. Lempel and Moran [2003] utilized another AltaVista query log containing around 7.7 million queries as part of their research into improving search engine throughput by caching popular query results. In this case, the statistics reported were focused towards patterns of page views.

There have been few large-scale studies into large volumes of clickthrough data, and it is in this respect that we feel our current work provides a contribution. Our results here can be seen as supporting recent work in connection with implicit relevance feedback, which contain some limited statistics regarding clickthrough outcomes [Agichtein et al., 2006, Joachims et al., 2005]. In this paper we have combined analysis of queries and clickthroughs in tandem, in order to draw out correlations between these two data streams.

## 9 Discussion and future directions

Much of what we have presented here simply confirms what has been found on other query streams – that queries are short; that a few queries (often completely inane) are very frequent in the query stream; and that there is a lot of mechanized access to search services. However, we can also draw a number of additional observations based on the clickthrough logs:

- Long queries have a smaller clickthrough rate;
- Users will sometimes take long jumps between consecutive clicks, and are also almost as likely to move backward as forward;
- Users dislike going beyond the first results page;
- Users are capable of making quick decisions about pages they have clicked on; and
- Users may click on the same answer page immediately after they have just viewed it.

One key issue that we may not yet have properly dealt with is that of spam removal within the queries, and possibly also within the clickthroughs (something which we have not considered). Similarly, the issue of session segmentation also needs to be addressed in order to create sessions of a finer granularity with more information value. In the absence of definitive information about the intentions of the user, such distinctions will remain elusive.

The natural application of our evaluations is to apply the understanding gleaned to try and improve search quality. Research by Joachims [2002] and Joachims et al. [2005] has shown that in a controlled setting, clickthrough data can be used to form pairwise relevance judgments, which in turn can be used to extract feature vectors for determining relevance of unseen documents. We will seek to explore these and related themes, possibly including a user study so that we have knowledge of user intention.

**Acknowledgments** Microsoft Research provided the logs described in this paper, and associated funding, via their "Accelerating Search" Project. Andrew Turpin (RMIT University) provided helpful input.

## References

E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR '06: Proceedings of the 29th ACM SIGIR Conference*, pages 3–10, New York, NY, USA, 2006. ACM Press.

A. Broder. On the resemblance and containment of documents. In *Sequences'97: Proceedings of the Symposium on Compression and Complexity of Sequences*, pages 21–29, Los Alamitos, CA, USA, 1997. IEEE Computer Society.

M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM Trans. Inter. Tech.*, 3(1):1–27, 2003.

T. Fagni, R. Perego, F. Silvestri, and S. Orlando. Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data. *ACM Trans. Inf. Syst.*, 24(1):51–78, 2006.

B. J. Jansen and A. Spink. How are we searching the world wide web?: A comparison of nine search engine transaction logs. *Inf. Process. Manage.*, 42(1):248–263, 2006.

T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the Eighth ACM SIGKDD Conference*, pages 133–142, New York, NY, USA, 2002. ACM Press.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th ACM SIGIR Conference*, pages 154–161, New York, NY, USA, 2005. ACM Press.

R. Lempel and S. Moran. Predictive caching and prefetching of query results in search engines. In *WWW '03: Proceedings of the 12th International Conference on the World Wide Web*, pages 19–28, New York, NY, USA, 2003. ACM Press.

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

A. Spink, S. Ozmutlu, H. C. Ozmutlu, and B. J. Jansen. U.S. versus European web searching trends. *SIGIR Forum*, 36 (2):32–38, 2002.

A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):226–234, 2001.

R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *SIGIR '05: Proceedings of the 28th ACM SIGIR Conference*, pages 35–42, New York, NY, USA, 2005. ACM Press.