An Investigation on a Community's Web Search Variability

Mingfang Wu **Andrew Turpin** **Justin Zobel**

School of Computer Science and Information Technology
RMIT University
Melbourne, Australia
Email: {mingfang.wu, andrew.turpin}@rmit.edu.au, jz@acm.org

Abstract

Users' past search behaviour provides a rich context that an information retrieval system can use to tailor its search results to suit an individual's or a community's information needs. In this paper, we present an investigation of the variability in search behaviours for the same queries in a close-knit community. By examining web proxy cache logs over a period of nine months, we extracted a set of 135 queries that had been issued by at least ten users. Our analysis indicates that, overall, users clicked on highly ranked and relevant pages, but they tend to click on different sets of pages. Examination of the query reformulation history revealed that users often have different search intents behind the same query. We identify three major causes for the community's interaction behaviour differences: the variance of task, the different intents expressed with the query, and the snippet and characteristics of retrieved documents. Based on our observations, we identify opportunities to improve the design of different search and delivery tools to better support community and individual search experience.

Keywords: Web Search, Search Context, Search Log Analysis, Community Search Behaviour.

1 Introduction

A major limitation of traditional information retrieval systems is that they focus on queries and documents, and neglect the users of the systems. This is primarily because the relationships between queries and documents, and the relationships among documents, are much easier to capture, model, and compute than relationships among queries, documents, and a user's search context. Consequently, documents are retrieved because of evidence such as that they contain the query words, and are frequently referred to by other authors in the web context, instead of matching users' search intentions. This often leads to unsatisfactory search experiences.

Recently, there is a trend to improve traditional information retrieval by leveraging users' actions. This includes explicitly asking users to give relevance scales to retrieved documents (White et al. 2001), and implicitly capturing users' interactions with retrieval systems, such as eye tracking (Granka et al. 2004, Joachims et al. 2007) and desktop application monitoring (Budzik & Hammond 2000)). Among these methods, one of the most popular is trying to predicate a user's search intention/interest through mining past *clickthrough data* from Web based search engines. Clickthrough data includes past queries issued by users, the set of retrieved pages for those queries, and the set of pages chosen (clicked) for viewing by users from the list of search results.

The premise behind exploiting clickthrough data is that past interaction history could reveal a user's current search intention, assuming that there was a good reason for the user to click on a link and visit a page. If, after reading the snippet of that page, the user clicks on a document because they find the page worth further investigation, then the "click" action could be interpreted as an implicit relevance judgment of the page. As such, this clickthrough data could help to reformulate the user's current query, or re-rank current search results. By incorporating this clickthrough data into retrieval and ranking criteria, the original query and document based ranking can be enhanced with users' contextual information. Of course, the click could be accidental, or the user was distracted from the original information need and clicked a page that was off topic, and so on, so clickthrough data must be employed with caution.

When we collect and use users' clickthrough data, we can treat each user as an individual, and personalise a user's current search result based on the user's own action; alternatively, we can aggregate the clickthrough data from the communities with which the user is assoicated, and use this information to tailor search results to meet the information needs unique to that user's community. Community clickthrough data has been used in collaborative filtering (Wang et al. 2006) and social recommendation (Smyth et al. 2004). The underlying assumption is that users from a community have similar backgrounds and like minds, thus their needs for information behind a query are likely to be similar, hence a document clicked by many users in the past should be useful to future searches of the same query by other users from that community.

However, this assumption has yet to be validated: would users from a community who submitted the same queries have the same information needs? If so, would they be interested in the same set of search results? In this paper, we present a study that examines this assumption.

Through studying a community's search history in its everyday search environment, we aim to investigate the variability of a community's search behaviour and identify key factors that could influence search performance within a community. We selected a well-defined community whose members are students (or staff) from the School of Computer Science of our University. The clickthrough data (including the queries and their associated clicks) were captured through a web proxy over a period of more than 9 months. The data log captured 540,424 queries (327,064 are unique) to Google search engine. To reflect the shared interests of the community, we only extracted a set of 135 queries where each query has been issued by at least ten users. We have also restricted the queries to those occurring in the computing domain; in our investigation, queries of a personal or non-computing nature were ignored. We believe that the findings from this real community with a large set of clickthrough data would help us to reveal and understand better the nature of community search, thus allowing better informed design of information search and delivery tools for community-based search.

We present a review of the background and related work in Section 2 followed with a description of the cache log used in this study. Section 4 explores a series of research questions; and Section 5 discusses our findings and their implications to the design of information search and delivery tools. Finally Section 6 concludes the paper.

2 Background and Related Work

Substantial research, in particular from the information science perspective, has investigated users' information needs, search behaviours and processes, and perceptions of relevance (Dervin & Nilan 1986, Ingwersen 1992, Saracevic 1997), with the aim of better understanding how humans process and retrieve information. The findings from these research areas played an important role in the design and development of traditional information retrieval systems and current web search engines. Of the most relevance to our work are studies on web search context: how users search the web, what they are searching for, what the characteristics of their search queries are (Jansen et al. 2000, Spink et al. 2001), how users with different search expertise, domain knowledge, and cognitive approach search the web (Hoelscher 1998, Kim & Allen 2002), and which features of web pages may influence users' search tasks (Tombros et al. 2005). Broder (2002) and Rose & Levinson (2004) analysed users' goals for searching the web and developed a web search taxonomy to classify such goals. These studies provide a broad understanding of how the general population use web search tools, and the requirements for a search engine to satisfy this web population.

At a lower level of investigation, clickthrough data is used to implicitly capture an individual's search context, with a view to using this information to personalize search results. This method assumes that a user's past queries and their associated clicks would reveal the user's interests, and thus it could be used to predict the user's future preference. This clickthrough data can be used to model a user's immediate information need or long term preferences, depending on the period of time over which the clickthrough data was captured. For example, Shen et al. (2005) infer a user's immediate information need by her recent queries and the snippets of clicked search results. When this model of a user's short term interests is updated by a new query or click on a new page, the user's longer-term interests could be inferred (Sugiyama et al. 2004).

When the clickthrough data from users with similar information needs is aggregated, it could plausibly be used to tailor search results for the members of that community. By doing so, the privacy of individual users is also protected. Here the community refers to groups who share similar interests or information needs. This community may be predefined; for example, because the members of the community have the same or similar social background, such as a same job role in a working environment; or could be interest-based, for example dynamically inferred through users' search history (Almeida & Almeida 2004).

The usual way to use a community's clickthrough data is to treat a click associated with a query as a vote for the page's relevance. For example, Smyth et al. (2004) used a hit matrix that records the relative click frequency of retrieved pages per query, and this information is used to re-rank future search results for the same query or similar queries. They found that the current users using lists reordered with their approach could answer more fact finding questions in a given time limit, and that more questions are answered correctly.

In some work the boundary between the personal search history and community search history is blurred. In the methods of Agichtein et al. (2006) and Joachims (2002), ranking algorithms are trained based on aggregated search history obtained over a large number of users. The search history includes not only the usual clickthrough data, but also fine-grained features such as query-text features, and browsing features such as page dwell time.

Almost all of these studies (and many others that can not be mentioned here due to space considerations) report positive results, and the use of clickthrough data is a key component of all of these methods. It is natural to ask whether clickthrough data is sufficient reliable as an indicator of user preference. Would a user's own past search history or a community's search history predict the user's current interests? Teevan et al. (2005) show that a group of people from the same company and with similar IT background had different intents even when they issued the same query to a search engine, and thus they rated the retrieved pages differently. Joachims et al. (2007) conducted a controlled experiment to study the reliability of clickthrough data through manipulating the relevance ordering of search results and comparing explicit feedback against manual relevance judgment. They concluded that clicks are informative, but biased.

However, there are few studies on the characteristics of a community's search behaviour. In this study, we examine such search behaviour by analysing click-through data from a well-defined community. We focus on users' search variability behind the same queries, and the factors that may cause search variations.

3 A Community Web Cache

The data set used for our study is originally from the cache logs from our school's web proxy server. This log recorded all web activities of those students and staff for the period from 1st January 2006 through to 6th October 2006.

User Identification One of the difficulties in search log analysis is that de-identification of data to protect privacy can remove information from the log. Most studies (Spink et al. 2001) use IP addresses as an identity marker, but, in a shared computing area, a computer can be used by many different users, and a user can have access to many computers. In our case, according to the school's policy, users need to use their personal identifier to access the web. This enables us to assign each activity clearly to a distinguishable individual, and thus we can trace an individual's search history. To preserve privacy, each user's account information was replaced with an anonymous ID prior to us receiving the data.

Search Session Identification We divided the data set into search sessions. In principle, a search session should start when a searcher submits a query and end when the searcher gets information to satisfy her need or otherwise gives up the search. However, it can be difficult to rigorously detect such search session boundaries automatically from query logs. Previous studies have used various timeout periods for session segmentation, ranging from 15 minutes (He et al. 2002) or 30 minutes (Mat-Hassan & Levene 2005) to a whole day (Spink et al. 2001) according to different research goals. As our purpose of using a session is to identify those search activities of a query, we examined our data and found 15 minutes to be a reasonable boundary — our users usually shifted their search topics within 15 minutes. Later, for our targeted queries, we combined neighbouring 15 minute sessions manually where we believed a search session may have been split by the 15 minute cut-offs.

Query Statistics The Google search engine is the most frequently used search engine in our proxy logs, hence we focus our attention to those queries sent to Google. There are 540,424 Google queries in the collection (after removing empty queries and those queries from subsequent result page requests). These queries were submitted by 3,574 unique users. On average, each user submitted 151 queries over the 9 month period. The average query length is 2.64 words. This is very close to that of the general web user population (Spink et al. 2001).

Among the 540,424 queries, 260,786 (48.3%) were issued only once, with the remainder (51.7%) re-occurring at least once. Overall, there are 66,279 distinct re-occurring queries, so on average each of these was submitted 4.2 times. Table 1 shows the number of users that submitted each re-occurring query. We can see that 70% of queries that occurred more than once in the log were always submitted by the same person (though a different person for each recurring query; for example, one user issued the query "bbc news" once or twice every day), while the other 30%, which occurred multiple times in the log, were issued by more than one user.

Data Set The users recorded in our cache log searched a wide range of topics from sports, music, news, computers, science and so on. Although the majority of them have the same study major and in a similar age group and economic status, their interests as expressed in searched topics were diverse. However, they are expected to form a close-knit community when they search on the topics related to their studied major, that is, computer science and information technology. For example, when searching a particular topic, it is likely that they were taking the same lecture or doing the same assignments.

We selected queries that can meet the following criteria: 1) the query is in the computer science and information technology domain and was submitted to the Google search engine; 2) the query has been sent by at least ten users; and 3) the search sessions can be reconstructed.[1] In the end, we collected 135 such queries.

To identify the clicks associated with this set of queries, we first located all search sessions that had any of these queries, then cleaned up the following clicks by filtering out those pages that either resulted from browsing actions within a site, or were not associated with the selected query. After the clean-up, we found that these 135 queries were searched 3,480 times by 1,115 users, each query re-occurring 25.8 times on average. There were 4697 clicks in total — 1.3 clicks per search on average. There were 14.9%, 56.5%, and 28.6% searches that had zero clicks, one click, and more than one click respectively. The reasons behind the queries with zero clicks are unknown; the searchers could be satisfied (or unsatisfied) by looking at snippets only, or users were handling multiple tasks (Spink et al. 2006) at the same time.

4 Community Search Analysis

4.1 Overall Click Behaviour

It is often assumed that a user's action of clicking on a URL indicates the page's relevance to the submitted query. However, studies of user clicking behaviour show that, while a user's decision to click is mainly influenced by the relevance of the snippets associated with the pages, it may also be biased due by the order

[1] As the retrieved document set was not recorded at the time when a user issued the query, we reconstructed this set by sending the same query to the Google search engine at a later date (from 2nd December 2006 to 7th December 2006), and recorded the URLs of the top ten retrieved pages. In order to minimise the differences between the original and reconstructed retrieved sets, we don't include queries where most of their clicked pages are not in the reconstructed top ten list. Here we take ten as a threshold because previous studies showed that most users do not access search results past the first page (Jansen et al. 2000).

No. of users	1	2	3	4	5	6	7	8	9	10+
Queries (%)	70.2	19.9	4.3	1.8	1.0	0.6	0.4	0.3	0.2	1.1

Table 1: The proportion (%) of users responsible for re-occurring queries.

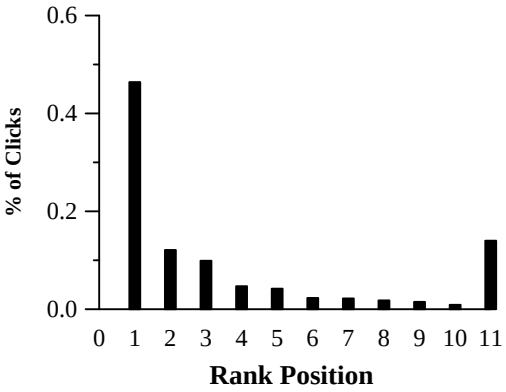

Figure 1: Percentage of clicks at each rank position.

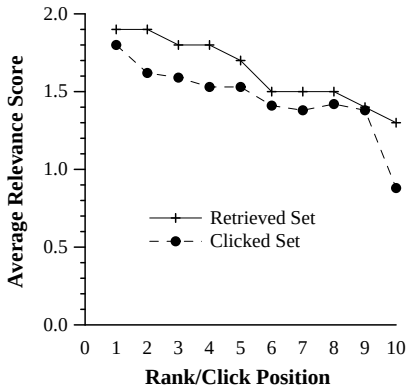

Figure 2: Average relevance score at each position in two ranked lists.

of a page in a ranked list of search results (Joachims et al. 2007). In this section, we presented our investigation if the above claims still hold true for our selected close-kit community. In particular, we set out to investigate the following three research questions.

1. Is it true that a page with higher rank would be clicked by users more frequently than a page with lower rank?
2. Does a ranking of documents based on *click frequency* correlate with a *relevance*-based ranking?
3. Are clicked pages relevant?

Q1. Is it true that a higher ranked page would be clicked by users more frequently?

For each page in the search result set (10 per query for 1350 in total), Overall, there are 762 retrieved pages that have zero clicks, and their average rank is 6.5, while the average rank of the other 588 pages with at least one click is 4.1. The ranks of clicked pages is significantly higher than that of un-clicked pages (un-paired t-test, $p < 0.01$). This shows that those clicked pages have higher rank on average.

Figure 1 shows the distribution of clicks for each ranked position. Clicks that do not select a page in the top ten answers for a query are assigned rank eleven. Nearly half of the clicks (46.7%) are on the top-ranked URL, 12.3% of clicks are on the second-ranked URL, and 14% of clicks are at 11th position. These figures indicate that higher-ranked pages are selected more frequently than lower-ranked pages, confirming that our query log shares characteristics of that used by Joachims et al. (2007).

Q2. Does ranking based on click frequency correlate with relevance-based ranking?

We re-ranked each of the top ten search results in the order of their click frequency from high to low, and use Kendall's-τ rank correlation coefficient to measure the degree of correspondence between two ordered lists. The Kendall's-τ coefficient ranges from -1 (perfect disagreement — one list is the reverse of the other), through 0 (the ranks of two lists are independent), to 1 (perfect agreement — the ranks of the two lists are exactly the same). As there are ties in the re-ranked list, the Kendall τ_b is used (Fagin et al. 2003).

Among the 135 queries, there are 86 queries whose paired lists have strong tendency of agreement ($\tau_b >= 0.4$, of which 40 are significant at $p < 0.05$ level, by the two-tailed Z-test), and for the other 49 queries the paired lists are independent ($|\tau_b| < 0.4$). There was no perfect disagreement found for any query.

Q3. Are the clicked pages relevant?

The assumption behind utilizing clickthrough data for ranking is that a click is a form of relevance judgment — the clicked pages are more relevant than those not clicked. Did our users click on a page selectively or just click on the top-ranked pages? To answer this question, we examined the relevance of the top ten pages from each search result.

Each page was judged for relevance (by either one of authors or a postgraduate student) on a three-point scale: highly relevant, relevant and irrelevant. These corresponded to scores of 2, 1, and 0 respectively. Figure 2 shows the mean relevance scores over all search sessions for the set ranked using click frequency, and the set ranked as per the Google result. We can see that the clicked data are of high relevance to the queries. Over 49,240 clicks, there are only 591 (12%) that were judged to be on an irrelevant page.

We have seen that our users tend to click the top-ranked search pages, and Figure 2 shows that the top ranked pages are also of high quality. A remaining question is whether our users clicked the top-ranked search results blindly or clicked relevant URLs that happened to be highly ranked.

We have only three queries in our data set whose first ranked page was judged irrelevant. We found that, in these cases, the majority of our users clicked lower ranked, yet relevant, pages. For example, for the query "ssi", the relevant pages appear at positions 4, 5, and 7 (the pages at position 4 and 5 are from the same site). Of the 25 users that issued this query:

- 11 users did not click any page;
- 6 users clicked on the fourth or seventh ranked page;
- 2 users first clicked on the first ranked page and then clicked on the fourth search page; and
- For the remaining 6 users, their first click is not on the top ten list but 5 of them are judged relevant or highly relevant.

For the 11 sessions without any click, there are query modifications in 10 sessions to either expand the query to "server side includes" or include more contextual words such as "ssi in html". In these cases, we assume that the users make relevance judgments by reading the snippets only. From these observations and the finding from a systematic evaluation (Thomas & Hawking 2006) that users were able to reliably distinguish between high- and low-quality result sets, we can be confident that our users did not click a relevant page just by chance.

4.2 User Click Variability

From the above discussion, we see that users demonstrated a tendency to click on highly ranked documents. However, we observed a difference in click patterns among users even for the same query. Here we measure this difference by using inter-rater agreement[2]. We calculate it in two ways. First, we treat the clicks from a query as a simple *click-list* ignoring the position of the click in the ranked list. The average inter-rater agreements over all queries is 0.36. That is, only about a third of all possible pairs of users over all queries clicked the same set of pages.

Second, we calculate the inter-rater agreement amongst the first click made by all users, then the second click, and so on (we refer to this as the click position of the click, as opposed to the click rank). As shown in Figure 3, the inter-rater agreement for click position 1 is 0.52. It then dropped dramatically to 0.25 for click position 2. The decreasing inter-rater agreement as click position increases indicates that users have a tendency to click the top-ranked page (hence the high agreement for click position one) and then accessed the remaining search results in different orders. This could be because they interpret the snippets on the results pages differently, or because their search intentions differ for identical query strings.

4.3 Task Variability

To understand why the members of this close-knit community sent the same query but clicked on different pages, we scrutinised those queries and their associated clicks. We found that search task variation

[2]Inter-rater agreement gives a score of how much consensus there in the ratings given by judges. Here we treat each user as a rater, and her click on a URL as a judge.

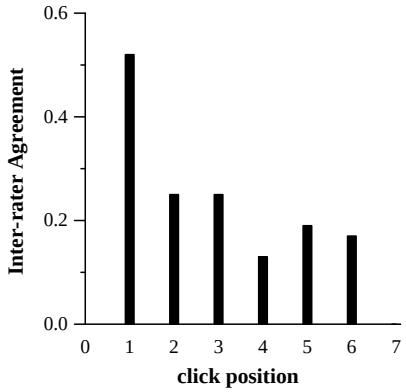

Figure 3: Inter-rater agreement at each click position.

Query type	Examples
navigational	cygwin
	delicious
	java api 1.5
	apache
	ctrl alt del
broad informational	big o notation
	css tutorial
	software requirement specification
	project management
	bioinformatics
	xml schema
specific informational	c printf
	matrix multiplication
	ascii chart
	sql max
	php date

Table 2: Examples of queries in each categories

is one of the factors that may cause the click difference. Referring to search task classification in the web search context (Broder 2002, Rose & Levinson 2004), we classified our queries into three categories: navigational queries, specific informational queries, and broad informational queries. Here the navigational queries are those whose intent is to reach a particular site (for example, "cygwin" and "delicious", while the informational queries are aimed to acquire information or facts that may be contained in one or more web pages (which are unknown to the user). Here we further divide the informational queries into two groups: specific and broad information queries, depending on whether a query implies a multitude of facets or interpretations. For example, "binary tree", "c tutorial", and "test plan" would be regarded as broad informational queries while "binary search tree in c", "c strtok", and "sql max" are specific informational queries. Table 2 shows some example queries in each of categories.

We conjecture that users' click patterns may be more similar for navigational and specific informational queries, and less on broad informational queries, on the grounds that the former two types of queries embody a clear information goal, while the motivations behind the broad informational queries may be quite diverse and rich. A user might want to investigate a broad topic ("c tutorial"), or a user

	Inter-rater agreement			
	High	Medium	Low	Overall
Navigational	12	2	0	14 (14.4%)
Specific Inf.	17	12	3	32 (23.7%)
Broad Inf.	9	30	50	89 (66.9%)

Table 3: Number of queries categorised by type and inter-rater agreement of click-patterns. High inter-rater agreement is larger than 0.7, Medium is between 0.4 and 0.7, and Low is less than 0.4.

could target a particular concept ("c strtok") as part of a typical search strategy that ranges from broad to narrow (Spink et al. 2001). Alternatively, a user may first issue a specific informational query, but, on receiving poor results, then re-issue a broad query.

Table 3 shows the distribution of queries at each level of inter-rater agreement (of click-lists) among the three query categories. Overall, there are about 10%, 23%, and 67% of the queries in navigational, specific informational, and broad informational categories respectively. In the high inter-rater interval, where agreement is at least 0.7, 76.3% of queries are in the navigational and specific informational query categories, as opposed to the low inter-rater interval (agreement less than 0.4), where 94.3% queries are broad informational queries.

Note that some broad queries have unexpectedly high inter-rater agreement. We examined the search results and clicked pages for these nine queries and found that all these queries lead to clicks on comprehensive resource pages (four of them are `www.w3.org`); these pages provide easy-to-navigate links to most facets of a topic domain, and so users with diverse information needs can find their information through navigation rather than search. This also indicate that a user's click behaviour is also influenced by the characteristics of a retrieved page.

Three specific informational queries have low inter-rater agreement, and it is not obvious why this is the case. We suspect that the difference in snippet quality for these queries may be the major reason. For example, for the query "c strtok" the users' clicks are divided into the top three documents, which all have the same quality of information (example and explanation), yet the snippets of the three documents are slightly different as shown in Figure 4. The first is highly generic, while the second shows an example code line and the third has a problem diagnosis. This may explain why some users skipped the first and clicked either the second or the third.

We find that both navigational and specific queries have significantly more clicks than broad information queries (un-paired two tailed t-test, $p < 0.03, 0.002$, respectively); the low inter-rater category also has more clicks than the high-inter category, as shown in Table 4. Figure 5 also shows that the click distributions for each query type are different: the clicks from navigational queries and specific information queries are skewed towards the top-ranked page and the top three pages respectively, while the clicks from broad information queries are scattered, although the top-ranked pages attract more clicks.

Tables 5 and 6 also show that the average ranks of first clicks and all clicks for the broad informational

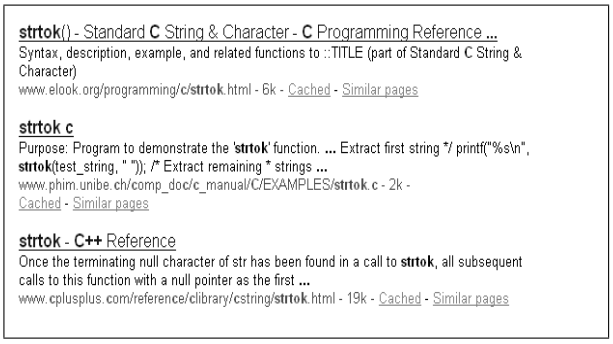

Figure 4: The snippets for the top three results for the query "c strtok".

Inter-rater	High	Middle	Low	Overall
Navigational	1.44	1.36		1.43
Specific Inf.	1.32	1.57	1.47	1.43
Broad Inf.	1.61	1.62	1.79	1.72

Table 4: Average number of clicks per query.

queries are significantly lower than the other two categories ($p < 0.0001$), even more so for the broad informational queries with low inter-rater agreement. This may indicate that these two measures could be taken to identify query types, thus allowing application of different search and relevance feedback strategies to the queries of each category.

4.4 Query Reformulation Variability

Given that our users show different search patterns for different tasks but can still find a set of relevant retrieved pages, does the set of relevant pages satisfy our user's information needs? Relevance of a page can range from topic relevance and situational relevance to cognitive relevance (Saracevic 1996). The relevance judgements we have used in this study are at the topic relevance level, that is, whether a search result is relevant to the search query topic — a TREC-like (Text REtrieval Conference) assessment (Voorhees 2005). The ultimate goal of an information system is to satisfy users' informational needs at a situational and cognitive level; that is, whether a search result is *useful* to a user's task at hand and right to her knowledge level. The best way to answer this question is to interview users at the time of search. In the absence of this information, we can estimate satisfaction by examining users' query reformulation history. We assume that if a user keeps reformulating her query, most likely the information she has so far does not satisfy her information need.

For each query in the selected query set, we located the search sessions, and manually examined whether a query has been modified. For each query, the query modification rate is calculated as the number of sessions with modified queries divided by the total number of sessions. Thus, the higher the query modification rate, the more users modified the query. Overall, the average query modification rate is 41.5%. This high query modification rate indicates that merely delivering a list of highly ranked, topically relevant

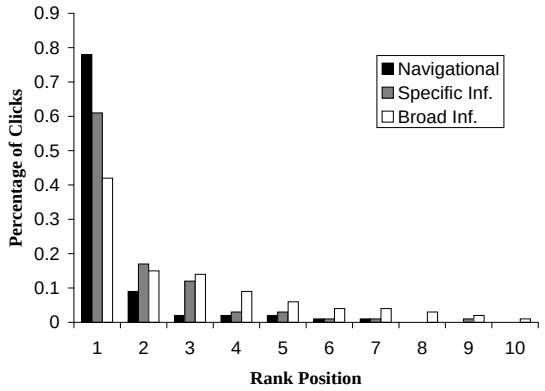

Figure 5: Click distribution at each rank position.

Inter-rater	High	Middle	Low	Overall
Navigational	1.23	1.63		1.29
Specific Inf.	1.22	2.27	2.91	1.77
Broad Inf.	1.34	2.59	4.23	3.38

Table 5: Average rank of first click.

documents is not enough to satisfy 41.5% of information needs.

By closely examining the query modification history, we found that, for queries with high query reformulation rates, the initial query is reformulated by different users into different facets of the original query topic. Take the query "test plan" as an example. This query was sent 17 times (by 12 users) and was modified 12 times. The subsequent modified queries include: "test plan template", "what is test plan", "test plan sample", "test plan technique", "test case distribution", and "test plan acceptance criteria".

A strong correlation is observed between the query modification rate and the inter-rater agreement of click-lists ($r = -0.41, t = -5.26, p < 0.01$) or agreement amongst the first clicked page ($r = -0.38, t = -4.68, p < 0.01$). Task by task, navigational and specific informational queries have significantly lower query reformulation rates than broad informational queries, as shown in Table 7. This may indicate that, if the inter-rater agreement among users' click set is low, there may be a high chance that the query carries multiple information intents.

5 Discussion and Implications

We have explored the community's overall search pattern. Our results confirm that users tend to click on top ranked pages, and consequently that those top ranked pages also have a high click frequency. A detailed analysis of users' click history revealed that the majority of users clicked on pages that are topically relevant to search queries.

However, we inferred that a users' decision to click on a page was limited by what snippets were presented. In most cases, users' clicked pages are relevant but might not be useful, as evidenced by low inter-rater agreement on clicked pages and a high query reformulation rate. We observed that different users

Inter-rater	High	Middle	Low	Overall
Navigational	1.64	1.93		1.68
Specific Inf.	1.56	2.59	3.02	2.08
Broad Inf.	1.92	3.05	4.59	3.80

Table 6: Average rank of all clicks.

Inter-rater	High	Middle	Low	total
Navigational	0.26	0.54		0.30
Specific Inf.	0.31	0.35	0.38	0.34
Broad Inf.	0.36	0.46	0.49	0.47

Table 7: Query reformulation rates in each query category

reformulated their queries into queries on different facets of their search topic. This branch out from the same query to different facets indicates that users' information needs may be different even though they have a similar background and submitted the same query.

In Section 4.3 we examined the task variation on clicking variability and identified a number of different click patterns for different search tasks. We found that users clicked significantly more pages for broad informational queries than those for navigational and specific informational queries. The rank of clicked pages from those broad information queries are significantly lower than those from navigational and specific informational queries. Users agree more on navigational and specific information queries than the broad informational queries. These findings indicate that the value of community clickthrough data varies for different search tasks.

Using clickthrough data to alter rankings will most likely benefit the specific informational search tasks and the homepage finding task, as these tasks are precision oriented and a user's information need can usually be satisfied by just one web page. If a search result list already has a high precision, then incorporating community clickthrough data may not help much, but would not do any harm either. However, if a search result list has a poor precision, then using the community's click frequency data would most likely bring relevant pages to the top of the list as the majority of community members click on the relevant page (especially when the snippet of the page is of high quality). An alternative use of the clickthrough data is to automatically expand the query by using the pages with a high click frequency.

Care should be taken when using community clickthrough data in relevance feedback for broad informational queries. For these type of queries, it may be preferable to deliver search results that cover as many facets as possible (width first). We tried a traditional relevance feedback method (using the text or snippet of clicked pages as a source for query expansion) for two broad queries. The effect was to raise the ranking of pages that are similar to the clicked pages (increasing the depth), without increasing the coverage of more facets of the searched topics. In our opinion, for this type of query, not only the relevance of a page but also the novelty of the page should be considered; and the snippets of each retrieved page should also differentiate one page from each other. Other

work (Carbonell & Boldstein 1998, Zhai & Lafferty 2003) gives examples of how to increase the diversity of search results.

To accommodate the diverse search intentions behind a broad informational query, a search system should support not only the querying activities but also the after-query browsing activity. We could use a domain specific taxonomy to categorise search results as demonstrated in DynaCat (Pratt et al. 1999). In case there isn't a ready-to-use taxonomy for a particular domain, we could use the community query reformulation history to guide the search and search result organisation. Take the query "test plan" from Section 4.4 as an example; we can derive all facets of a broad query from the community's query reformulation history. If we could take one or two top-ranked pages from each of these reformulated queries or facets to form a new list, it would implicitly show a diversified list that covers many facets. As shown in Figure 6, we may even explicitly show the pages under headings derived from the query reformulations, to give users a clearer view what has been retrieved and help users navigate this retrieved document space to get information they need. The more members of the community search on a topic, the more comprehensive an answer list would be.

Finally, to support various search tasks, a search engine should have the ability to automatically identify each query type so that it can apply the optimal ranking scheme for each task. Click distribution and anchor-link distribution have been explored to predicate a users' search goal (Lee et al. 2005). Here, in a community search context, we could use a variety of criteria to classify a query: inter-rater agreement among users' clicks, query reformulation rate, the average number of clicks, or the average rank of first click. All these measures are significantly correlated ($p < 0.01$). Different thresholds should be tested for each measure and data set. For example, in our data set, 74.9% queries with query modification rate greater than 0.4 are broad queries, and 75.0% with query modification rate less than or equal to 0.2 are navigational and specific queries; 83.3% queries with the average rank of the first click greater than 1.5 are broad queries, and 76.3% queries with the average rank of the first click less than or equal to 1.5 are navigational and specific queries.

6 Conclusion

We have explored community search history aiming to identify opportunities to better support community members' information searching tasks. We found that: users tend to click on highly ranked pages and consequently the highly ranked pages also have a high click frequency; the community shows diverse search patterns for different search tasks; and the information needs behind broad informational queries are different even for members of the close-knit community.

Our findings indicate that, the gain of using a community's search history to improve future search experience mainly from the specific informational searches and the navigational searches. For broad information searches, using clickthough data can only bring together similar pages, and this will not satisfy the diverse information needs of the community. We found that users' query reformulation history may provide a potential source for query expansion to broaden the range of web pages returned, and to organise those pages clearly to different facets to highlight the diversity and thus to support browsing activities. Further experiments with users will be necessary to determine the benefit of this claim.

In this study, we focused on the community search of web content. The characteristics of communities and searched domains may vary in other situations. In our future work, we will also investigate the search behaviour of close-kit communities with a closed set document collection, as well as the search behaviours of dynamically bonded community with various document collections.

Acknowlegements

This work was supported by the Australian Research Council.
We would like to thank Yanghong Xiang for collecting some data for this study.

References

Agichtein, E., Brill, E. & Dumais, S. (2006), Improving web search ranking by incorporating user behaviour information, *in* S. Dumais, D. Hawking & K. Jarvelin, eds, 'Proceedings of the 29st ACM-SIGIR Conference on Research and Development in Information Retrieval', Seattle, Washington, USA, pp. 19–26.

Almeida, R. & Almeida, V. (2004), A community-aware search engine, *in* M. Najork & C. Wills, eds, 'Proceeding of the 13th ACM-WWW Conference on on World Wide Web', New York, USA, pp. 413–421.

Broder, A. (2002), 'A taxonomy of web search', *ACM SIGIR Forum* **36**(2).

Budzik, J. & Hammond, K. (2000), User interactions with everyday applications as context for just-in-time information access, *in* D. Riecken, D. Benyon & H. Lieberman, eds, 'Proceedings of ACM-IUI Conference on Intelligent User Interfaces', New Orleans, Louisiana, pp. 44–51.

Carbonell, J. & Boldstein, J. (1998), The user of MMR, diversity-based reranking for reordering documents and producing summaries, *in* W. B. Croft, A. Moffat, C. J. van Rijsbergen, r. Wilkinson & J. Zobel, eds, 'Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval', Melbourne, Australia, pp. 335–336.

Dervin, B. & Nilan, M. (1986), 'Information needs and uses', *Annual Review of Information Science and Technology* **21**, 3–33.

Fagin, R., Kuman, R. & Sivakumar, D. (2003), 'Comparing top k lists', *SIAM Journal on Discrete Mathematics* **17**(1), 134–160.

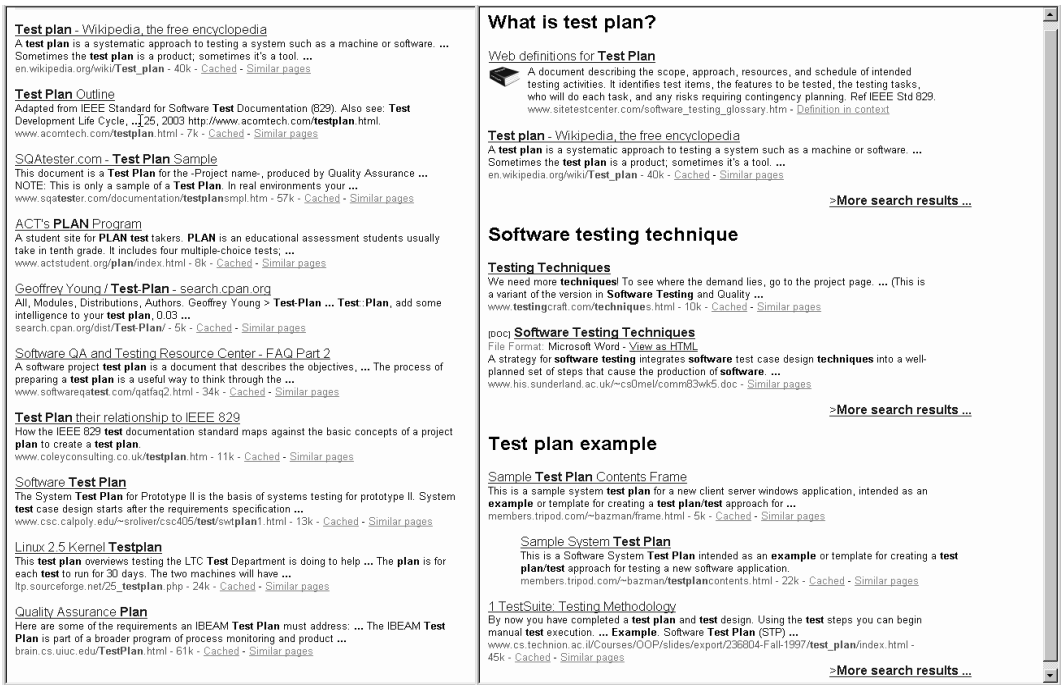

Figure 6: Using query reformulation history for comprehensive answer constructing

Granka, L. A., Joachims, T. & Gay, G. (2004), Eye-tracking analysis of user behavior in www search, *in* K. Jarvelin, J. Allan & P. Bruza, eds, 'Proceedings of the 27st ACM-SIGIR Conference on Research and Development in Information Retrieval', Sheffield, UK, pp. 44–51.

He, D., Goker, A. & Harper, D. J. (2002), 'Combining evidence for automatic web session identification', *Information Processing and Management* **38**, 727–742.

Hoelscher, C. (1998), How internet experts search for information on the web, *in* 'Proceedings of WebNet'98'.

Ingwersen, P. (1992), *Information Retrieval Interaction*, Taylor Graham.

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**, 207–227.

Joachims, T. (2002), Optimizing search engines using clickthrough data, *in* 'Proceedings of ACM-SIGKDD Conference on Knowledge Discovery and Datamining', Alberta, Canada, pp. 133–142.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, P. & Gay, G. (2007), 'Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search', *ACM Transactions on Information Systems (TOIS)* **25**(2), 1–26.

Kim, K.-S. & Allen, B. (2002), 'Cognitive and task influences on web search behavior', *Journal of the American Society for Information Science and Technology* **53**(2), 109–119.

Lee, U., Liu, Z. & Cho, J. (2005), Automatic identification of user goals in web search, *in* 'Proceeding of the 14th ACM-WWW Conference on on World Wide Web', Chiba, Japan, pp. 391–400.

Mat-Hassan, M. & Levene, M. (2005), 'Associating search and navigation behavior through log analysis', *Journal of the American Society for Information Science and Technology* **56**(9), 913–934.

Pratt, W., Hearst, M. A. & Fagan, L. M. (1999), A knowledge-based approach to organizing retrieved documents, *in* 'American Association for Artificial Intelligence', pp. 80–85.

Rose, D. E. & Levinson, D. (2004), Understanding user goals in web search, *in* M. Najork & C. Wills, eds, 'Proceeding of the 13th ACM-WWW Conference on on World Wide Web', New York, USA.

Saracevic, T. (1996), Relevance reconsidered, *in* P. Ingwerson & N. O. Pors, eds, 'Proceedings of the Second International Conference on Conceptions of Library and Information Science (CoLIS): Integration in Perspective', Copenhagen: Royal School of Librarianship, pp. 201–218.

Saracevic, T. (1997), 'The stratified model of information retrieval interaction: extension and applications', *Proceedings of the American Society for Information Science* **34**, 313–327.

Shen, X., Tan, B. & Zhai, C. (2005), Context-sensitive information retrieval using implict feedback, *in* G. Marchionini, A. Moffat & J. Tait, eds, 'Proceedings of the 28st ACM-SIGIR Conference on Research and Development in Information Retrieval', Salvador, Brazil, pp. 43–50.

Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. & Boydell, O. (2004), 'Exploiting query repetition and regularity in an adaptive community-based web search engine', *User Modeling and User-Adaped Interaction* **14**(5), 383–423.

Spink, A., Park, M., Jansen, B. & Pedersen, J. (2006), 'Multitasking during web search sessions', *Information Processing and Management* **42**(1), 264–275.

Spink, A., Wolfram, D., Jansen, B. J. & Saracevic, T. (2001), 'Searching the web: the public and their queries', *Journal of the American Society for Information Science and Technology* **52**(3), 226–234.

Sugiyama, K., Hatano, K. & Yoshikawa, M. (2004), Adatptive web search based on user profile constructed without any effort from users, *in* M. Najork & C. Wills, eds, 'Proceeding of the 13th ACM-WWW Conference on on World Wide Web', New York, USA, pp. 675–684.

Teevan, J., Dumais, S. T. & Horvitz, E. (2005), Beyond the commons: Investigating the value of personalizing web search, *in* 'Proceedings PIA 2005: Workshop on New Technologies for Personalized Information Access', pp. 84–92.

Thomas, P. & Hawking, D. (2006), Evaluation by comparing result sets in context, *in* V. Tsotras, E. Fox & B. Liu, eds, 'Proceedings of the 15th ACM-CIKM Conference on Information and Knowledge Management', Virginia, USA, pp. 94–101.

Tombros, A., Ruthven, I. & Jose, J. M. (2005), 'How users access web pages for information seeking', *Journal of the American Society for Information Science and Technology* **56**(4), 327–344.

Voorhees, E. M. (2005), Overview of trec 2005, *in* 'The 14th Text REtrieval Conference (TREC 2005) Proceedings', Gaithersburg, MD, USA.

Wang, J., de Vries, A. P. & Reinders, M. J. T. (2006), A user-item relevance model for log-based collaborative filtering, *in* M. Lalmas & A. Tombros, eds, 'Proceedings of the Annual European Conference on Information Retrieval (ECIR)', London, UK, pp. 37–48.

White, R. W., Jose, J. M. & Ruthven, I. (2001), Comparing explicit and implicit feedback techniques for web retrieva: Trec-10 interactive track report, *in* 'Proceedings of the 10th Text REtrieval Conference (TREC)', Gaithersburg, Maryland, USA.

Zhai, C. & Lafferty, J. (2003), Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval, *in* J. Callan, D. Hawking & A. Smeaton, eds, 'Proceedings of the 26st ACM-SIGIR Conference on Research and Development in Information Retrieval', Toronto, Cannada, pp. 10–17.