

Cost and Benefit Analysis of Mediated Enterprise Search

Mingfang Wu James A. Thom Andrew Turpin
School of Computer Science & Information Technology
RMIT University, Melbourne, Australia

Ross Wilkinson
Australia National Data Service
Melbourne, Australia

{mingfang.wu, james.thom, andrew.turpin}@rmit.edu.au ross.wilkinson@ands.org.au

ABSTRACT

The utility of an enterprise search system is determined by three key players: the information retrieval (IR) system (the search engine), the enterprise users, and the service provider who delivers the tailored IR service to its designated enterprise users. Currently, evaluations of enterprise search have been focused largely on the IR system effectiveness and efficiency, only a relatively small amount of effort on the user's involvement, and hardly any effort on the service provider's role. This paper will investigate the role of the service provider. We propose a method that evaluates the cost and benefit for a service provider of using a mediated search engine – in particular, where domain experts intervene on the ranking of the search results from a search engine. We test our cost and benefit evaluation method in a case study and conduct user experiments to demonstrate it. Our study shows that: 1) by making use of domain experts' relevance assessments in search result ranking, the precision and the discount cumulated gain of ranked lists have been improved significantly (144% and 40% respectively); 2) the service provider gains substantial return on investment and higher search success rate by investing in domain experts' relevance assessments; and 3) the cost and benefit evaluation also indicates the type of queries to be selected from a query log for evaluating an enterprise search engine.

Categories and Subject Descriptors

H 1.1 [Systems and Information Theory]: Value of Information
H.3.3 [Information Search and Retrieval]: Relevance Feedback
H.3.5 [Online Information Services]: Web-based Services

General Terms

Measurement, Performance, Human Factors, Economics

Keywords

Information Retrieval, Enterprise Search, Relevance Feedback, Evaluation, Mediated Search, Cost and Benefit Analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL '09, June 15-19, 2009, Austin, Texas, USA.
Copyright 2009 ACM 978-1-60558-322-8/09/06... \$5.00.

1. INTRODUCTION

Information retrieval (IR) systems aim to deliver users with relevant answers highly ranked within the first 10 items of their search results. Search engine companies ensure that they deliver the “right results” – including the “right” advertisements for top frequent queries. Large search engine companies are able to leverage the massive numbers of queries to mine correct answers, for example, this could be done by mining query logs to see which results are selected by users, and listing the ones with the highest number of clicks first. Naturally, they are also able to use human intervention if they wish, again due to scale.

An enterprise search service provider would also wish to ensure that the right results are delivered for those “important” queries with high search frequency. Large enterprises might be able to use their search logs to provide enough data to reliably place the right results at the top of a search list. In many other cases human intervention may also be needed. For example, an enterprise could employ a relevance feedback mechanism to ask users to provide relevance assessment explicitly, or have “expert users” manually select the “right answers” for frequent queries.

However human intervention incurs potentially high cost to the enterprise. This cost must be outweighed by the benefit it produces to be viable – it is therefore not good enough to state simply that “precision has been improved by 10%”. The question is: how can we measure the cost and benefit of such human intervention in an enterprise search environment? There is little guidance reported in the literature that might inform an enterprise of when, and to what extent, it is desirable to have human intervention. This is particularly problematic in that many enterprises will not have the level of usage that enables statistical approaches to ensure reliable improvement. Therefore in this study we attempt to answer a simple question: “can we profitably exploit a small group of users' expertise in selecting answers to queries for the benefit of later large number of searchers? And if so how much should we do?”

Arguably there is plenty information in the literature to inform the first part of the question but little that we are aware of that can help a search service provider to determine the level of cost balanced against the level of benefit. This study shows where an enterprise search service should concentrate its effort. There is greater potential savings of searchers' effort with the more frequent queries. Consequently, experts' effort should be invested on these frequent queries. However, at some stage this effort outweighs the benefit. This leads to the formulation of our main research question:

- Is there an optimal point for handcrafting query responses based on query frequency?

with three sub-questions:

- What is the cost of adding expert judgments?
- What is the benefit of adding expert judgments?
- Where is the optimal point for handcrafting query responses?

We attempt to answer the above questions by proposing a cost and benefit analysis method and applying it to a search service provider of a large insurance company.

The rest of this paper is organized as follows. In Section 2, we review related work. In Section 3, we discuss the approach of using domain experts' relevance assessments for improving the search result. We present our proposed cost/benefit evaluation method in Section 4, and demonstrate how to use this method through a case study in Section 5. We discuss limitations and how these may be addressed by future work in Section 6, and our conclusions in Section 7.

2. RELATED WORK

In this section, we review the current efforts in improving search engine performance by exploiting human interactions, namely: relevance feedback based on individual user's interaction, collaborative search based on a group of users' interaction and mediated search based on expert users' interaction. We then discuss evaluation of search engines and associated information systems.

2.1 Relevance Feedback

The effectiveness of relevance feedback in information retrieval has been extensively studied [19]. With this method, we need to identify a set of documents that are relevant (or irrelevant) to an initial query, and then use this set of documents to train a search engine to extract good evidence (another set of words). It is expected that the newly identified evidence together with the initial query would deliver more relevant documents with better ranking to the user. Broadly there are two methods to identify a set of relevant documents.

One method is to treat a number of top retrieved documents from an initial query as relevant – so called pseudo-relevance feedback [2]. This method tries to improve search quality of a query disregarding who the user is.

The other method is to involve the user in the relevant document selection activity and search for documents that are relevant to the user. In this second method, the user can be explicitly asked to make relevance judgment on retrieved documents [27], or the user's interaction with a search engine (such as clicked documents and reading times) is implicitly captured to infer which documents are relevant or irrelevant to the user [9][22][25]. Comparatively, the explicit approach introduces less noise but requires user's time and mental effort in reading through a large quantity of documents and marking them as relevant or irrelevant; thus the implicit approach has been widely researched and recommended.

2.2 Collaborative Search

Collaborative search utilizes the interaction behavior of a targeted community to tailor the search result to that community. Here a

community refers to a group of users who share similar interests or information needs. A community may be predefined, for example, the community members have the same or similar social background, such as the same job role in a working environment; a community could also be dynamically formed, for example, a group of users who send the same set of queries and click on the same set of searched documents [1].

There are two assumptions behind the collaborative approach: 1) users who send the same query may be interested in the same set of relevant documents; 2) a user's click represents a vote to the relevance of the clicked document; thus the more clicks a document attracts, the higher chance that the document is relevant. The documents with a high click frequency can then be gathered as the training set for relevance feedback as discussed in Section 2.1; or the click frequency can be used directly to re-rank a search list. As a result, a search list is improved to suit the community better.

The first assumption – that the documents accessed by some users as a result of a query might be relevant to other users – has been supported by some experiments. For example, Smyth et al. [23] used a hit matrix to record the relative click frequency of retrieved pages per query, and used this information to re-rank future search results for the same query or similar queries. They tested this approach with computer science students for a fact-finding task and found that the subjects using re-ranked lists could answer more fact-finding questions correctly within a given time limit. However, the risk of collaborative search is that the same people may use the same queries for different information needs and therefore find different sets of relevant documents. For example, a study by Wu et al. [29] shows that although a group of computer science students who sent the same query (which was related to their studied subjects), their clicks varied greatly for the broad informational types of queries but less so for the navigational and specific informational types of queries. This study indicated that the effectiveness of the collaborative search approach would depend on the nature of users' information search task.

2.3 Mediated Search

Unlike the above mentioned relevance feedback and collaborative search approaches, where the evidence on relevance is collected directly from user interaction; the mediated search approach (usually used in a library context) utilizes intermediaries (e.g. librarians) who mediate between information seekers (or library patrons) and a search system [24]. Intermediaries interact with information seekers to clarify their search context and attempt to understand what is important for the information seekers' information need; they then apply their knowledge of the available collections and search knowledge to form their strategic search plans, and negotiate a set of search results with information seekers. Comparing to the unmediated search approaches, the mediated search has a higher success rate [14]. It is reported that, with help from librarians, patrons were successful in getting their needed information in 90% cases, judged by the fact that the patrons actually leave with one or more documents and some expression of satisfaction [13].

In mediated search, the knowledge of a user's search context and the knowledge of the document collections or digital libraries play an important role in the information searching process. Some attempts have been made to have a system-simulated mediated search. For example, Muresan and Harper [12] proposed the use of a clustering interface and language model to support users in

exploring a collection, refining an information need and formulating good queries.

2.4 Evaluation

Evaluation has long been a driving force for research on information retrieval. However, the evaluation has been largely focused on the system part; the question often asked is, given a set of queries and associated relevance judgment, how good are the search results from a search system. This is usually assessed in terms of precision, recall and some related measures.

With information retrieval systems being widely used by more and more people for their daily tasks nowadays, the development and evaluation of an information retrieval system that could support a user's search context and task is gaining more and more attention. As Ingwersen and Järvelin ([7], p.314) observed: "The real-life issue in IR systems design and evaluation is not whether a proposed method or tool is able to improve recall/precision by an interesting percentage with statistical significance. The real issue is whether it helps the searcher better solving the seeking and retrieval tasks." Drawing on the traditions of library science and information science, Ingwersen and Järvelin [6] proposed four sets of criteria to evaluate an IR system along four contexts of relevance, namely: system, user and system interaction, work task, and broad socio-organizational and cultural context. Paris al. [16][17] also argued that it is important to measure a natural language generation system not only by the quality of its output but also by the cost and benefit that the system could bring to all its stakeholders.

Evaluating an IR system within the work and organizational context is even more important when the system is used by an enterprise: we need to ask how much a good IR system would impact on the enterprise as a whole. This question has also been widely researched in the information system community. For example, DeLone and McLean's [3] taxonomy of information system's success includes six categories: system quality, information quality, use, user satisfaction, individual impact, and organizational impact. The first five categories overlap with Ingwersen & Järvelin's four contexts of relevance above. However, the aspect of organizational impact is largely neglected by IR community. We seek to address this issue in this paper.

3. MEDIATED ENTERPRISE SEARCH

There are two issues with utilizing the click data in refining or re-ranking a search result. First, click-through data, whether collected at an individual user level or aggregated from a group of users, is un-avoidably noisy. Users are constrained to what documents are searched and presented to them; therefore their clicks might be biased towards those top-ranked but not necessary relevant documents [10]. Guan and Cutrell's study shows that searchers correctly selected relevant web pages less than 20% of the time when relevant web pages were placed below position 2, and not a single subject correctly selected a relevant web page when it was at position 8 [5]. Second, in the Web search context, even though a group of users have sent the same query in the past, they may have had different intents then or in the future [26].

Now we consider searches in an enterprise context, where users search information for their work related tasks from a managed collection of documents. Here the relevance of a document is influenced little by a user's personal interests, but largely by the user's domain knowledge. If we assume users (or employees of the enterprise) have the capability to identify needed information

should a relevant document be presented, how we can develop a search engine that nearly always ranks those "relevant" documents on the top of the search list?

If we could put all our efforts into developing a more advanced or perfect search engine with better natural language understanding and better weighting and ranking methods, we may one day be able to deliver only the relevant information to user. However, until such a perfect search engine could be developed, we should leverage human expert's knowledge into this ranking process. Here, we propose to simply have human domain experts (or subject matter experts) identify relevant documents to a certain number of queries, we can then use these relevance judgments to re-rank the search result of the query should the query be issued in future.

This approach is very much like the mediated search in a library: here domain experts act as intermediaries. Compared to intermediaries from libraries, domain experts need to be more knowledgeable and experienced with their specific application domains and well versed in the problem area as well as the contents of documents being searched for, but probably have less formal training on formulating search strategies. Although domain experts do not interact with information seekers, domain experts are part of the user population that the search engine is designed for. We refer to this approach as mediated enterprise search in the following sections.

4. COST AND BENEFIT ANALYSIS

In this section, we develop a cost and benefit analysis method to evaluate the cost effectiveness of the mediated enterprise search approach.

When an enterprise makes its decision to deploy a new information system or improve its existing systems and services, the enterprise needs to justify such an investment with a detailed analysis of the associated costs and benefits. Some of cost and benefit can be measured quantitatively – such as time and money, while some may be measured qualitatively – such as customers' and employees' satisfaction and their stress level of using the system. Some elements of cost and benefit can be measured directly while some cannot – for example, the impact of a lost business opportunity due to the failure to give customers the right information on time.

Sassone [20] surveyed eight generic methodologies which have evolved to quantify the cost and benefit of information systems, namely: decision analysis, structural models, breakeven analysis, subjective analysis, cost displacement or avoidance, cost effectiveness analysis, time savings times salary, and the work value model. Each methodology has its strengths and weaknesses and is applicable to different evaluation contexts and purposes. For example, the subjective analysis approach asks decision makers to subjectively determine whether the prospective benefits of an information system are worth the projected costs and this approach is used when the benefits are intangible or uncertain; whereas the cost effectiveness approach is used for choosing among similar information systems or system components.

Of these eight methodologies, the Time Savings Times Salary (TSTS) methodology is most suitable for assessing the direct labor cost and benefit involved in the utilizing additional information service of mediated search. Two premises behind the TSTS method are: 1) a worker's value to an enterprise equals his or her cost to the enterprise; and 2) saving X percent of a worker's

time is worth X percent of the worker's cost. So if an enterprise has Y workers, and is expected to save an average X percent of each worker's time by introducing an information system, and if each worker costs the organization an average of $\$S$ per year, then the annual value of the system is calculated as $X \cdot Y \cdot \$S$. This method is easy to calculate and very intuitive in a time intensive information service environment such as a call center – it is quite straight forward to translate the time saved by the mediated search into the number of additional customers that could be served.

Cost

By applying the TSTS method, the cost of having domain experts' intervening is to convert their time spent on relevance judgments into a monetary value. This can be calculated as follows:

$$\text{Cost} = W_E \cdot T_J \cdot N \quad (1)$$

where

W_E = average wage per minute for expert employees involved in relevance judgments;

T_J = average time spent to make relevance judgment for each query; and

N = the number of queries to be judged.

Benefits

We propose two methods to calculate the benefit:

- 1) Time could be saved by moving each relevant document from a lower rank to a higher rank;
- 2) Increased chances to find needed information or gained successful search as a result of re-ranking.

When a relevant document i of a query j is at a rank r_{ij} from an initial ranked list from a search engine, after manual relevance feedback is applied, the document's rank is promoted to r'_{ij} where $r'_{ij} < r_{ij}$.

In the first method, we assume that a searcher would read retrieved documents one by one from top to bottom, and then the benefit is the cost of reading irrelevant documents before the searcher reaches a relevant document from the original search result:

$$\text{Benefit} = \sum_{j=1}^N QF_j \cdot (\sum_i (r_{ij} - r'_{ij}) \cdot T_d \cdot W_a) \quad (2)$$

where

QF_j = number of times a query j is repeated;

T_d = time spent to judge relevance of a document (here it could be measured by average reading time of a document, or elapsed time between two clicks);

W_a = average salary of all employees; and

N = the number of queries.

For a search task of finding all relevant documents, the benefit as calculated in Equation 2 would be the sum of the benefit of each relevant document whose position is promoted. For a search task of finding a relevant document, then we only need to calculate the benefit of finding the first relevant document from a list.

An assumption behind Equation 2 is that a searcher would read all documents before a relevant document is found. In reality, there are a decreasing number of searchers who would read and click on

lower ranked documents. For example, it is reported that, for an informational query, the probability of documents at ranked position 1, 2, 4, 5, 7 and 8 being clicked is 89%, 33%, 17%, 17%, 6% and 0% respectively [5]. So if the rank of the first relevant document is lower than 7, it is more likely that a user might give up earlier before getting there and thus miss a chance to access the relevant document. Thus the gained successful search (GSS) resulted from re-ranking is:

$$\text{GSS} = \sum_{j=1}^N QF_j \cdot \sum_i (P(r'_{ij}) - P(r_{ij})) \quad (3)$$

where

$P(r'_{ij})$ = the probability that a document i from a query j at rank r'_{ij} would be clicked.

We can transfer gained successful search into monetary value:

$$\text{Benefit of GSS} = \text{GSS} \cdot T_f \cdot W_a \quad (4)$$

where

T_f = the time spent to resolve a failed search.

In an enterprise environment, when a searcher gives up a search, s/he usually resorts to her/his colleagues for information – as a result, two or more people's time would be spent to solve a problematic situation. In this case, the benefit of GSS would be at least doubled.

5. A CASE STUDY

To validate the above method and to determine if having a domain expert marking relevant documents would lead to a significant benefit to other staff issuing the same queries, we approached a major insurance company that runs a call center for serving their existing and potential customers. The call center has about 230 staff; their job title is Customer Service Consultant (CSC). The center has about 38 senior consultants serving as mentors to other CSCs. Mentors are comparatively more knowledgeable with the company's insurance products, processes and several customer service information systems that are in place.

The company recognized that their call center employees (most of them are aged between 25 and 35 years old) learn and work differently from older generations: they are technology-savvy and have a habit of on-demand learning; thus accessing the right information on time is critical for the call center. The company puts all its insurance products and processes into easy-to-understand and act-on documents and makes them accessible through a software application called MAX. CSCs are told that all information they need to interact with customers is contained in MAX and MAX should be their first point of reference on the job.

5.1 Search Environment

MAX provides two access points to information: one is through a browsing shelf: all documents in the collection are categorized and are displayed as in a file directory; the other is through search. There is a "favorite folder" function that is used very often by CSCs: they save those frequently used insurance products and procedures into the folder for their quick access in the future.

When a CSC is assigned to an incoming call from a customer, s/he may answer the customer's question straight away if s/he has the necessary knowledge; otherwise s/he may resort to MAX to get the needed information. If the CSC has trouble in finding the needed information for answering a call, s/he may raise a flag to

request help from mentors on duty. Mentors then will walk to the CSC's desk to provide help. In most cases, a mentor would be able to tell the CSC what search query should be issued and which documents are relevant. Afterwards, mentors may record the cases and notify knowledge managers that some needed information is not in the collection or is hard to locate.

It is crucial for a search engine to rank relevant documents high in a search result list. CSCs have very limited time to examine search result. They are in an intensive work environment – when searching they also have to keep up conversation with the customer and may also need to check the customer's insurance type and past interactions with the company through other applications. Furthermore, when a CSC cannot find the needed information, it would result in not only extended call time but also taking mentor's time.

5.2 Experiment Settings

In order to validate our proposed evaluation method, we set up an experimental system to measure the cost and benefit that the mediated search brings to the insurance company. Details about the experimental system, search query selection, subjects and relevance assessments are as follows.

Experimental System

Our experiment system guides users through four stages.

1. **Entry stage** After experimental subjects log into the experimental system, they first read an instruction page that states the purpose of the experiment and shows a step by step description how to use the experimental system.
The end of instruction page also includes an entry questionnaire that gathers the subjects' search experiences with the company's embedded search system in MAX and their familiarity with searched collection.
2. **Pre-assessment stage** Whenever subjects start a new query, they will be required to fill in a pre-judgment questionnaire. In this questionnaire, they will be asked if they have searched using the particular query before and describe as many likely scenarios as possible for issuing that query.
3. **Assessment stage** Subjects carry out the relevance assessment using the assessment interface shown in Figure 1.
4. **Post-assessment stage** Subjects are then presented with those documents they have judged to be relevant or partially relevant. They could re-judge these documents and adjust their scenarios for the query if they wished.

All subjects' interactions with this experimental system were recorded and time-stamped automatically.

Search Query Selection

The experimental queries were extracted from the company's query log recorded from 2008-04-11 to 2008-06-04. The query log has 74,270 entries – about 1,768 queries per day (42 days in total – logging function was not turned on for 10 consecutive days during this period.). Query length is 1.9 words on average. There are 14,632 distinct query entries in the query log. Like many other search logs [30], the frequency of searched queries follows Zipf's law: a few queries were searched very frequently, while a large number of queries were searched less often, as shown in Figure 2.

In order to answer our research question about where is the optimal point for handcrafting query responses based on query frequency, we need to select a set of queries with different search

frequencies. We ranked the queries from the query log from highest frequency to lowest frequency and picked 50 queries at various positions according to the following sampling rule:

$$\text{query}(i) = \begin{cases} b \cdot \text{query}(i-1) & i > 1 \\ 1 & i = 1 \end{cases}$$

The aim of this rule is to achieve a balanced sampling of high and low frequency queries. The coefficient $b (=1.21)$ was picked to fit the sampling within the range. A decimal fraction is rounded up to the next highest integer. Queries are numbered from 1 to 50 accordingly. As vertical lines shown in Figure 2, this sampling method would take more queries with high frequency, but queries with lower frequencies would also be sampled. Together, the 50 queries had been searched 14,077 times (mean = 281.54, std = ±400.2) over the 42 days in the query log.

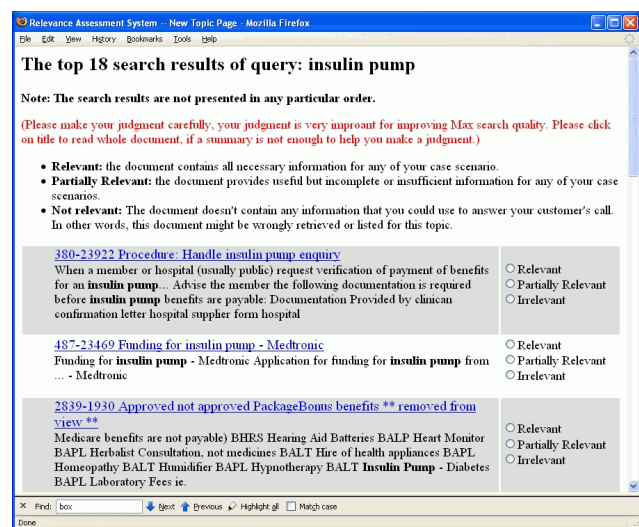


Figure 1: Relevance Assessment Interface

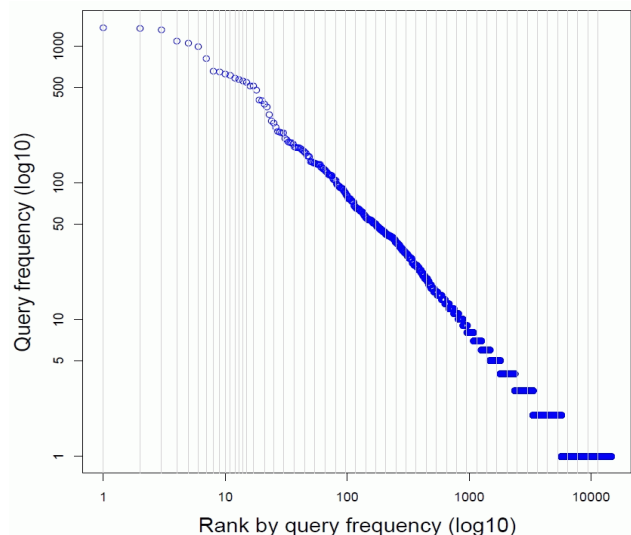


Figure 2: The distribution of query frequency with logarithmic scale on both axes. Vertical lines indicate the points where sample queries are chosen for the study

For each query, the top 20 documents were retrieved by using our open source search engine Zettair¹, except for one query which had 16 documents retrieved. The search engine used the Okapi BM25 ranking model [18] (with $k1=2$, $b=0.75$) for document weighting and ranking. The retrieved documents were used as the candidate documents for relevance assessment.

Experimental Subjects

Ten team leaders or mentors from the insurance company were recruited as experimental subjects to make relevance judgments and were paid for their time. In the entry questionnaire, questions included:

- “How well are you familiar with the contents in MAX?”: not at all(1), a little bit(2), somehow(3), familiar(4) and extremely familiar(5);
- “How often do you use MAX search function?”: every time when I use MAX(4), most of times(3), occasionally(2) and never(1).

Their answers showed that they were very familiar with the content (mean = 4.1, std = ± 0.6) and used their MAX embedded search engine very often (mean = 3.25, std = ± 0.8).

Relevance Assessment

When subjects did their relevance assessments, they could judge the relevance of a document based on the snippet of the document. If a snippet did not provide enough information to make a judgment, they could click on the document title and that would lead them to the full document. They could then make a relevance judgment at the document level. The relevance judgments were on the following three-valued scale:

- **Relevant:** The document contains all necessary information for any of your case scenarios.
- **Partially Relevant:** The document provides useful but incomplete or insufficient information for any of your case scenarios.
- **Not relevant:** The document doesn't contain any information that you could use to answer your customer's call. In other words, this document might be wrongly retrieved for this query.

To avoid ranking bias, the documents from the original ranked list were displayed in a random order. Subjects were told that these documents were not ranked in any particular order.

Using the expert relevance judgments, the original ranked list from the search engine is then re-ranked in the order: relevant documents come in the first tier, partially relevant documents in the second tier, and irrelevant documents in the last tier; the rank of documents within a tier is the same as their rank in the original list.

5.3 Results

To apply the proposed cost and benefit evaluation method, we need to understand how much time experts need to assess relevance; how consistent they are – a measure of reliability; whether the queries actually are returning relevant documents; and

where the relevant documents appear in the ranked list. This section discusses these issues.

Judgment Reliability

In our experiment setup, 14 out of the 50 queries were judged by two CSCs. Overall, the Kappa coefficient² of inter-rater reliability is 0.64, which indicates substantial agreement between two CSCs. In particular, among 280 paired judgments (14 queries · 20 documents), two CSCs agreed on 217 (78%) of them. Most of disagreements (63 pairs) occur in grey areas: relevant versus partially relevant (28) and partially relevant versus irrelevant (31). There are only 4 pairs where one CSC judged a document relevant while the other CSC judged it irrelevant.

Relevance judgment consistency could vary quite widely in different contexts. For example, Lesk and Salton found only 30% agreement in relevance judgment between the authors of queries and non-authors [11]; however, O'Connor found that by using the same document corpus, the agreement rate could be as high as 80% between two documentation experts [15]. Our high agreement rate here between two mentors resembles O'Connor's case: where relevance of a document to its query has less variation and the experienced CSCs know the search context well.

Precision and Discount Cumulated Gain

First we show evaluation of the original search results as output from the search engine and the re-ranked search results based on the domain experts' judgments. Figure 3 shows a substantial improvement by leveraging the domain expert's judgment. If we fold the three graded relevance judgments into binary judgments: regarding partially relevant as irrelevant, then the improvements of re-ranked list over original list at position 1, 3, 10, 15 and 20 are 144%, 90%, 41%, 17%, 0% respectively.

More realistically, we can use Discount Cumulated Gain (DCG) [8] to combine three levels of relevance judgment and a document's rank. By using the DCG measure, relevant documents that are ranked low would get less value than those ranked high. Here we assign weight 2, 1 and 0 to relevant, partially relevant and irrelevant document respectively; and we use a discount factor of 2 to model an impatient or time-poor searcher for whom the value of relevant documents drops rapidly if the documents come late. Figure 4 shows the result of applying DCG measure. The improvements of the re-ranked list over the original list at position 1, 2, 3, 10, 15 and 20 are 40%, 23%, 24%, 25%, 18% and 10% respectively.

Cost

Of 1276 relevance judgments collected, 1036 are explicitly recorded where a judgment was made either at a snippet level or at a document level³. Among these explicit recordings: 997 (96%) of them were made by looking at snippets only, only 39 (4%) were made when documents were read. We believe this is because our subjects as mentors are experienced CSCs and familiar with the contents. In fact, data gathered from the pre-judgment questionnaire shows that 41 out of 50 queries were searched before by the subjects.

¹ <http://www.seg.rmit.edu.au/zettair/> The search engine embedded in the company's software is different from the Zettair search engine we used in our experimental system.

² Kappa coefficient measures the proportion of observed agreement above that expected by chance [4]; a coefficient less than zero indicates no agreement and 1 almost perfect agreement.

³ This logging feature was not turned on for the first two users.

On average, it took CSCs 162.88 seconds (that is 2.71 minutes) to judge a query, which is 8.20 seconds per snippet or document. The salary of these experienced CSCs is approximately \$1.00 per minute (about \$50K per year). According to the Equation 1, the cost of making judgments for 50 queries is $\$1 \cdot 2.71 \cdot 50 = \135.5 – which is about \$2.71 per query.

Benefit

To apply Equation 2 for estimating benefit, we assume that a CSC would spend the same amount of time to read a snippet/document as mentors, i.e. 8.20 seconds. A CSC’s average salary is about \$0.90 per minute. According to Equation 2, the total benefit by saving the cost of locating the first relevant document would be \$4,217 in total, about \$0.30 per search on average. This benefit is significant – the saving is 31 times of cost. Considering that the data gathered covers only 42 days, so the potential benefit can be much higher. If we account for every relevant and partially relevant document that have their rank position moved up, then the benefit would be \$45,337, about \$3.20 per search on average.

If we count the clicking probability⁴ of each rank position and apply Equation 3, among the 14,077 searches in total, there are 6,044 (43%) successful searches gained, should the relevant document be moved up to the top. In our on-site observation, we observed that if a CSC failed to find needed information, it would take the CSC another 2 to 3 minutes to either browse the shelf or ask help from a mentor. So the money saved by this increased success rate would be at least about $6,044 \cdot 0.9 \cdot 2 = \$10,879$.

Discussion

Our case study has shown that the insurance company would get substantial benefit by investing in relevance judgments. Since the cost for assessing a query is fixed, the more a query is searched, the more benefit the company would gain. Figure 5 shows the relationship between a query’s search frequency and its benefit: queries that appear in the top right area are those with high search frequency and gain the most benefit: the Pearson correlation⁵ between query frequency and benefit is 0.67 with 95% confidence ($p < 0.0001$).

Back to our research question about where is the optimal point for handcrafting responses to queries based on query frequency. If we put cost and benefit on each side of equation, we could get the answer that the cost for judging a query would be justified if the query is searched 9 times more (\$2.71 cost per query judgment, \$0.30 gain per search). As shown in Figure 5, there are 12 queries from the bottom two areas that were searched less than 9 times, among them, query 39 already has relevant documents ranked on the top of its candidate list; queries 43, 47, 49 and 50 do not have any relevant document retrieved in their corresponding candidate list; the remaining seven queries (40, 41, 42, 44, 45, 46, 48) have relevant documents moved from lower ranks to the top, however, because of their low search frequency, their gained benefit is only marginal. The 38 queries from the top two areas were searched at

⁴ We used click probabilities of documents at various rank position as reported in [5]: $P(1) = 0.89$, $P(2) = 0.33$, $P(3 \sim 6) = 0.17$, $P(7) = 0.6$ and $P(8 \sim 20) = 0.06$.

⁵ Pearson correlation is a common measure of the correlation between two variables. Its value ranges from -1 to 1: the closer to 1 a value is, the more positive correlated the two variables.

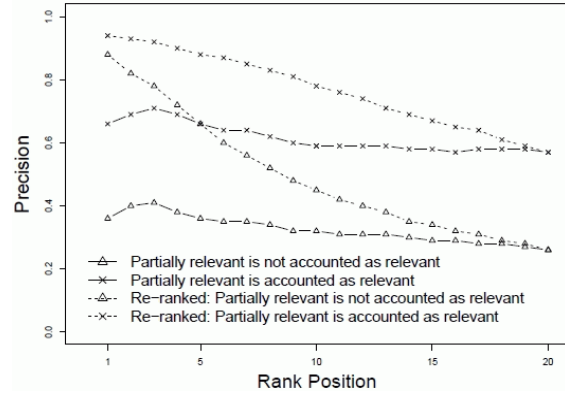


Figure 3: Precision of search results with and without domain expert judgment

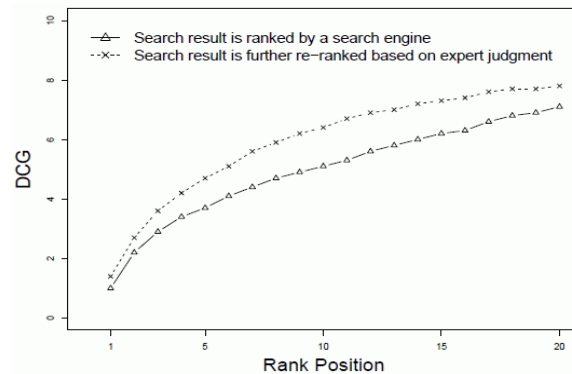


Figure 4: DCG of search results with and without domain expert judgment

least 9 times. The 19 queries from the top-left area gain zero benefit as 17 of them already have the relevant documents ranked on the top (except 35 and 38 that do not have any relevant documents in their candidate list). The remaining 19 queries from top-right area are those queries that benefit most: these 19 queries together (that is 38% of total selected queries) contribute 99.5% (\$4,198) of total gain, moreover the top 15 queries contribute 94.5% (\$3,986) of total gain.

The duration of our log data is only 42 days. Naturally, we would like to know whether our selected queries are representative of the on-going business so we have confidence of scaling up these findings to a longer term, say a whole year. We divide the query log of 42 days into 6 consecutive weeks and we observe a similar distribution pattern over 6 weeks as shown in Figure 6 – our top 15 sampled queries had higher search frequency almost every week. Thus we can scale the benefit gained in 42 days from judging the 50 queries on an annual base, that is: $\$4,217 / 42 \cdot 365 = \$36,648$.

For queries that have their relevant documents moved from a lower ranked position to the top of their candidate list, the benefit for these queries is only marginal because they had low search frequency in the logging period. These queries might gain much more if we had a log of a whole year. In fact, there are 199 queries that were searched every week, and 734 queries every two weeks, although some of these queries were searched less than 9 times in 42 days, their re-occurring pattern indicates that these queries would have been searched more than 9 times over an extended

Table 1. Estimated benefit for queries that could be searched 9 times and more over a year time

	Number of queries	Cost (\$)	Estimated Benefit (\$)		
			1 year	2 years	3 years
Weekly	199	539	30,755	61,511	92,266
Fortnightly	734	1989	42,176	84,353	126,529

logging period. Table 1 shows the estimated benefit of these queries. The yearly benefit is calculated as following: taking weekly queries as an example, according to the above analysis, about 38% of these queries can benefit from the mediated approach, the average search frequency of these queries during 42 days is 156 (58 for queries searched fortnightly), each search would save \$0.30 by moving the first relevant document to the top, thus the total benefit would be: $0.30 \cdot 199 \cdot 0.38 \cdot 156 \cdot 365 / 42 = \$30,755$. This number would help the company to decide if they should invest on mediated enterprise search.

Our data also indicates that those queries with a lower search frequency tend to have irrelevant documents ranked high: there are 17 out of 50 queries whose first ranked document is irrelevant; and nearly half of them (8) are in the bottom 10 queries – these queries brought only marginal gain. This might not necessarily mean that these lower ranked queries should be ignored. In fact, if we take one year as the breakeven period, some lower frequency queries that were only searched once a month would be searched more often over a year and therefore be potentially profitable. In our on-site observation, we also observed that when a CSC met such a search result, s/he would either spend a lot of effort to find the relevant document or resorted to mentors to find the relevant document and then bookmarked the relevant document for later reference. In the entry questionnaire, they were asked in what circumstances they added an item/document into their favorite's folder, out of eight responses: six subjects chose "the document that is referred very often" and seven chose "the document that takes time to search for". When a person's favorite folder is getting larger, that brings the problem to find the needed document again, although this may not be reflected in the query search log. Further study needs to be conducted to determine if improved ranking of a query with low search frequency would bring more value in the longer term.

The cost and benefit analysis indicates that providing precision/recall figures to an information service provider tells only half of the story. The distribution of queries with high precision along search frequency is also important: a query with high precision but low search frequency is of much less value than a query with high precision as well as high search frequency.

6. LIMITATION AND FUTURE WORK

Our case study shows that it is highly beneficial for a company to adopt the mediated search if the company has a rich digital library to support its business. However, further work would be necessary before we could generalize the findings from this study, as their applicability may be limited due to some omissions and assumptions used in our case study.

The benefit as calculated in Equations 2 and 3 does not take into account some indirect factors such as the learning effect. A user may read those irrelevant documents before a relevant document when s/he issues a query the first or second time, s/he might go to

the relevant document directly if s/he issues the query repeatedly and the ranking of a search result is stable over time. In this case, the benefit from a repeated query of the same user might be lower than we estimated in Equations 2 and 3.

Our cost and benefit method only measures the value of improved search result lists in terms of direct labor cost. It does not measure the intangible costs and benefits and consider users' search experience with a search engine. For example, would an enterprise search engine with better ranking method be able to improve users' search experience and increase usage of the search engine, thus reduce customers' waiting time and increase customer's satisfaction? If these are evident, then we would expect these indirect benefits to outweigh the direct monetary benefits. In the future we will try to investigate these issues by conducting a subjective evaluation through questionnaires and monitoring usage pattern over a longer period of time.

Our cost and benefit method also omitted the possible one-off cost associated with the procurement or development of the facilities to enable the mediated search. Our consideration for this omission was that the facilities would likely be made available as a new feature of the newer version of the software, and therefore no direct cost associated. This might not be true or feasible in other software platforms. If the one-off cost is high, it might take a longer time to reach the breakeven point.

We also note the context in which the mediated search is applied. In our case study, there is a community of users who are working in the same well-defined domain. The information needs are reasonably easy to be interpreted and expressed in query words and the relevance assessment of a document is less influenced by the users' personal characteristics. Would we get the same effectiveness or cost-benefit justification if we apply the same method to a community of users (either pre-defined or dynamically formed) who search on an open collection and with their vague information needs? These are research questions which will require future study.

In this paper, we analyzed the cost and benefit of mediated search. We are aware, however, that some other means can also be used to improve search quality, such as the use of a community's clickthrough data. We are going to conduct further studies to compare the effectiveness of a ranking based on the mediated search and a ranking through learning from clickthrough data.

7. CONCLUSION

In this paper, we presented an evaluation of the impact of introducing a specific search feature on an organization. In particular, we proposed a method to assess cost and benefit for an information service provider to invest in mediated enterprise search. We also presented a case study that applied our cost and benefit method. The case study shows: 1) domain experts have high agreement on a document's relevance to its searched query; 2) by having domain experts involved into the improvement of search results and putting more effort on frequently searched queries, the service provider gains substantial benefit.

This study demonstrates that it needs more than recall/precision to determine whether a particular ranking strategy is valuable or not. The proposed cost and benefit analysis method would provide researchers a more complete picture of search; enable service providers to understand whether some technology investment might be of value and how to select a suitable set of queries to evaluate an enterprise search engine.

It is clearly worthwhile for expert intermediaries to invest their effort in mediating the search results of high frequency queries issued by a community of users such as we studied here. It would be interesting to apply a similar analysis to a broad context such as publicly accessible digital libraries.

8. ACKNOWLEDGEMENT

We thank Ken LeHunt, Simon Parker, Alan Thomson and Geri Overberg for facilitating the experiment.

We would also like to thank Cecile Paris and anonymous reviewers for their valuable suggestions.

The project is funded by Australia Research Council.

9. REFERENCES

- [1] Almeida, R. and Almeida, V. (2004). A community-aware search engine. In *Proceedings of the 13th ACM-WWW Conference on World Wide Web*, New York, pp.413-421.
- [2] Buckley, C., Salton, G. and Allan, J. (1992). Automatic retrieval with locality information using SMART, In *Proceedings of the first Text Retrieval Conference*, Gaithersburg, USA, pp.59-72.
- [3] DeLone, W. H. & McLean E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*. v3(1), pp.60-96.
- [4] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, v76(5), pp.378-382.
- [5] Guan, Z. & Cutrell, E. (2007). An eye-tracking study of the effect of target rank on Web search. In *Proceedings of CHI 2007*, San José, pp.417-420.
- [6] Ingwersen, P. and Järvelin, K. (2004). Information retrieval in contexts. In *Proceedings of the SIGIR 2004 IRIx workshop*, Sheffield UK, pp.6-9.
- [7] Ingwersen, P. and Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context* (The Information Retrieval Series), Spinger-Verlag, New York, Inc. p.314.
- [8] Jarvelin, K. and Kekalainen, J. (2004). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, v20(4), pp.422-446.
- [9] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of ACM-SIGKDD Conference on Knowledge Discovery and Data Mining*, Alverta, Canada, pp.133-142.
- [10] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, P. and Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, v25(2), pp.1-26.
- [11] Lesk, M. and Salton, G. (1968). Relevance assessments and retrieval system evaluation. *Information Storage Retrieval*. v4. pp.179-189.
- [12] Muresan, G. and Harper, D. (2001). Document clustering and language models for system-mediated information access. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, Darmstadt, pp.438-449
- [13] Nordlie, R. (1999). "User revelation" – a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proceedings of SIGIR 1999*. California, USA, pp.11-18.
- [14] Nordlie, R. (1996). Unmediated and mediated information searching in the public library. In *Proceedings of ASIS Annual Meeting*, v33, pp.41-46.
- [15] O’Conner, J. (1969). Some independent agreements and resolved disagreements about answer-providing documents. *American Documentation*, v20, pp.311-319.
- [16] Paris, C., Colineau, N. and Wilkinson, R. (2007). NLG systems evaluation: a framework to measure impact on and cost for all stakeholders. Position Paper presented in *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Processing*, Arlington, Virginia, USA.
- [17] Paris, C., Colineau, N. and Wilkinson, R. (2007). Bang for buck in exploratory search. *CSIRO ICT Centre Technical Report*. Report Number: 09/197.
- [18] Robertson, S., Walker, S. Jones, S. Hancock-Beaulieu, M. and Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of the Third Text Retrieval Conference*, Gaithersburg, USA.
- [19] Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback from information access systems. *The Knowledge Engineering Review*. v18(2), pp.95-145
- [20] Sassone, P. G. (1988). Cost benefit analysis of information systems: A survey of methodologies. In *Processings of the ACM SIGOIS and IEEECS TC-OA 1988 conference on Office Information Systems*. Palo Alto, USA, pp.126-133
- [21] Sassone, P. G. (1987). Cost-benefit methodology for office systems. *ACM Transactions on Office Information Systems*, v5(3), pp.273-289
- [22] Shen, X., Tan, B. and Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback, In *Proceedings of ACM-SIGIR 2005*, Salvador, Brazil, pp.43-50.
- [23] Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. and Boydell, O. (2004). Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction*, v14(5), pp.383-423.
- [24] Spink, A., Wilsin, T., Ford, N., Foster, A. and Ellis, D. (2002). Information seeking and mediated search study. Part 1. Theoretical framework and research design. *JASIST*, v53(9), pp.695-703
- [25] Sugiyama, K., Hatano, K. and Yoshikawa, M. (2004). Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th ACM-WWW Conference on World Wide Web*, New York, USA, pp.675-684.
- [26] Teevan, J., Dumais, S. and Horvitz, E. (2005). Beyond the commons: Investigating the value of personalizing web search. In *Proceedings of PLA 2005: Workshops on New Technologies for Personalized Information Access*. pp.82-92
- [27] White, R. W., Jose, J. M. and Ruthven, I. (2001). Comparing explicit and implicit feedback techniques for web retrieval: TREC-10 interactive track report. In *Proceedings of the 10th Text Retrieval Conference*, Gaithersburg, USA.

[28] White, R. W. and Kelly, D. (2006), A study on the effects of personalization and task information on implicit feedback performance, In Proceedings of CIKM 2006, Arlington, Virginia, USA, pp.297-306

Australian Computer Science Conference 2008, Woollongong, Australia, pp.117-126

[29] Wu, M., Turpin, A. and Zobel J. (2008). An investigation on a community's web search variability. In *Proceedings of*

[30] Xie, Y. and O'Hallaron, D. (2002). Locality in Search Engine Queries and Its Implications for Caching. In *Proceedings of IEEE INFOCOM 2002*. pp.1238-1247

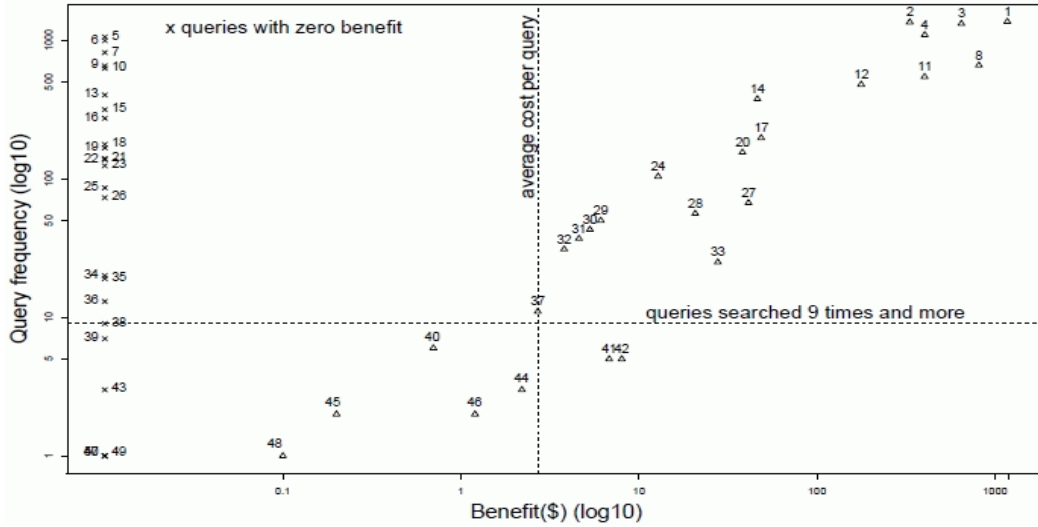


Figure 5: The relationship between a query's search frequency and its benefit.

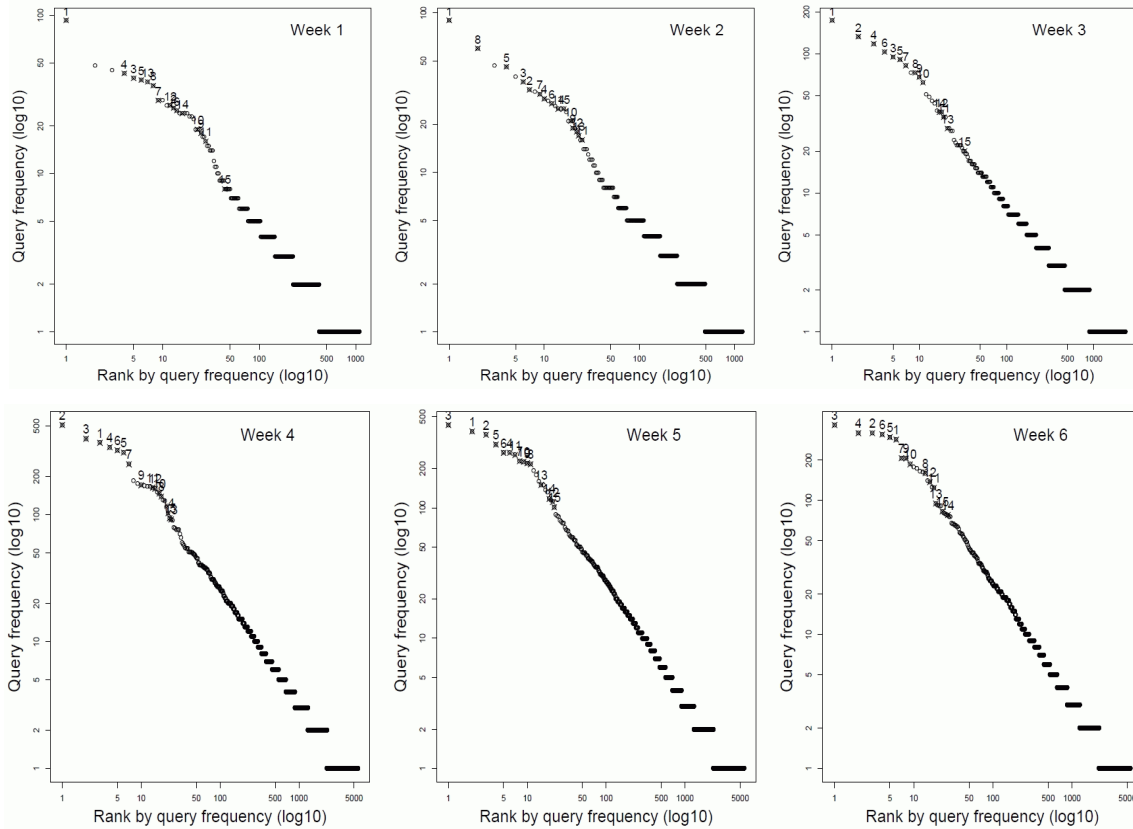


Figure 6: The weekly distribution of top 15 selected queries