Journal of Information Science OnlineFirst, published on April 10, 2007 as doi:10.1177/0165551506077407

JIS

Employing log metrics to evaluate search behaviour and success: case study BBC search engine

Paul Huntington, David Nicholas and Hamid R. Jamali

School of Library, Archive and Information Studies (SLAIS), University College London, London UK

Received 3 September 2006
Revised 6 January 2007

Abstract.

This paper argues that metrics can be generated from search transactional web logs that can help evaluate search engine effectiveness. Search logs from the BBC website were analysed and metrics extracted. Two search metrics – the time lapse between searches and the number of searches in a session – were developed to see whether they could measure search success or satisfaction. In all, 4 million search statements by 900,000 users were evaluated. The BBC search engine possessed a number of functional attributes which sought to improve retrieval and these were subjected to the two metrics to help determine how successful they were in practice. There was some evidence to support the proposition that the search outcome metrics did indeed indicate the effectiveness of engine functionality. The authors argue that this result is significant in that the identification of search outcome metrics will pave the way for assessing the effectiveness of site specific search engines and a greater understanding of the effectiveness of search engine functionality.

Keywords: transaction log analysis; user behaviour; search behaviour; BBC; search engines

1. Background

The authors have developed a powerful technique, which provides robust evaluations of the usage of digital information services, called deep log analysis, which allows for a more detailed and robust analysis than is the case with traditional log analysis [1, 2]. The results of this work have been widely published and essentially show that, in general, the demand for digital information is high and rising but that a relatively high proportion of users (typically 50–66%) appear to make only limited or irregular use of a website [3]. It is not clear whether this has anything to do with problems of retrieving or

Correspondence to: David Nicholas, School of Library, Archive and Information Studies (SLAIS), University College London, Henry Morley Building, Gower Street, London WC1E 6BT, UK. Email: david.nicholas@ucl.ac.uk

navigating to information in the crowded consumer digital environment, or whether, for instance, user search terms and the ubiquitous search engine, which our studies have shown to be a major method of retrieving scholarly material (especially among younger users), were not proving successful in yielding relevant documents [4]. Alternatively, it might be simply a manifestation of massive choice and limited search time. This leads us to hypothesize as to what search or retrieval data could be obtained from transaction logs and whether searching behaviour, especially that associated with search engines, could be measured in the same way as we had described usage.

2. Aims

The major aim of the paper is to investigate search engine searching via transactional logs to see what it discloses about the searching of the general public in a digital environment. The secondary aim is to examine server generated metrics to determine their potential for measuring search success or outcome and in this regard two possibilities are investigated: the number of searches conducted in a session and the lapse in time between searches in a session.

The two metrics presented here for consideration could provide information scientists and web managers with a quick, easy, automatic and 'big picture' means of measuring search success and search engine effectiveness. The number of searches metric possibly indicates the ease or difficulty the user is having finding what they want – the greater the number the greater the difficulty. However, the metric might be read a different way – the number of searches demonstrating a complex interactive search or the fact that the user has stacked a number of individual searches to be conducted in the same session. The lapse of time metric says something about satisfaction/success by establishing whether the user had found anything of interest – the longer the time the more likely they were to have found something. Of course, it too could be interpreted differently; for instance, a lapse in time indicating the user has moved to another task or has gone offline. The paper investigates these various interpretations.

2.1. Scope

A Department of Health funded research investigation of the health users of the BBC website provided the vehicle for the research.[1] The BBC's encyclopaedic website is arguably the most popular website in the UK for factual information of all kinds, and it is also regarded as being highly authoritative. Given the BBC site's stance on effective and quality retrieval the site was also a good one to choose to see what the logs could tell us about the nature and success of health information consumer searching in a general consumer environment, and also about the working of the BBC's search engine which provided additional functionality in order to make the search process more effective.

The BBC announced in 2000 with some fanfare the launch of its very own search engine which would create a trusted and effective environment in which its users could search the web. The BBC engine was of particular interest as:

(1) it is both an alternative search engine of the world wide web (WWW) and a search facility for the BBC's own website;

(2) it attempts to provide a regulated and controlled search environment in order to protect the user from the worst excesses of unregulated web searching – lots of irrelevant or indigestible links – and we could test the effectiveness of the BBC attempts using our success metrics;

(3) there was evidence that for certain groups of health users the BBC engine was the preferred method for retrieving data [5].

We believe this is the first published study of the BBC engine.

The BBC search facility aims to be the 'safest' search engine in the UK. To this end editors research popular topics and recommend the best websites, labelled 'best links', available. These returned 'best links' triggered by selected keywords are an important part of the functionality of the search engine designed to make searching easier for users. It also highlights useful links from BBC News, Sport, Entertainment and Education. The search engine is dedicated to serving a UK audience and results have a UK slant, the intention being to provide added search precision. Also in the attempt to improve precision the BBC search engine features two additional functions. Firstly, there is a function called 'scope of search'. This enables users to conduct their search within a specific subject area or context (e.g. health). The second function is 'tab of search', which refers to the digital universe the search is limited to. The default option is 'allbbc' and the other main user options are 'WWW', 'recipes' and 'sport'.

3. Previous literature

Given endemic end-user searching (disintermediation), the huge popularity of search engines like Google and the availability of search logs, it is surprising that more research has not been conducted on search success. We shall however look at the research that has been conducted (for a detailed review of the literature on web searching research see [6–10]). Jansen et al. [11] analysed search transaction logs of a set of 51,473 queries posted by 18,113 users of Excite and looked at the number of search terms used, changes in queries during a session and the rank frequency distribution of the most used search terms. They found that, on average (arithmetic mean), users made 2.84 queries and that the queries contained 2.35 terms. The first query was labelled a unique query and subsequent queries a modified query. They found that 67% of users had conducted just one query. The research we present in this paper develops some of the work of Jansen et al. and takes their work a stage further by hypothesizing that derived metrics can be outcome metrics of the search process. Jansen et al. [12], using the same dataset, also looked at the misuse of Boolean operators. Jansen and Spink [13] conducted a two-year study of AlltheWeb.com European users. They found some decline in query length, with extremely simple queries. The majority of web searchers viewed fewer than five web documents, spending only seconds on a web document. Web searching by these European users tended toward greater simplicity from 2001 to 2002. Queries decreased in length and sessions were shorter.

Jansen and Spink [14] undertook an analysis of nine search engines and looked especially at the trends in the numbers of one-query sessions and one-term queries. They found that session lengths were not increasing. However, they did not investigate the number of queries made per session. Chau, Fang and Sheng [15], whose research focused on a search engine that was limited to a website, included the number of queries per session in their study. They concluded that their finding was consistent with those of other studies (e.g. [16 and 17]), in which most sessions were very short and contained only one or two queries. The authors reckoned there were two possible reasons for the lower number of queries submitted to their search engine per session when compared with results reported in previous studies. First, their study used a different definition of sessions involving time-out and cookie information, which would have resulted in a larger number of sessions. Second, it was possible that many users were able to find the results they wanted in the first query and therefore did not need to modify their queries to perform a new search. They speculated that this result might indicate a special characteristic of website search engines because they need to index only a limited number of pages; higher precision and recall in the search results may be achieved more easily. At the same time, users may have a better idea of what they want specifically when using a website search engine, so they may be able to formulate better queries in the first step. Therefore, there is a smaller need for users to modify their search queries.

Researchers of online searching and search behaviour have, however, generally neglected the analysis of server transaction files. Research has nonetheless developed models of search behaviour based on task observation. Thus Navarro-Prieto, Scaife and Rogers [18], for example, modelled cognitive strategies in web searching. They identified three different general patterns of searching: a top-down strategy, a bottom-up strategy and a mixed strategy. A top-down strategy is where users start from the general and then narrow their search. The bottom-up strategy is when users searched on a specific keyword, while in a mixed strategy users use both. Hsieh-Yee [7] presented a simple model to show that information retrieval can be distilled into three major components: system content, system capabilities, and users' personal and environmental characteristics. Choo, Detlor and Turnbull [19] also developed an integrated model of browsing and searching on the web using log analysis and interview data. Spink and Sollenberger [20] examined 1241 mediated searches and developed an enhanced information retrieval task model. They identified four major tasks during the mediated interactions that were contributed to by both the search intermediary and information seeker, that is, the information problem modelling task (covering the successive searching subtask and the multitasking search subtask), search system task, interactive task, and social task.

Authors have identified the importance of assessing search engines although no effort appears to have been made in assessing engines based on user online behaviour. Pearce [21] observed that the search engine used on a company's site often does not return the most appropriate page. He further notes that the return on investment from websites is highly dependent on how efficiently those sites can be searched, a fact that highlights the need for a methodology to assess the effectiveness of search engines. Sanchez [22] noted the importance of search engines and argued that more than half of all web users were search-dominant, meaning that they will go straight to the search box when they enter a site. He also suggested that users were not interested in browsing but rather in finding what they wanted as quickly and easily as possible. More recently, Aula and Nordhausen [23] conducted an observational study in an attempt to model search success with task completion speed. They studied 22 participants who performed three fact finding tasks and two broader tasks. They found different variables, including web experience, to have effects on performing different tasks.

4. Methods

Logs of user search queries were supplied by the BBC for 12 days in 2004. Altogether this yielded 4,048,137 search queries made by 891,129 'users' across 1,035,514 sessions, a testament to the sheer popularity of the BBC website and search facility. For the outcome or success analysis a subset of 18,085 search statements by 1073 users were evaluated. There were no agent queries.

The log file relates to search queries entered by users and provides 10 data fields (Table 1). The first three fields and field 8 are (anonymous) user and session identification fields; fields 4 and 5 record the date and time. Field 6 records the search expression entered, field 7 gives the number of

Table 1
Search log data fields

Field 1: Machine (user) id (cgiper13)
Field 2: Process number
Field 3: Process number
Field 4: Date
Field 5: Time
Field 6: Search expression entered
Field 7: Number of BBC best links returned – best link may be an external or a BBC link
Field 8: Cookie
Field 9: Scope of search – a subject category which represents a specific area within which the user initiated the search. A scope search confines the results returned within a specified number of directories (urls)
Field 10: Tab of search user defined to specify a search within the BBC or the WWW

best links returned, field 9 the scope of search used and 10 the tab of search. The process number fields define a search session.

The data was processed using deep log techniques working with raw logs, which enabled us to extract more from the logs than can be obtained from proprietary software. The logs were loaded into SPSS and the data analysed. More details of the methods can be found in [24].

4.1. Limitations

The research team only had access to the search logs of the BBC site and not to the pages viewed on the BBC site. This is important as search logs provide just part of the search story. Clearly users had a choice as to whether to use menu navigational tools to access BBC information or the search facility to access either BBC or WWW information. In fact users might use both, going so far with menus and then jumping to use the search engine. Alternatively, and perhaps more significantly, the user might get so far with using the search process, find nothing of interest in the returned hits, and start using menus instead.

5. Results

5.1. General search profiling

During the two-week study period between 120,000 and 210,000 users searched the BBC site each day and while they were there they made between 400,000 and 850,000 search requests a day.

5.1.1. Microanalysis In order to obtain an insight into the methodology employed and the way that data are generated, it is best to turn the spotlight on one user's interaction with the search engine. The one-day log of this person's actions is shown in Table 2. It can be seen that this particular person used the search facility 11 times over a 17-minute period between 2.27 p.m. and 2.44 p.m. on 12 March. The decision to return to the search time and time again was taken relatively quickly. Generally there was a gap of 10 to 30 seconds between the re-framed search queries, which argues that the user had quite quickly made up their mind that nothing relevant was forthcoming from the choices on offer.

The user first searched for 'cure of diabetes' within the a–z index; 26 seconds later they re-entered their search expression, presumably because they did not find anything, and typed in 'diabetes'. Just over 2 minutes later they returned to their first search expression 'cure for diabetes' this time searching 'allbbc'. This was followed by a gap of two and half minutes; the user then typed in the expression 'type 2 diabetes' within the scope of search 'news'. This search was followed by a gap of 5

Table 2
Microanalysis of user 1

6b240e635912ebc1ce0e1a570599ecf4			12th March 2003	
		Best link returned	Scope of search	Area of search
14:27:26	cure for diabetes	0	a_zindex	a_zindex
14:27:52	diabetes	0	a_zindex	a_zindex
14:30:05	cure for diabetes	0		allbbc
14:32:39	type 2 diabetes	0	news	allbbc
14:37:41	diabetes	3		www
14:37:53	diabetes	2		allbbc
14:43:30	type 2 diabetes	0		allbbc
14:43:58	type 2 diabetes	0		allbbc
14:44:06	type 2 diabetes	0		allbbc
14:44:07	type 2 diabetes	0		allbbc
14:44:47	type 2 diabetes	0	newsifs	news

minutes. This break might result from the user reading a link returned by the search engine; alternatively the user might have gone elsewhere or just decided to have a break. The user returned to the search facility at 2.37 p.m. and typed in the expression 'diabetes', widening the search first to the world wide web and then 12 seconds later to just 'allbbc'. In the first case the engine returned three best links and in the second two best links. About five and half minutes after the second the user returned to the search screen and typed the expression 'type 2 diabetes'. The user returned to this search five times, the time lapse between each one varying from one second to half a minute, sufficient time to visit a returned link and return to the search screen. In the final search the user restricted their search to 'news' and with an 'international news' (newsifs) scope of search.

5.1.2. Number of search queries conducted Well over a third (37%) of user sessions (visits) recorded only a single search query being entered; nearly half of all sessions (49%) saw between two and five queries made; 9% between six and 10 queries and 5% over 10 queries. On average (Huber's M-estimator) [25], users employed 1.9 queries in a session. Huber's M-estimator used here is not strictly comparable with Jansen et al.'s [11] estimate of 2.84 queries as Jansen appears not to have used a robust estimator. The distribution of searches in a session is skewed and in such cases estimates based on the arithmetic mean will be biased and cannot be relied upon to give an accurate estimate. To compensate for this the robust estimator, Huber's M-estimator, was used. Huber's M-estimator is a weighted mean estimate where extreme values are given less weight. Using the arithmetic mean (so as to make a comparison with Jansen's figure) for our sample our figure turned out at 2.64 and Jansen's estimate falls in the upper boundary of the 95% confidence interval, suggesting that the number of queries in a session has not changed all that much since Jansen's research six years ago (1998), despite the arrival of Google etc. However, using the arithmetic mean delivers a biased estimate of the mean and it seems more likely that, on average, users complete about two searches in a session.

Those limiting their search to 'health' via the scope facility undertook a greater number of single query sessions. Hence, just under a half (47%) of them undertook only one query in a session, 44% made from two to five queries and 9% made six or more. The greater proportion of single query sessions among 'health' limited usages may reflect the way that scope of search function operates. This procedure limits the topic area of the search based on the search expression used and at this point provides user relevant health menus. In part, the menus make it more likely for the user to find content and there is less need to go on to compose additional search queries (in other words the menus take over). Users were also more likely to be limited by a scope of search if they used a single word query, as scope of search is triggered by a list of key words, rather than multiple words in their search expression.

5.1.3. Words in a query Over a third of queries (36%) employed one word, 34% two words, 16% three words and 14% entered four or more words. For those limiting the scope of their search to health, the proportions were very different – 49% described their health search in one word, 28% in two words, 13% in three words and 10% in four or more words. As mentioned above, we might have expected a greater proportion of single word search expressions among health scope section users as this is influenced by the way a health scope search operates.

The relationship between the number of queries in a session and the number of words used to compose a query was investigated and it was found that those people just using one word in a search were more likely to limit their session to just one query. This probably argues that those composing a single word query were more likely to find relevant content and they did not then need to go on to compose additional queries. It also probably reflects the fact that those using a single word to compose their query were less likely to be looking for a complicated topic.

The sheer variety of search topics sought was illustrated by the fact that, of the top 20 search words entered by those limiting their search to the health scope, only two, smoking (1.3%) and sex (1%), accounted for 1% or more of the terms entered.

5.1.4. BBC search engine functionality The BBC search engine sought to improve retrieval by returning best links for searches. However, the logs told us that for most search queries (74%) no best link was returned, for 10% of queries just one link was returned, for 15% of queries two to five and for 1% of queries six links or more were recorded. However, a session analysis provides more illumination. Thus 45% of sessions recorded at least one such link. This suggests that some search queries were poorly phrased and that rewording their enquiry resulted in a best link being returned.

The tab of search function was available to all users. The default option was 'allbbc' so it is no surprise to find that 'allbbc' proved the most popular, being used 42.9% of times; 'www' accounted for 32% of tab use and 'news' 10.1%. The scope of search function is also available to all users.

5.2. Determining search success and outcome

Two metrics were considered:

(1) the lapse of time following a search; and

(2) the number of searches conducted in a session.

In regard to the first metric a long time break following a search might be interpreted as a user reading a link returned by the search engine – perhaps, a sign of user satisfaction. Of course, the metric can be interpreted differently. Thus the user might have gone elsewhere or just decided to have a break (a sign, perhaps, of dissatisfaction or taking stock). In regard to the second metric, the conduct of a number of searches in a session could be argued to reflect the fact that the searcher did not find anything the first time around and hence repeats the search. This would be a sign of dissatisfaction and search failure. However, again there is another way we can read it – the user may be developing a search as a result of the feedback obtained from each search; trying out a range of terms could be regarded as a relatively sophisticated form of searching and users anyway might simply want to return to previous hit lists. There is also the fact that what the user is doing is actually conducting separate and different searches.

We sought to understand the meaning and test the effectiveness of the two metrics. With the BBC search facility options were offered to improve retrieval: scope of search and best link. These options would be expected to have an impact on search metrics, as they were intended to improve retrieval. That is, they were expected to lead to greater satisfaction, as demonstrated by a greater time interval between searches, and less dissatisfaction as demonstrated by searches. If scope of search and best link did indeed have the expected impact on metrics then we would have gathered evidence to support the contention that the metrics did measure what we hypothesized.

5.2.1. Lapse of time Lapse of time is calculated as the difference in time between searches in a session. The time lapse of the last search or only search in a session cannot be computed. About a quarter (21%) of users only made a single search and a time metric could not be established for these users. In all, time lapses could only be estimated for 57% of the sample queries. Still, given the large number of transactions covered we are still working with a large dataset.

Figure 1 records the distribution of the lapse of time between searches for the same topic. Two-fifths of the searches where a time lapse could be established recorded a lapse of 20 seconds and under (Figure 2). Clearly 20 seconds is only sufficient time to scan the resulting returned hits and to recompose the search query or consider a variation in the search strategy. Not enough time to read anything in detail or with care.

There were some big differences between topics in the health field, thus, for instance, queries including the word 'coronary' recorded the highest average lapse of time, about 94 seconds (Huber's M-estimator), while searches including the word 'diabetes' recorded a much shorter time lapse of 49 seconds.

Table 3 records the result of tests to determine whether scope of search and best link had an impact on the lapse in time between searches. The time lapse groupings were compared in a logistic regression model: the grouping 'up to 19 seconds' was compared with '20 seconds to under 3 minutes'; and 'up to 19 seconds' was compared with the gap grouping '3 minutes & over'. In addition to scope of search and best link it was decided to also include the number of words in a search. It was hypothesized that searchers using more words in a search query might be undertaking more complicated searches, or over-complicating their search.

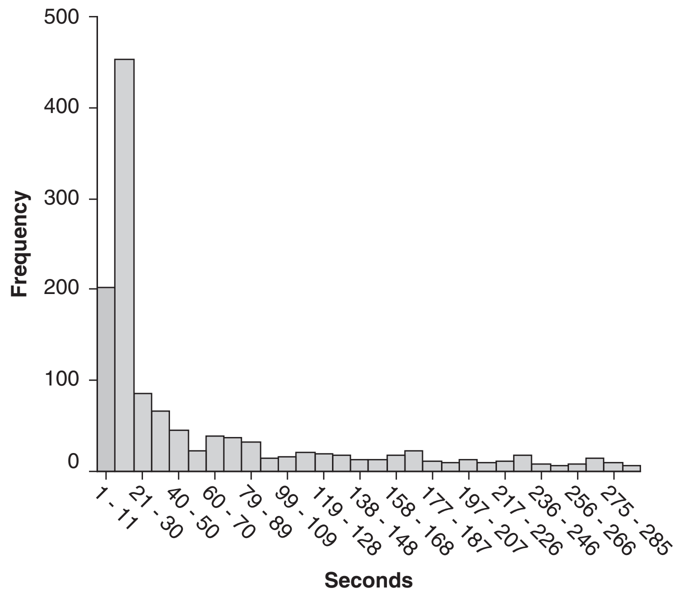

Fig. 1. Lapse of time between searches for the same topic, 1–300 seconds.

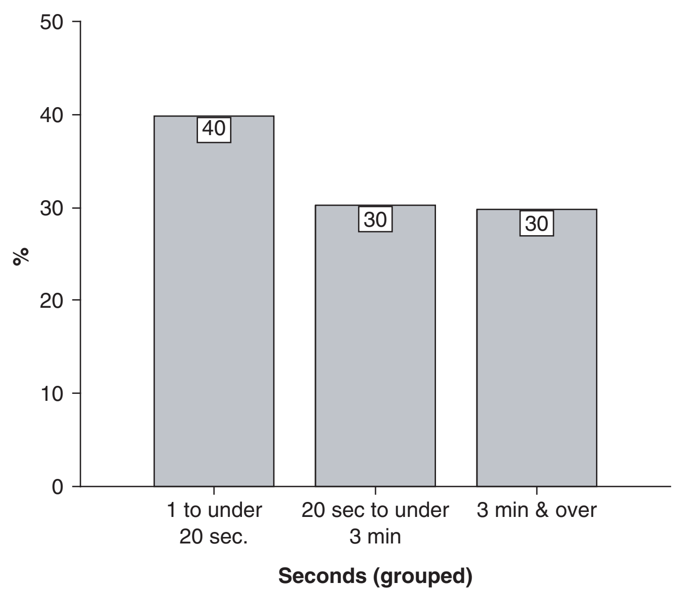

Fig. 2. The percentage distribution of the lapse of time between searches for the same topic – grouped.

The best link and the number of words used were not significant; however, scope of search was significant and impacted on lapse of time. The metric in the case of scope of search was positive and resulted in a greater time lapse between searches. This argues that one of the enhanced retrieval functions did result in a greater time lapse between searches. The result also lends support to the hypothesis that time lapse measures search satisfaction. However, the result is not conclusive as best link was not found to be significant.

Examining the results in more detail, respondents who had used scope to limit their search were nearly four (3.7) times more likely to record a gap of between 20 seconds and under 3 minutes and

Table 3
Comparing time lapses between searches with scope of search and best link using logistic regression

	Between 'up to 19 seconds' and '20 seconds to under 3 min'	Between 'up to 19 seconds' and '3 min & over'
Number of words	Not significant	Not significant
Scope of search		
No		
Yes	3.7*** (0.14)	1.8*** (0.14)
Best link		
No		
Yes	Not significant	Not significant

Levels of significance (Wald's statistic): *$p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 4
Lapse of time between searches by individual health topic

	N	Mean	5% trimmed mean	Median	Huber's robust mean
Diabetes	843	407.6	86.2	13.0	18.6
Cataract	51	161.8	83.8	21.0	28.2
MMR	389	469.8	123.4	22.0	31.0
Cervical	159	707.6	90.5	25.0	31.6
Coronary	111	413.1	198.3	37.0	55.4

nearly twice (1.8) as likely to record a gap of over 3 minutes compared to respondents who had not used scope of search.

The average lapse of time across search topics varied considerably (Table 4). Searches for queries including the word 'coronary' recorded the highest average lapse of time (about 55 seconds) while searches including the word 'diabetes' recorded the shortest time lapse of 13 seconds.

5.2.2. Number of searches conducted in a session Under half of users (44.2%) made one health search, about a quarter (27.8%) made two searches, 18% repeated their searches three to five times and 10% repeated their search six times or more (Figure 3).

Figure 4 provides a cross comparison of the number of searches conducted (grouped) by topic of search. There is evidence to show that people using 'coronary' and 'MMR' in their search expressions conducted more searches: respectively, 63% and 60% conducted more than one search.

Figure 5 shows the number of searches by the average number of words used in a search expression (grouped). Those using more than one word in their search expression were about three and half times more likely to carry on searching their searches than those using a single word. Those using just one word were less likely to go on to repeat their search, just 45% did so compared to 56% of those using two words and 70% of those whose queries consisted of three or more words. This suggests – contrary to expectation – that the more words a user employs the more likely the user is to conduct another search. There is not an obvious explanation for this because one might have expected better precision from searches constructed of more words, and thus a more relevant return. Maybe single word searchers are the bouncers we have encountered elsewhere, people just happy to make broad brush comparisons on the basis of a quick dip into the database [3].

No bivariate relationship was found between best link and number of searches conducted[2] but a relationship was found with scope of search (Figure 6). Those users searching within a subject scope

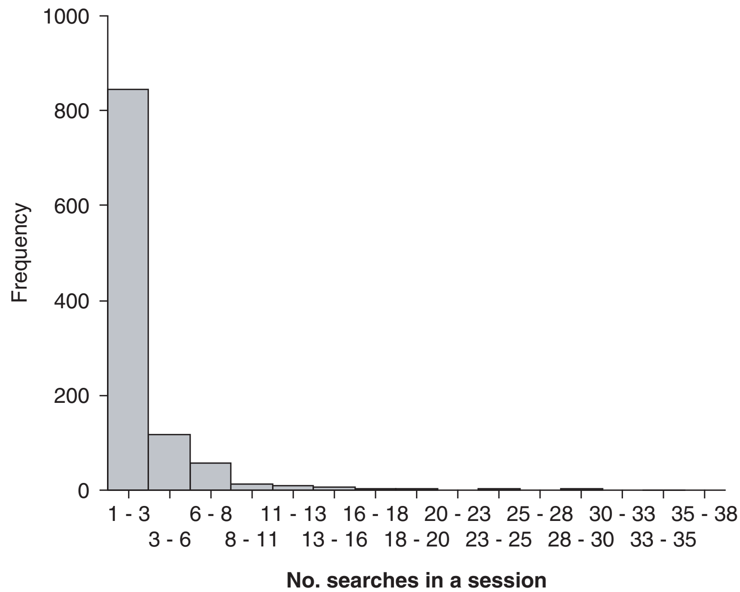

Fig. 3. Number of searches in a session for the same topic.

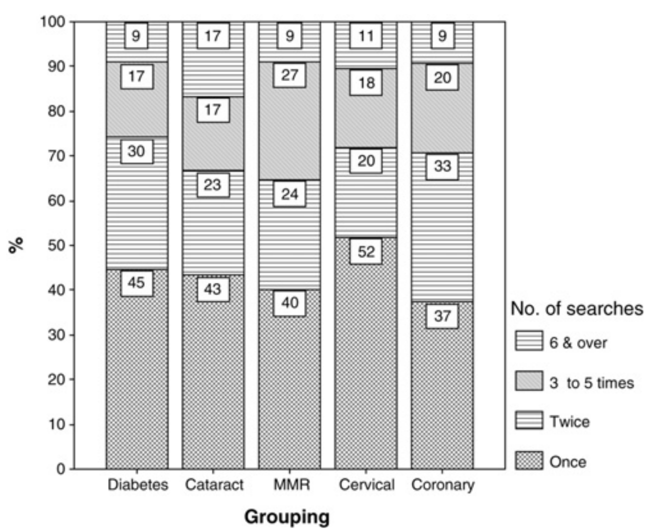

Fig. 4. The number of searches conducted by topic of search %.

limiter were less likely to carry on searching. Nearly two-thirds (62%) of users who applied a subject limit searched just once; however, this was only true of 23% of users who applied no subject limits to their search. Over two-fifths (41%) of no-limits users made three or more searches and this was true for only 17% of those who applied a limit. This seems to indicate that, when a scope of search had been applied, users were far less likely to continue their search.

There was no difference between types of scope search; users employing health and other subject scope limited searches were similarly less likely to repeat a search.

The number of times users repeated their search was divided into three groups and compared. Comparisons were made between dependent variable groupings 'One search' and 'Two searches' and 'One search' and 'Three or more searches'. Comparisons were made using logistic regression

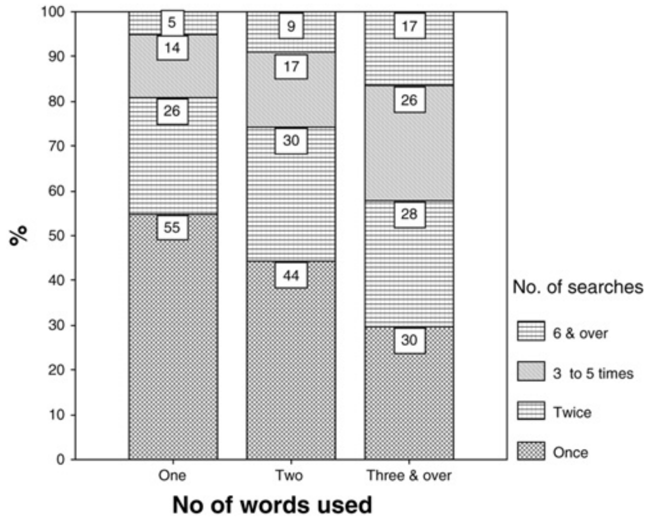

Fig. 5. The percentage frequency share of the number of searches (grouped) by average (median) number of words used in a search expression (grouped).

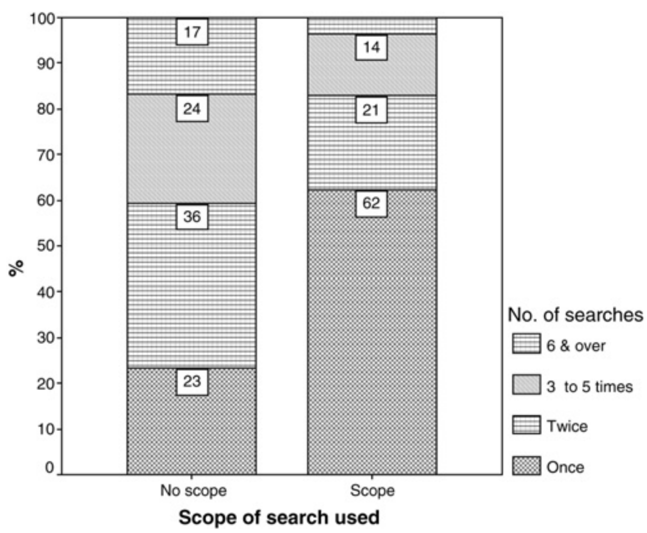

Fig. 6. Percentage frequency share of the number of searches by used scope of search.

and the results are given in Table 5. The independent variables in each case were number of search words used, if scope of search was used or if best link was used.

Again, best link was not found to be significant, although the number of words used to compose a search query and the scope of search were. Those using more than one word were more likely to carry on searching compared to users composing their query with one word. Those using two words were just one and a half (1.6) times more likely to have completed two searches and were roughly the same number of times as likely (1.7) to have made three or more search queries compared to those using one word. Those using more than two words to compose a search query were twice as likely to go on to do two searches and were nearly four (3.7) times more likely to conduct three or more searches. Possibly, some people, using over-wordy queries, were becoming increasingly lost, conducting searches apparently without finding anything. This could be interpreted as users perhaps needing several goes to get

Table 5
Comparing number of searches in a session with the number of words used, scope of search and best link using logistic regression

	'One search' and 'Two searches'	'One search' and 'Three or more searches'
Number of words		
One word		
Two words	1.6* (.19)	1.7* (.21)
More than two	2.0** (.20)	3.7*** (.20)
Scope		
None		
Yes	.21*** (.16)	0.17*** (0.17)
Best link	Not Significant	Not Significant

Levels of significance (Wald's statistic): *$p < 0.05$, ** $p < 0.01$; *** $p < 0.001$.

the desired results from their more complex search queries. Alternatively, it might indicate that the BBC search engine is less able to handle complex queries, forcing the user to repeat their search.

Scope of search was found to be significant. Those using scope of search were less likely to carry on searching; they were five (0.20)[3] times less likely to go on to two searches and six (0.17) times less likely to go on to do more than three searches compared to users who did not use scope of search. This supports the expected view that the facility to improve retrieval did have an impact in the expected direction. Importantly, this result supports our hypothesis that the number of searches conducted is a metric of search dissatisfaction.

6. Conclusion and discussion

In regard to the potential success metrics:

(1) The lapse of time following a search varied considerably depending on the topic, but the number of terms employed in the search statement had no impact, nor did the return of a best link. However, there was evidence that those people using scope of search did go on to record longer lapse times, possibly an indication of search success or satisfaction.

(2) Those using more than one word in their search expression were about three and a half times more likely to carry on searching than those using a single word. No relationship was found between best link and number of searches conducted. However, those people using scope of search were less likely to continue searching.

We proposed two metrics which might be used to assess the effectiveness of search engine searching – one that we have hypothesized estimated user satisfaction and was measured by time lapses and another that was hypothesized to measure dissatisfaction, measured by number of searches in a session. We found evidence to support the hypotheses in that one of the two search facility options (scope of search) had the expected impact on the metrics. That is the metrics did measure the improved retrieval in the correct direction. The result is not conclusive as best link, the other facility option, was not found to be significant and hence no improved retrieval was calculated by the metrics. Arguably, though, best link may not actually lead to improvements in retrieval. Unpublished work supports this and found that users did not trust the best link option [5].

An additional weakness in the argument is that the number of words used in a search expression was found significant in one model (number of searches) but not the other (time lapse). We might expect the metrics to be sensitive and consistent to the same variables as they both measure the same event – search outcomes. The different finding between the metrics in regard to the variable number of words suggests that perhaps the metrics are not as robust as we have argued. However, it should be noted that a time gap could not be estimated for users just completing one search and that users just making one search were more likely to use one word in their search expression. The metric estimating search satisfaction based on a time gap is compromised in that this cannot be computed for all cases. Further research is needed to clarify the findings.

There is a need for easily collected metrics that provide an assessment of site search engine performance and to facilitate a comparison between competing search engine options. The authors propose that such metrics can be derived from transactional search log files. These are routinely collected at little or no cost by websites. Previous research has reported descriptive metrics that can be derived from search logs but this research has failed to recognize that the metrics might reflect user satisfaction or dissatisfaction and further have not taken on board the problems associated with statistics generated from search logs. The failure to address these problems would limit both analysis and findings. In particular, the distribution of time and count values are skewed and in such cases the arithmetic mean is biased and distribution free statistics such as the median should be applied or values should be grouped. Here we have opted for the second approach.

The results provide an encouraging first step towards the development and understanding of metrics for assessing search engines. There are few log transaction based metrics that have potential for assessing search engine use and search behaviour. The two most obvious metrics: the time lapse between searches and the repeating of search expressions, are not wholly robust for the reasons explained. However, further research is needed; firstly a repeat study using a different site and internal search engine is required to clarify generality, and secondly backup qualitative research is needed to understand user intention specifically with regard to session length.

Notes

[1] The digital health consumers of the BBC website. Department of Health, 2004–2005.
[2] Chi = 1.25, 3 d.f., $p = 0.74$.
[3] 1 divided by 0.20.

References

[1] P. Huntington, D. Nicholas, H.R. Jamali and A. Watkinson, Obtaining subject data from log files using deep log analysis: case study OhioLINK, *Journal of Information Science* 32(4) (2006) 299–308.

[2] P. Huntington, D. Nicholas, H.R. Jamali and C. Tenopir, Article decay in the digital environment: a usage analysis by date of publication employing deep log methods, *Journal of the American Society for Information Science and Technology* 57(13) (2006) 1840–51.

[3] D. Nicholas, P. Huntington, H.R. Jamali and T. Dobrowolski, Characterising and evaluating information seeking behaviour in a digital environment: spotlight on the 'bouncer', *Information Processing & Management* (forthcoming) doi:10.1016/j.ipm.2006.08.007.

[4] D. Nicholas, P. Huntington, H.R. Jamali and A. Watkinson, The information seeking behaviour of the users of digital scholarly journals, *Information Processing & Management* 42(5) (2006) 1345–65.

[5] D. Nicholas, P. Huntington and P. Williams, *Digital Health Information Consumers and the BBC Website* (London, UCL SLAIS, 2005).

[6] B.J. Jansen and U. Pooch, A review of Web searching studies and a framework for future research, *Journal of the American Society of Information Science and Technology* 52(3) (2000) 235–46.

[7] I. Hsieh-Yee, Research on Web search behavior, *Library & Information Science Research* 23(2) (2001) 167–85.

[8] A. Spink, Web search: emerging patterns, *Library Trends* 52(2) (2002) 299–306.

[9] I. Fourie, Learning from web information seeking studies: some suggestions for LIS practitioners, *Electronic Library* 24(1) (2006) 20–37.

[10] B.J. Jansen, Search log analysis: what it is, what's been done, how to do it, *Library and Information Science Research* 28(3) (2006) 407–32.

[11] B.J. Jansen, A. Spink, T. Saracevic, Real life, real users, and real needs: a study and analysis of user queries on the web, *Information Processing & Management* 36(2) (2000), 207–27.

[12] B.J. Jansen, A. Spink and T. Saracevic, Failure analysis in query construction: data and analysis from a large sample of Web queries. In: *Proceedings of the 3rd ACM International Conference on Digital Libraries, June 23–26, 1998, Pittsburgh, PA, USA*, (ACM, New York, 1998) 289–90.

[13] B.J. Jansen and A. Spink, An analysis of Web searching by European AlltheWeb.Com Users, *Information Processing & Management* 42(1) (2006) 248–63.

[14] B.J. Jansen and A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, *Information Processing & Management* 41(2) (2005) 361–81.

[15] M. Chau, X. Fang and O.R.L. Sheng, Analysis of the query logs of a Web site search engine, *Journal of the American Society of Information Science and Technology* 56(13) (2005) 1363–76.

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

[17] A. Spink, D. Wolfram, B.J. Jansen, and T. Saracevic, Searching the Web: the public and their queries, *Journal of the American Society for Information Science and Technology* 52(3) (2001) 226–34.

[18] R. Navarro-Prieto, M. Scaife and Y. Rogers, *Cognitive Strategies in Web Searching* (1999). Available at: http://zing.ncsl.nist.gov/hfWeb/proceedings/navarro-prieto/index.html (accessed 15 June 2001).

[19] C.W. Choo, B. Detlor and D. Turnbull, Information seeking on the Web: an integrated model of browsing and searching, *First Monday* 5(2) (2000). Available at: www.firstmonday.dk/issues/issue5_2/choo/ (accessed 2 September 2006).

[20] A. Spink and M. Sollenberger, Elicitation purposes and tasks during mediated information search, *Journal of Documentation* 60(1) (2004) 77–91.

[21] J. Pearce, CSIRO blasts enterprise search engines, *ZDNet Australia* (2003) Available at: www.zdnet.com.au/news/business/0,39023166,20275971,00.htm (accessed 2 September 2006).

[22] M. Sanchez, Adding search functionality to your site, *SiteProNews* 413 (2003). Available at: www.sitepronews.com/archives/2003/nov/24.html (accessed 2 September 2006).

[23] A. Aula and K. Nordhausen, Modeling successful performance in Web searching, *Journal of the American Society of Information Science and Technology* 57(12) (2006) 1678–93.

[24] D. Nicholas, P. Huntington, N. Lievesley and A. Wasti, Evaluating consumer Website logs: case study The Times/Sunday Times website, *Journal of Information Science* 26(6) (2000) 399–411.

[25] P.J. Huber, Robust estimation of a location parameter, *The Annals of Mathematical Statistics* 35(1) (1964) 73–101.