Software Engineering and Management Software Engineering and Management Bachelor Project TIG040 2008-05-20	

Search Analytics

Written by: Jens Bengtsson

Course Responsible: Björn Olsson

Supervisor: Carl Magnus Olsson

IT-universitetet i Göteborg
Box 8718, SE-402 75 Göteborg
Tel 031-772 48 95
Fax 031-772 48 99
www.ituniv.se

Introduction

Information retrieval (IR) as a research field was considered to be rather narrow and mainly of interest to librarians and information scientists. The growth of the World Wide Web (web) escalated during the early 1990s which led to a change as it gave non IR-professionals access to information retrieval through web search (Spink et al., 2001; Jansen & Spink, 2003). Parallel to the evolution of the web, corporate intranets (i.e., intra-organizational networks based on web technology) became an important and growing information environment and to be able to exploit this environment there is a need for efficient IR-tools (Stenmark, 2006; Offsey 1997). One way to improve the performance of these tools that has been recognized as useful is through transaction log analysis (Jansen 2006; Spink & Ozmutlu 2003).

However, there has still been limited research done within the area of information search; the research that does exist has mainly focused on web related search (Spink et al. 2001; Jansen & Spink 2003; Silverstein et al. 1998) although there are some related to intranets (Stenmark, 2005a; 2005b; 2006; 2007; Fagin et al., 2003). Their research has created an initial body of knowledge related to how casual users interact with search engines. To analyze these kinds of data is a useful for search engine performance optimization (Spink & Ozmutlu, 2003) which in the end will lead to better exploitation of resources that exist but are underused (Abecker et al, 1999).

The study presented in this paper relies on using transaction log analysis (TLA – cf. Jansen 2006). TLA has been used to evaluate library systems, information retrieval systems and lately Web systems. It is based upon three stages: collection, preparation and analysis. Collection refers to the process of collecting search data for a specific period of time in a transaction log. Preparation is the part where you clean and prepare the transaction log for analysis and finally you have the analysis which is when you analyze the prepared data. By using TLA I can relate my results to previously performed research (Stenmark 2005b; 2005a; 2006; Spink et al., 2001), I will use a modified version of an already existing software solution to support my TLA research. By doing this I will have control over what data is inserted into my transaction log and that will give me more freedom when doing the analysis.

To be able to meet these research objectives, contact was made with a company, Findwise, who had started to show a deeper interest in search analytics and is an established actor in the enterprise search solution market. Being able to analyze how the users search at their different customers will make them able to enhance their search implementations and thereby increase the customer satisfaction. It will also help them create a model of their work process when implementing and maintaining a search implementation. The way that analysis of user search behavior can help Findwise is that certain conclusions can be drawn when looking at search statistics and these conclusions can lead to practical actions referred to as content tuning, which means that you configure certain aspects of the search implementation to gain more beneficial results for the users search experiences.

This work will contribute in three ways. One, a practical contribution in aiding Findwise with a study of how you can analyze search logs to improve an already existing search implementation and make it more efficient. Two, a theoretical contribution by answering the call for further studies related to the search behavior of casual users (cf. Stenmark 2005a; 2005b; 2006; Jansen et al. 2005). Three, a methodological contribution through the adaptation of TLA and reflections upon its contributions towards the findings made during this study.

In the next section some background work that is relevant to the study is presented. Thereafter the characteristics of TLA are described which then leads to the presentation of Findwise model for information retrieval. This is followed up by the result analysis and thereafter the conclusion.

A Background to Information Retrieval

Studies about information searching have existed for many years but they have been related to conventional IR-systems (Jansen et al. 2000) however during the last decade research geared towards how casual users interact with web search engines started to emerge. For example Silverstein et al. (1998) studied the usage of Altavista by analyzing a transaction log of six weeks which contained more then one billion entries. However, the most consistent work about web searching has been conducted by Spink and Jansen, individually or in cooperation with others, which has established a good research base (Stenmark 2005b).

Spink et al. (2001) studied web queries by users of the excite search engine where they found that the most people use few search terms, view few web pages, and barley use the advanced search features. A small number of queries are used frequently while there is an extensive quantity of unique terms. The findings of the study are compared to two other studies on web search. The web search associated research continued when Jansen & Spink (2003) examined the viewing patterns of commercial web search engine users. The research addressed three questions: How many pages of results do web search engines users' examine? How many documents do the users' view when searching the web?, and how relevant are the documents that they are viewing? The study was performed on transaction logs from Alltheweb.com, a search engine owned by FAST. Their conclusion that the overall information needs of web searchers are not to complex since they often only require one query, that the search engines do a good job when indexing and ranking results since the majority of users only view one results page and that about 50% of the document view will be relevant implying that an average of two documents needs to be viewed to find a relevant result.

Two of the ways search behavior has been studied are to look at variations during different times of the day, and characteristics of the individual queries. An analysis on variations in user search behavior during different times of day was conducted by Ozmutlu & Spink (2003) on search queries from US-based Excite and Norwegian-based FAST web users. Their findings suggested among other things that there were fluctuations in web user behavior over the day and that the characteristics of a query remain steady during the entire day. Jansen et al. (2005) studied the characteristics and the changes in web search behavior on Altavista from 1998 to 2002 using transaction log analysis. The study showed that the users engaged in more interactivity with longer sessions and expanded queries, also it was noted that the most frequent terms only accounted for less then 1% of the total term usage which corroborates Spink et al. (2001) study.

Stenmark analyzed the search performed within a large corporate intranet with the intent to observe search patterns during different time periods. One study (Stenmark 2005a) focuses on the search done during one week which concluded that the number of active users and the number of sessions are higher early in the week and declines as week goes on, the primary contribution was a baseline for more intranet research. Another study (Stenmark 2007) is a longitudinal analysis of search log data from three different years. The primary goal was to further the understanding of intranet users by studying their search behavior and by comparing the results from research done on the public web both similarities and differences between the users could be found. Stenmark (2005b) also conducted a study with the intent to analyze intranet user queries. He goes into more detail by analyzing pairs, triplets and full queries in a later research article (Stenmark 2007) where he draws the conclusion that there are major differences between search terms on the public web and the ones used within the intranet he studied. Although it has a more technical focus Fagin et al. (2003) performed a study of the difficulties with intranet search, where they created a research prototype for intranet search. They also discuss the differences between intranet and internet search and also present the view that there are defining features though they both share a great deal, e.g. there is a conflict between a good searchable intranet and the diverse methods that the information is presented which in many ways mirrors the internet. Intranet queries are also very often jargon-heavy with various abbreviations and acronyms which could be company specific but it is probably not unique.

Jansen (2006) performed a review and created a foundation for search transaction log analysis. He elaborates on the different steps of TLA and also discusses the strengths and weaknesses of the methodology. Of the above mentioned related studies, several make use of TLA to understand the needs and behavior of those using the search engines. (cf. Stenmark 2005a; Spink et al. 2001; Jansen et al. 2005).

Characteristics of Transaction Log Analysis

Transaction Logs are a common way of capturing patterns of users' interactions with IR systems, it is reasonably so because it is a non intrusive way to collect user-searching data from a large number of users (Jansen & Pooch 2001). So what is a transaction log? Jansen (2006) describes it as an electronic record of interactions that have occurred during a searching episode between a search engine and users searching for information on that search engine. The users can be human or computer software searching on behalf of humans, the interactions is the communication exchanged that takes place between the users and the system.

After the data is collected it needs to be analyzed to provide any substantial information, the process of analyzing the data is referred to as transaction log analysis (TLA) (Jansen 2006). For search related research, TLA is used as a tool for investigating the interactions between users, search engines or the content accessed during a search episode. The end goals are to gain a better understanding of how users, content and system interact with each other and with the help of this information improve system designs, search assistance or identify information searching behavior (Jansen 2006).

TLA also works well together with a grounded theory (Glauser & Strauss 1967) approach which focuses on a systematic discovery of theory from data using methods of comparison and sampling, these theories are grounded in interpretation of the "real world" rather then being thought out beforehand (Jansen 2006). Library systems, traditional IR-systems and now, more recently, web and intranet systems have been evaluated by researchers and practitioners with the help of TLA (Jansen 2006). Jansen & Pooch (2001) report on a group of studies that have used TLA while studying web search engines and searching on the web. There is also a thorough review of web-search TLA studies by Jansen and Spink (2005)

There exist some problems with using TLA, as there are no standardized metrics which makes straight number comparisons hard to achieve because data and analysis differ from study to study. Instead, the suggestion is that you compare similarities in trends (Jansen & Pooch 2001; Stenmark 2005b; Spink et al. 2001). Another issue that is present when using TLA is that you do not get the context in which the users are executing their queries and we do not know their reason for engaging with the search engine, nor do we get any information about how satisfied the users are with the results (Stenmark 2005b; Jansen 2006).

The TLA process is described by Jansen (2006) as a three stage procedure; the three stages consist of collection, preparation and analysis. The following table presents the three different stages in a condense format.

Collection	• Transaction log collection gives robust data without being obtrusive to the user
	• The data is collected from real users interacting with the search engine
	• Example data in a transaction log:
	  o IP-address/Session ID
	  o Date**/**Time
	  o Search URL
	• More extended log formats are common

Preparation	• Transfer the data into a database or other analysis software • Cleaning the data, e.g. checking each field for bad data • Cleaning of bad/corrupted data is often done automatically since transaction logs often contain a lot of posts • Normalize the data e.g. turn all queries into lowercase

Analysis	• There are tree levels of analysis
	• Term[1] level analysis
	  o Different term related analysis e.g. high usage terms
	• Query[2] level analysis
	  o Analysis based on complete queries e.g. unique queries
	• Session[3] level analysis
	  o E.g. session interaction, the actions performed by a user within a session

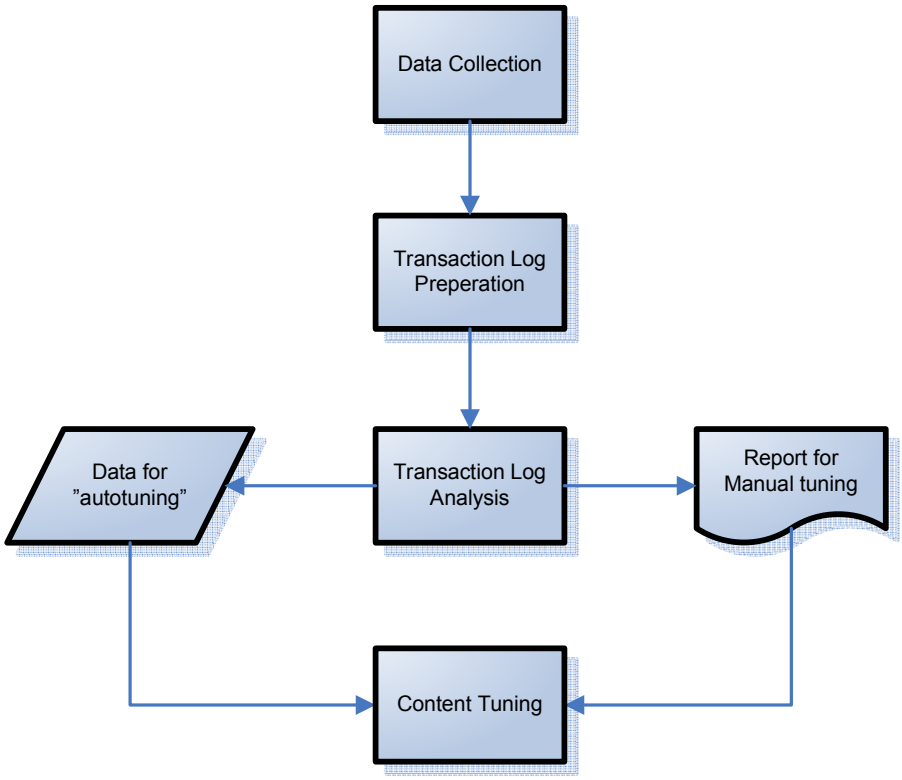
[1] Term: any broken string of alphanumeric characters entered by a user. (Spink et al. 2001)

[2] Query: a set of one or more search terms (Spink et al. 2001)

[3] Session: the entire set of queries by the same user over time. A session could be as short as one query or contain many unique and repeat queries. (Spink et al. 2001)

A model of information retrieval at Findwise

Because of the needs to customize the TLA-model to this specific study I further present the model that I, together with Findwise, defined as the process for optimizing their information retrieval process. It starts out with the data collection; this is done for a defined time span that leads to a transaction log. The transaction log is then prepared by being cleaned and normalized based upon different criteria so that it thereafter can be analyzed. The goal of the analysis is to get a report that can be used to tune the content of a search engine so that it is more efficient and performs more effectively, another goal would be to get data from the analysis that could lead to the content being tuned without any manual labor.



Different search statistics correspond to different content tuning actions e.g. if you have a web site for a physical store there is a large chance that when you do a transaction log analysis you will se statistics that tell you that users are searching for a map. By doing a content tuning action which is called “Best bets” you can then provide a direct link when someone searches for “map” that is positioned as the first result on the search result page which leads the user directly to a map showing the way to the store. Another example that you can look at, if you have a search box that is reachable from anywhere on your site, is the location from which the query took place. If a large amount of people are searching from a page on your site called “Contact” then you need to look at what they are searching for to gain knowledge about what they are expecting to find on the “Contact” page and adjust your site accordingly (Inan 2006).

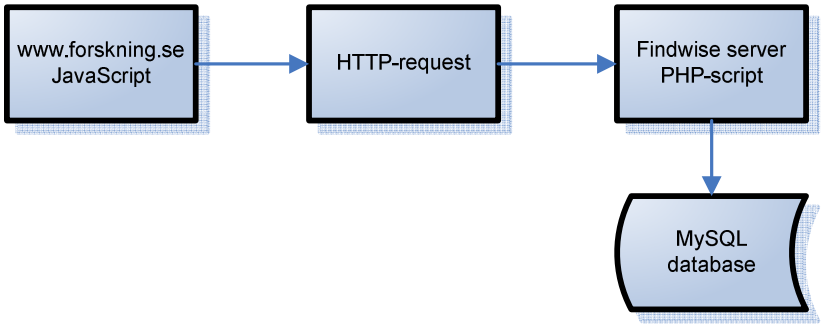
You can look at the search box on a site as a possibility for the users to answer the question “What do you want?” and search analytics as the process of finding out what their response was so that you can provide better answers to their requests in the future.

I will use TLA to analyze a transaction log and look at what kind of conclusions can be drawn from the data with the intent to optimize search engine implementations. I will also create a model that shows how search analytics is incorporated into search engine maintenance by providing an overview of the related parts in a search engine implementation process.

With the help of Findwise contact was made with the Swedish research council which is a Swedish authority under the department of education and the largest government financier of basic research in Sweden. The Swedish research council operations consist of financing and promoting basic research

within all research areas. They host the page www.forskning.se which is a national webpage for research information where they collect and sort material about research and complement it with synoptical information. A part of this is by having the responsibility for making Swedish research searchable. This combined with the fact, as they self put it: “willingly cooperate”, made them a perfect match for the research I wanted to conduct.

A JavaScript was implemented on the search page at www.forskning.se which added a transparent layer that collected data from the users while they were interacting with the search engine. The data collected was passed on to a server which with the help of a PHP[4]-script recorded the data into a MySQL [5]database. This provided a very non-intrusive way of gathering a rich collection of data to be used for analysis. The following model shows the collection of data:



[4] PHP: Hypertext Preprocessor - http://www.php.net/

[5] MySQL - http://www.mysql.com/

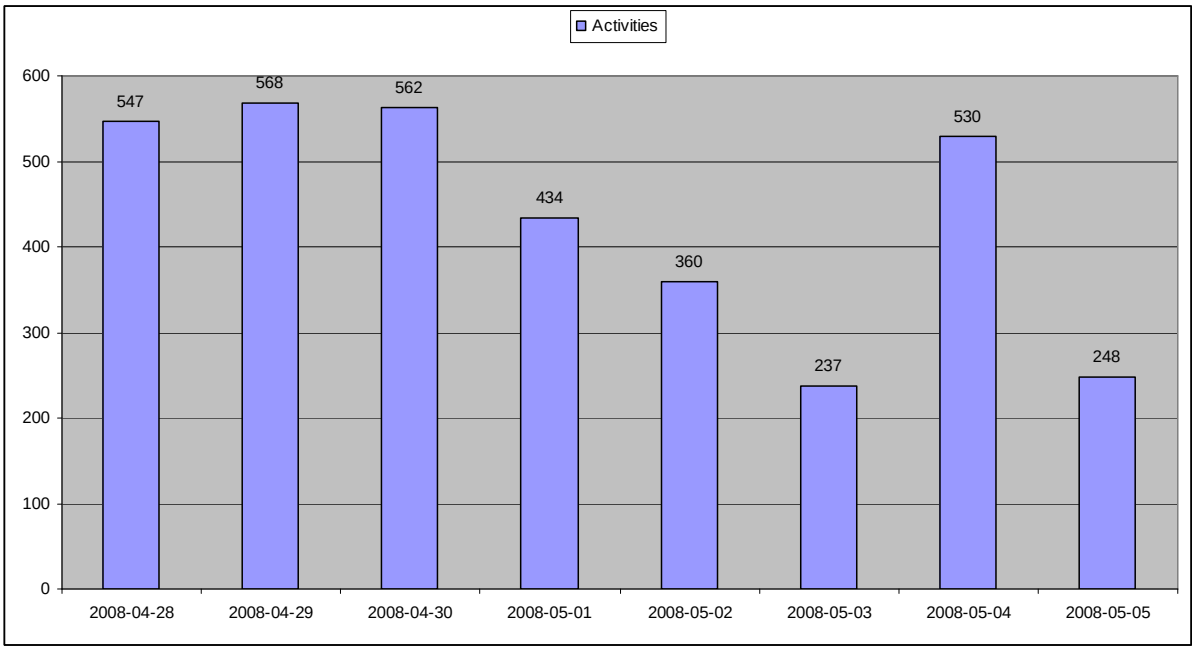
Analyzing Results

Activities

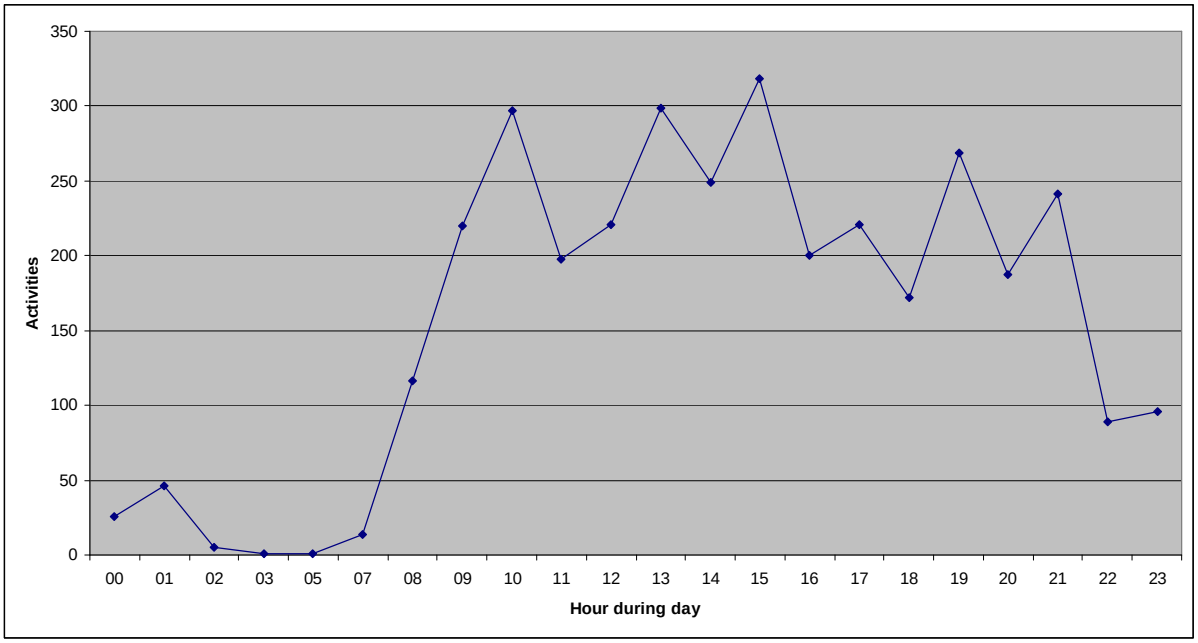
Every entry into the log is seen as an activity, depending on which type of activity that is performed different data is logged. An activity can be the loading of a results page after a query, a click on a result link, switching between advanced and simple search or changing of sorting method. Also navigations within the site (www.forskning.se) that are not related to search are logged as “other”. The parameters that can be recorded into the log for the activities are as follows:

- Query ID – A unique ID for every post in the log.
- Timestamp – With the format YYYY-MM-DD HH:mm:SS e.g. 2008-05-05 10:31:25.
- Session ID – A unique session id from the user.
- Query – The query as entered by the user.
- Ctype – A definition of the activities performed by the user e.g. a click on a result link equals the Ctype “c” and a click on the link for the next results page equals “nav.next”. This is used to distinguish the activities from one and another.
- Weight – A measure of how relevant a clicked result link is in accordance to the query entered by the user. Measured from 0 to 100.
- Page – The number of the current results page e.g. the results page viewed after executing a query is 1.
- Search type – Carries one of the values simple or advanced, used to tell if the advanced search function is used.
- Referring URL – From what page did the user engage the search.
- Clicked URL – If a result link is clicked on the clicked link is saved.
- Results – The amount of hits that match the query entered by the user.
- Sort type – The results page is sorted by relevance by default but can also be sorted alphabetically or by source.

I collected data during the time period between 2008-04-28 13:14:21 and 2008-05-05 12:51:06 (first and last timestamp in the log), this gave me a collection of 3486 activities. The average amount of activities per session id/user was 5,61 with the lowest being 1, a single query without any following activities, and the highest being 96, the effect of a large quantity of navigations to connective result pages. By dividing the activities based on the date they were performed you can see a fluctuation between the days with maybe the most interesting one being that there was less then 50% the activities on Friday (2008-05-03) compared to Saturday (2008-05-04). The following diagram shows the activity spread over the dates that data was logged.



As seen in the diagram there is a major difference between the two Mondays (2008-04-28/2008-05-05) in amount of activities. By plotting the amount of activities by hour of the day to a diagram it can be seen that the amount of activities is lower during the forenoon compared to the afternoon and the logging on 2008-04-28 started and continued during the afternoon while is ended during the forenoon on 2008-05-05. The peak in activities occurs between 10:00 and 16:00 with a drop at around 11:00 to 12:00 probably because of the general user having “lunch hour”. The activity fluctuation is very similar to the one experienced by both Stenmark (2005a) and Otzmutlu and Spink (2003).



Unique sessions

Every user carries a session id in a tracking cookie; this session id was fetched through the JavaScript and entered into the database as a way of distinguishing users from one and another. When the logging finished there had been 612 unique sessions ids recorded from the users.

Queries

By looking at the most frequent queries you can take not of what the users of the site often want more information about. Taking this into consideration you can boost certain types of information resources to give them higher relevance and therefore provide more appropriate results. It is also a good way of finding words and language that might be specific for the site in question and thereby add synonymous to that word so the search engine can give it its correct relevance.

The following table displays the 20 most common queries by unique session ids, the data collection is somewhat limiting because of the rather small activity during the time span of logging but there is a clear indication that the most popular things to look for are related to medicine and biology.

Query	Query count
se hjärnan	20
kloning	12
hjärnan	11
genetik	9
nya biologin	8
genteknik	6
etik	5
östersjön	5
permafrost	5
hålla andan	4
Dextegen	4
VATTEN	4
ADHD	4
stamcellsforskning	4
fetma	4
stroke	4
vetenskapslandet	4
stress fysisk aktivitet	4
hälsa	3
språk	3

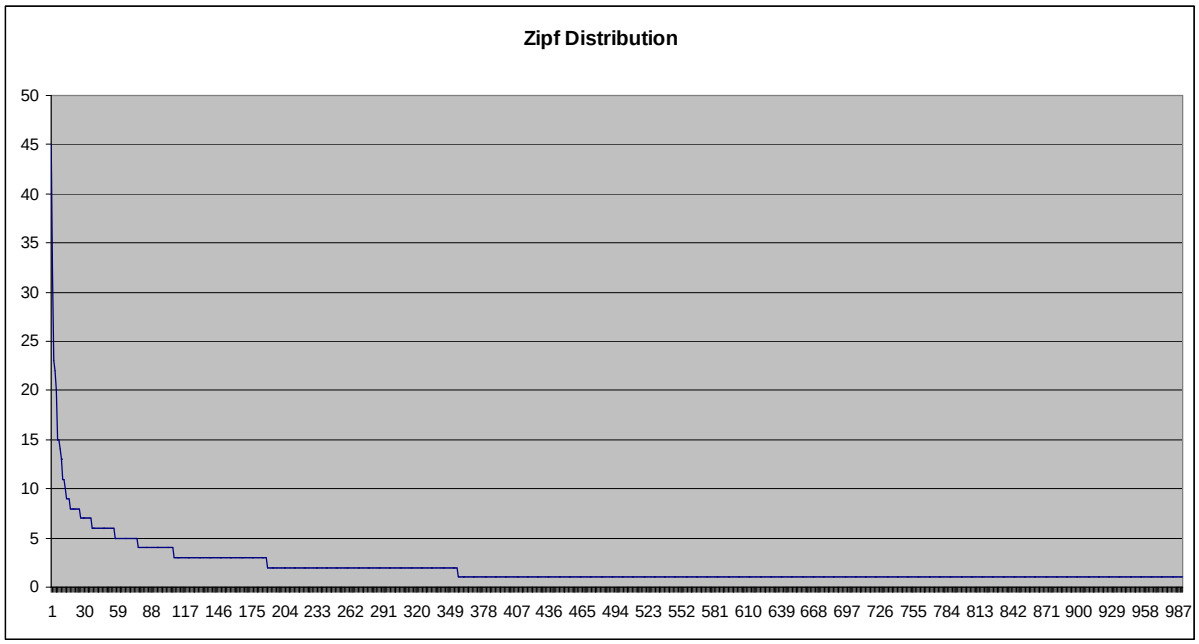
Term analysis

Another way of studying the vocabulary of the users is by breaking down the queries into terms; this can give light to words that are frequently used in multi term queries but not used on their own. Looking at the following table of the top 20 frequently used terms you find obvious entries like “och” that is a common word used for concatenating sentences but besides these kinds of words you can also see that there are others that can not be found in the top 20 queries e.g. “forskning”.

Term	Term count
hjärnan	45
i	32
och	23
se	22
kloning	20
hälsa	15
genteknik	15
genetik	14
hudcancer	13
forskning	13
på	11
nya	11
stroke	10
psykologi	9

med	9
biologin	9
självförtroende	8
permafrost	8
idrott	8

The following diagram shows the frequency of all the terms decomposed, it is plotted as a Zipf (Zipf, 1932) distribution and it gives a short head with a long body and tail, the initial steep fall in term frequency differs from what was seen by Jansen et al. (2000), probably because they had a larger data set to analyze. The information you gain from doing this is that you get a visual overview that shows that the most frequent terms make out for a large part of all the terms. It is less obvious here since there were a great amount of terms only used once.



Top-hits / Non-hits

Looking at the amount of hits that queries generate is interesting because it can show where there is an overflow or underage of information. An emphasis can be put on the queries that users often search for or new content can be added to support what is being searched for but not found.

Looking at the following table with frequent queries and the amount of hits they receive you can see that there is a significant amount of hits for several of the queries reaching as high 34879 hits. This is vast collection of information to parse through and taking into consideration that most users do not pass the first two to three result pages there should be actions taken to help the users narrow their search.

Query	Query count	Hits
se hjärnan	20	20023
hjärnan	13	1744
kloning	12	341
genetik	9	677
nya biologin	8	18954
genteknik	6	221
hålla andan	5	34879
art-training	5	7283
östersjön	5	1735
etik	5	324

In the same way you can look at the statistics for reoccurring queries that do not generate any results at all. From that you can analyze if content related to the queries should be found and then add information sources or synonymous that would make relevant results obtainable. In this case there were some queries that gave no hits but the count of queries was very small with a peak of 4 and that can be considered to be too small of an amount to require any actions to be taken.

Query	Query count	Results
Dextegen	4	0
kina	2	0
mikrokiurgi	2	0
avancerad inlärning	2	0
löntagafonderna	1	0
teckenkommunikation	1	0
Norrelgel	1	0
Fritjof Norrelgel	1	0
telomer	1	0
cellselning	1	0

Length of queries

The length of queries differed immensely with a span from the shorter of two characters to the longest with 85 characters but the longest one within the 90th percentile was 24 characters. This gives a good indication of what size the query box on your site should have to meet the users' general queries. Looking at a percentile measurement instead of the average length (13,58) corresponds more accurately with the users' query lengths since using the average number would leave out approximately 40% of the queries.

Referrer

A referrer in this case is the page from where the user performed his initial search query. The reason for monitoring this is to detect lack of information e.g. if a significant amount of users show a pattern of searching for seemingly similar things from a certain referrer you might be able to detect information gaps on that specific referrer page. Due to two circumstances I have deemed this data not to be relevant to show here, these circumstances consist of the data collection of referrers being to small to draw any conclusions from and the fact that www.forskning.se is a general purpose page for research and therefore often only provide broad information on its pages so searching for more information is natural.

Result page analysis

The amount of result pages viewed by users had a huge span, from 1 to 26, with a mean value of 3,50, which is higher then what Stenmark (2005b) reported where 90,8% only viewed one page and the mean value was 1,4 and also higher then Jansen et al. (2001) where the corresponding mean value was 2,35. More then half the users do however not proceed further then the second result page which strengthens the argument for the need to have relevant hits on the first result page. Closer examination of the data shows that navigation beyond results page 17 was performed by only one user.

No. Of result pages examined	Occurrences	Percentage
1	195	35,52%
2	108	19,67%
3	67	12,20%
4	46	8,38%
5	35	6,38%

6	29	5,28%
7	18	3,28%
8	10	1,82%
9	8	1,46%
10	5	0,91%
> 10	28	5,10%

Result page click through analysis

A click through is when the users clicks on one of the results that is presented after a query and by viewing from which result page users are doing click throughs you lay more strength to the claim that the results on the first result page is the most important ones. The data shows that over 80% of the click throughs at www.forskning.se took place from the first result page and more then 90% from the first three result pages.

Result page	No. Of clicks	Percentage
1	677	81,57%
2	60	7,23%
3	26	3,13%
4	22	2,65%
5	17	2,05%
6	7	0,84%
7	8	0,96%
8	3	0,36%
9	2	0,24%
10	1	0,12%
11	2	0,24%
12	1	0,12%
15	3	0,36%
16	1	0,12%

Weight analysis

Weight is a measure of how relevant a certain result is to the query specified by the user and it is calculated by algorithms within the search engine. It is presented as a number between 0-100 and with an average weight of 82,52 on the click throughs it shows that users frequently click highly weighted results which comes as no surprise since most clicks are from the initial result pages and they should be displaying the most relevant results, granted the search engine is implemented correctly and there is information resources that correspond to the query.

Advanced search

Measuring the usage of advanced contra simple search is interesting because it shows how the average user prefers to find the information he is looking for and also informs if the advanced search warrants its maintenance. The results, 1049 simple searches compared to 177 advanced searches, came as no surprise since simple search was the default and it is also the way that users are used to interacting with a search engine much because of how e.g. Google is designed.

Discussing the metrics

Activities as a metric is interesting because you can see when the user load is the highest and you get a good overview on if the search actually is used. If you look at the activities during hour of the day you get a clear indication of when you should be performing maintenance and when you can experience performance problems if the server hardware is not up to par with the peak user load. In addition to this looking at the unique session ids and comparing the amount of them to the amount of activities you can take note if the search engine is used by a large or small amount of users. When

the activities has been analyzed you with term and query analysis which is interesting because you can do further research into which click throughs they often result in, the combination of these to metrics gives you the ability to provide more directed results to the popular queries and for intranets and on site search where there often is a specific business or theme “language” you can gain great knowledge about the vocabulary used. For further analysis you can look at the weight of the results that are click throughed, are the results weighted with a relevance that corresponds to how often they are clicked? When you know what your users are searching for it is interesting to analyze what kind of results they are receiving from their queries, what hits are they seeing. Top-hits is very relevant when you have a large index of information especially when you look at the statistics from the result page analysis, if you provide over 20000 hits for a query and the users by average only look at 30 of those results, then there can be a major problem if those results do not generate any click throughs. At the opposite end, non-hits can together with the referrer show a lack of information in either your index or your site as a whole but it requires some manual labor with analyzing what information is missing, why it is missing or even why users are searching for it when it actually is not supposed to be found on the site. Bundle all of these metrics together with the length of queries and advanced search metrics which provides you with information on how you should design your search, how do the users prefer to search and how do we cater to their preferences.

All in all it is hard to point to a specific metric and say that this is the one metric that is more important then the others, maybe it is more about finding the collection of metrics that together give a good basis for drawing conclusions about user search behavior and how you can improve the search implementation based upon all of them. So instead of looking for specific metrics to be able to make comparisons (Jansen & Pooch 2001; Stenmark 2005b; Spink et al. 2001) the focus should instead be for a set of metrics that are general enough so they can be used for analyzing all types of search implementations on their own merits. For instance, the use of term/query analysis, click throughs and weight analysis provide a good overview of how the users search and interact with the results while the query length does not present an as important/general contribution. Compared to Stenmark (2005a) and Ozmutlu et al. (2003) it was chosen not to present some session and term related metrics e.g. session arrival and terms per query because it seems like there is no real explanation to why this data is relevant to measure, it is just measured. I am thus arguing that TLA studies should include a discussion on which sets of metrics have been of particular use, and why, over the more traditional presentation of data only. Therefore, use of TLA for specific cases may become a contribution to the study, while the study in itself is given a chance to contribute back to TLA as research approach. If you look at previous research (cf. Stenmark 2005a; 2007, Jansen et al. 2000; 2005) it seems like the way of performing TLA is stuck in the more traditional way of performing IR-research, maybe it should instead be taken forward and adjusted to fit a web/intranet/on-site search setting. Instead of doing focused time or term transaction log analysis a step back should be taken and a more general stance towards the metrics should be established. This could lead to a more cohesive transaction log analysis research basis in further work.

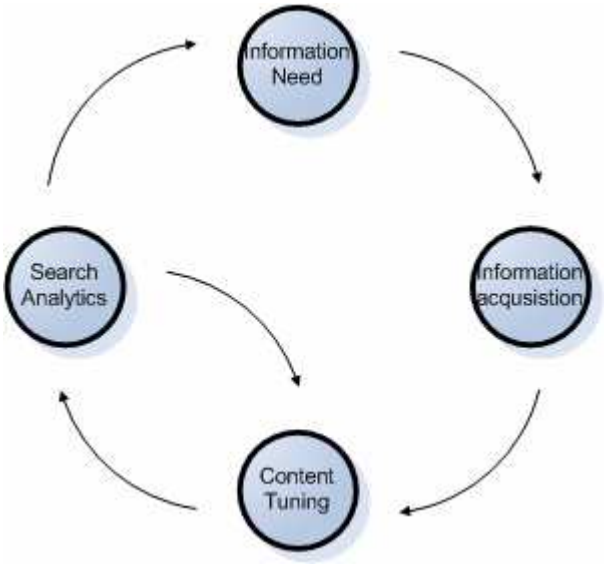
The practical contribution of this study has been to find out how/what to analyze from transaction logs to be able to improve an already existing search implementation and this requires that you use relevant metrics that provide substantial data. Another practical contribution was the customized transaction log which was created using the JavaScript and PHP-script to customize what parameters was logged, this gave great freedom because there were no constraints from the transaction log specification provided by the search engine vendor. This is probably the best way to conduct TLA since it gives you the ability to log relevant data and control the structure of it while also providing better ways to avoid “junk” data in the transaction log that needs to be cleaned in a later state. If this can be combined with a general metric specification it will give great strength to transaction log analysis.

Conclusion

I set out to study search behavior of users conducting searches at the on site search of www.forskning.se with the goal to contrast and relate the findings to previous search related research, finding what types of data that is interesting to take note of and how to look at them when you want to find ways to enhance a live search implementation and also to reflect upon how TLA contributed to the work.

Resulting from the TLA use, we have learned that my results showed both similarities and differences from previous studies, but mainly it showed that previous TLA studies have not had the same practical focus as this one and therefore have maybe not questioned the metrics that are used in the same way.

Contributions can thus be summarized as actions you can take to enhance the search engine implementation in regards to result analysis and following figure which will show how the action, search analytics, can be placed in a larger scale model/lifecycle of search implementation and maintenance, this is an abstract picture, but it does show how search analytics and content tuning interact with each other.



1. Identify information needs
2. Acquire information
3. Tune the acquired information so that you have clean data in your index
4. Perform transaction log analysis (search analytics) on gathered statistical data (transaction log)
5. Depending on the conclusions that can be drawn from the search analytics (4) you either tune the existing content or identify new information needs.

As for TLA, further analysis on what general metrics should be used when performing this sort of research could help to perform more controlled and coherent studies in the future. Also combining it with qualitative research could give more strength to it by finding out the users' intent and emotions while performing search.

References

Abecker, A. Bernardi A and Sintek S. (1999) "Enterprise Information Infrastructures for Active, Context-Sensitive Knowledge Delivery", ECIS'99 - The 7th European Conference on Information Systems

Fagin, R & Kumar, R & McCurley, K, Novak, J & Sivakumar, D & Tomlin, J & Williamson, D (2003), "Searching the workplace web", Proceedings of the Twelfth International World Wide Web Conference, Budapest, 2003.

Glaser, B., & Strauss, A. (1967, June). "The discovery of grounded theory: Strategies for qualitative research". Chicago: Aldine Publishing Co.

Inan, H (2006) "Search Analytics: A Guide to Analyzing and Optimizing Website Search Engines", Booksurge Llc

Jansen, B. & Spink, A. (2003) "An Analysis of Web Documents Retrieved and Viewed". Proceedings of ICIC'03, Las Vegas, NE, 65-69.

Jansen, B.J. & Spink ,A.(2005)."How are we searching the World Wide Web? A comparison of nine search engine transaction logs". Information Processing and Management, 42, 248-263.

Jansen, B. (2006) "Search Log Analysis: What it is, what's been done, how to do it", Library & Information Science Research Volume 28, Issue 3, Autumn 2006, Pages 407-432

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

Jansen, B. J., Spink, A., and Saracevic, T. (2000)." Real life, real users, and real needs: A study and analysis of user queries on the web". Information Processing and Management. 36(2), 207-227.

Jansen, B. J., Spink, A., & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. Journal of the American Society for Information Science and Technology, 56(6), 559-570.

Offsey S. (1997) "Knowledge Management: Linking people to knowledge for bottom line results". Library Hi Tech News incorporating Online and CD Notes, Volume 1, Number 2, 1997 , pp. 113-122(10), Emerald Group Publishing Limited.

Ozmutlu, S., Spink, A. & Ozmutlu, H. (2003) "A day in the life of Web searching: an exploratory study". Information processing and management, 40, 319-345

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1998) "Analysis of a very large altavista query log". Technical Report 1998-014, Digital SRC, 1998.

Spink, A., Wolfram, D., Jansen, B. & Saracevic, T. (2001). Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology, 52, 3, 226-234.

Spink, A., & Jansen, B. J. (2004). Web search: Public searching of the Web. New York: Kluwer.

Stenmark, D. (2005a). "One week with a corporate search engine: A time-based analysis of intranet information seeking". *Proceedings of AMCIS 2005*, Omaha, Nebraska, August 11-14, 2005, pp. 2306-2316.

Stenmark, D. (2005b). "Searching the intranet: Corporate users and their queries". *Proceedings of ASIS&T 2005*, Charlotte, North Carolina October 28-November 2, 2005.

Stenmark, D. & Jadaan, T. (2006). "Intranet Users' Information-Seeking Behaviour: A Longitudinal Study of Search Engine Logs", *Proceedings of ASIS&T 2006*, Austin, Texas, November 3-8, 2006.

Stenmark, D. (2007). "Analysing terms, pairs, triplets and full queries used in intranet searching", in Proceedings of 3rd WEBIST, Barceloina, Spain, March 3-6, 2007.

Zipf, G. K. (1932). Selected studies of the principle of relative frequencies in language. Addison-Wesley.