



Handelshögskolan

VID GÖTEBORGS UNIVERSITET

Institutionen för informatik

2006-01-16

Mining a corporate intranet for user segments of information seeking behavior.

Abstract

Our intranets are growing and the constant unstructured adding of documents and information calls for the need of state of the art search tools. Very little studies have been conducted with the focus on information seeking behavior on intranets. More importantly, only a handful of researchers and vendors see the users as a heterogeneous group. Instead, search engines are designed for one group of people, the user. By applying data mining techniques on a week's log file taken from a search engine of a large corporate intranet an explorative approach was taken to identify user segments in terms of information seeking behavior. Five rather homogenous segments of users were found and described. Some of the commonly used parameters in Transaction Log file Analysis on both Internet and intranet were examined regarding inbound correlation. The characteristics of these segments and the correlation among the parameters can be used as input when designing new and better fitting search tools.

Keywords: Information seeking behavior, segments, users, intranet, mining, Self Organizing Maps, Clustering, log file analysis.

Author: Henrik Strindberg

Supervisor: Dick Stenmark

Master Thesis, 20 Credits

ACKNOWLEDGMENTS

This thesis would not have been possible without the supreme guidance of my supervisor Dick Stenmark., thank you. I also wish to thank my fiancée Anna Andersson and two old friends, Magnus Skog and Jonas Öhlund, for their comments and proofreading.

TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	III
INTRODUCTION.....	1
<i>Objective and Research Questions</i>	<i>1</i>
<i>Research Needs</i>	<i>1</i>
<i>The BISON Project.....</i>	<i>3</i>
<i>The Volvo group & Violin.....</i>	<i>3</i>
<i>Delimitations</i>	<i>3</i>
RELATED RESEARCH.....	4
<i>Information seeking behavior on the Internet</i>	<i>4</i>
<i>Information Seeking Behavior on Intranets.....</i>	<i>5</i>
<i>Parameters of interests.....</i>	<i>5</i>
METHOD.....	9
<i>Literature search</i>	<i>10</i>
<i>Data Collection and Pre-processing</i>	<i>10</i>
<i>Visualization and analysis.....</i>	<i>13</i>
<i>The k-mean Clustering algorithm</i>	<i>13</i>
<i>The Davies-Bouldin Index.....</i>	<i>14</i>
<i>Putting it all together, the analysis</i>	<i>15</i>
RESULTS.....	16
<i>Common Statistical Key Values</i>	<i>16</i>
<i>Correlation matrix.....</i>	<i>16</i>
<i>Parameters as maps</i>	<i>17</i>
<i>Path of finding the clusters.....</i>	<i>21</i>
<i>The Clusters.....</i>	<i>22</i>
DISCUSSION	26
THE PAIR-WISE CORRELATION OF THE PARAMETERS.....	26
THE FIVE SEGMENTS.....	26
<i>The opposites</i>	<i>27</i>
<i>The middles.....</i>	<i>28</i>
DELIMITATIONS & FUTURE WORK.....	29
<i>Methodological Reflections.....</i>	<i>29</i>
<i>Future Work.....</i>	<i>31</i>
CONCLUSIONS	32
<i>Possible to find segments of intranet users be identified.....</i>	<i>32</i>
<i>Possible to describe segments of intranet users</i>	<i>32</i>
<i>Implications for different interest groups</i>	<i>32</i>
BIBLIOGRAPHY.....	34
APPENDIX : ULTRASEEK PARAMETERS.....	38

Chapter 1

INTRODUCTION

The document collections in our intranets keep growing (Heide, 2002) and these huge collections are goldmines when looking at them in terms of an organizational memory. But, the unstructured adding of documents to these distributed repositories makes them virtually impossible to find without the usage of a search engine for indexing and accessing them. Search engines have previously been reserved and designed for highly skilled and trained information retrieving professionals, a rather homogeneous group, but are now used in an everyday context by ordinary people (Spink et al., 2001; Jansen & Spink, 2003). These new users have different background in training and have different and more personalized information needs and thus cannot be seen as a homogeneous group – still, vendors design and most of the research conducted on search engines are done as if they were used by a homogeneous group. This is a problem and shall be addressed in this thesis.

Marketing people have solved a similar problem when identifying potential customers within a heterogeneous market. They divide the market into homogeneous segments of buyers with similar needs and wants, making the segments heterogeneous among themselves, but homogeneous within (Kotler et al, 1994). Thus, they make it possible to diversify product design, marketing strategies and other efforts to best suit each segment for maximizing the sale of a product or a service. With the same reasoning I suggest a similar approach to segmenting the users of intranets by looking at characteristics of their information seeking behavior and, instead of optimizing the sales of a product, maximizing the fitting of search tools.

Objective and Research Questions

The objective for this study is to examine patterns in the information seeking behavior of intranet users. The research questions are: 1) Can segments of intranet users be identified? 2) Can these segments be described? And 3) what are the implications of these findings? To answer these questions, several parameters will be examined simultaneously with the aid of clustering techniques and graphical maps of information seeking behavior will be generated and studied.

Research Needs

Earlier research in the field of information seeking behavior has been conducted on Web search engines (Jansen et al., 1998; 2000; 2001; 2003 Göker et al. 2001) but these researchers have not paid any attention to what goes on within the millions of intranets and ordinary business people's

everyday information need and behavior. The constant growth of the information stored in our intranets (Hawking, 2004; Heide, 2002) combined with the poor quality of today's enterprise search tools results in costs of lost productivity and loss of business opportunities (Hawking, 2004; Feldman, et al. 2001). This points to that the area of information seeking behavior research on intranets has received little attention by researchers. Except from the few studies (Stenmark, 2002; 2004; 2005a; 2005b) conducted at a Swedish industry corporation virtually no previous research has been carried out in this area. As stated in the referred studies this calls for more research.

With the exception of a few researchers (Shriver et al. 2002; Huang et al. 2004; Stenmark 2005b) much of today's information seeking research seems to assume that users are a homogeneous group. This approach has obvious limitations. By investigating intranet users, as segments will specifically give us a more enhanced understanding of the user's information seeking behavior. Therefore, instead of investigating one parameter at the time, this study takes several different aspects of search behavior in consideration simultaneously. This approach allows me to cluster behavior and visualize graphical maps of search behavior that can be used to identify similarities and differences among segments of intranet users. By identifying these segments and showing the characteristics of them can provide valuable knowledge for future design of search tools. This can result in improved system performance and an enhanced quality of delivery of search tools.

A way of segmenting customers is to use statistical and clustering methods. For instance the basket analysis, taken from reselling field of business administration, later of course accepted and utilized by the scholars within the e-commerce field (Ghani, 2002), can be described as to determine correlations between different products placed in the same shopping basket. Meaning an effort to segmenting customers by examining what they are buying and to study which products are bought at the same time. Furthermore, the correlation among products in baskets combined with demographic data for specific customers are used to determine and even predict buying patterns is now supplied by e-commerce vendors (Microsoft, IBM, NetGenesis etc). A recent study (Desmet, 2001) take aid of the Self Organizing Map (SOM) algorithm, which we will discuss later, to do a rough clustering of buying patterns to identify customer segments in an online bookstore. The same method shall be used in this thesis.

In a broader and from a more organizational point of view a good usage of these intranets will add valuable knowledge capital to organizations using it thus increasing the competitive advantages of them. Therefore it is of strategic importance to provide these organizations with state of the art search tools so they fully can make use of their hidden knowledge resources. To succeed in the mission creating these search tools we need to get both a wider and a more detailed picture of the user seeks in these intranets. In order to study this phenomenon I needed access to a company with a large-scale intranet and came in contact with the leader of the BISON project, which previously worked with

the Volvo Group and had a well-established cooperation with the company. Therefore I will now introduce the context of this study within the BISON project and a brief description of the Volvo Group and their intranet.

The BISON Project

This study is a part of an ongoing research project, BISON, which is a sub-project of a larger three-year research programme run by the Department of Informatics at Gothenburg university and the Viktoria institute, funded by FAS; the Swedish Council for Working Life and Social Research. In BISON we focus on information seeking behavior amongst ordinary "business" people. The outcome (Stenmark 2005c) of the project suggests that information retrieval tools for intranet may need to be designed differently.

The Volvo group & Violin

Volvo was founded in 1927 and has today approximately 81,000 employees, production in 25 countries and operates on the global market covering more than 185 countries. According to Volvo Group and as presented on their webpage¹ they are one of the world's leading manufacturers of trucks, buses, construction equipment, drive systems for marine and industrial applications, aerospace components and services. The Volvo Group also provides complete solutions for financing and service.

The Volvo Group's intranet, *Volvo Information Online (Violin)* was first created with the purpose of supplying top managers with corporate news – but also to gain acceptance of this new way of distributing information. During the years the Violin expanded more and more, and today around 50,000 of the employees have access to it. What was once a nice news feature has evolved to become a core information channel for supplying the employees with everyday information. As many other evolving networks the constant adding of documents led to that the IT-department (now known as Volvo IT) installed the Ultraseek search engine in 1998.

Delimitations

The approach and techniques used in this study are untested in the context of intranet and information seeking behavior. Therefore I choose to delimit the scope of this study to only examine the possibility to find segments of intranet seekers and describe them, based on a few parameters to make a solid foundation for future work. I do not examine any technical, contextual or content-specific aspects but look only at the user's information seeking behavior in the intranet.

¹ <http://www.volvogroup.com>

Chapter 2

RELATED RESEARCH

This chapter is divided into three parts; first I discuss the previous work done regarding information seeking on public search engines available towards a vast majority of Internet users. Secondly I examine previous work done on different intranets to build a foundation to relate my findings and choice of method to. Finally, I present different parameters, which are used in information seeking behavior studies.

As described by Stenmark (2005b) the field of information retrieval (IR), mainly studied by librarians and information science scholars, has changed due to the major adoption of the Web. The Web opened up the IR field to millions of users who had little or no knowledge of traditional search tools (Jansen & Spink, 2003). These users were not *retrieving* information - they were *seeking* it. Information seeking is more human-oriented and the user is unaware if his or hers information need can be fulfilled (Stenmark, 2005b).

Information seeking behavior on the Internet

On the Internet, studies have been conducted for aiding in planning the amount of hardware and bandwidth to support caching facilities with the goal of lower the need of these resources (Lempel & Moran, 2003) They studied different parameters in the context of how and how often users interact with search engines. Their research has provided knowledge about the characteristics of overall load of usage and how it is distributed trough different intervals (hours, days, and weeks). Beitzel, et al. (2004) has mapped what the users seek for, this to give an overall picture of what is searched for – this has also been utilized in the context of providing cache facilities for building better retrieval and search algorithms.

This thesis continues the work regarding the aspects of how different users interact with the search engine in terms on what kind of behavior they show (Jansen et al., 2000; 2002; 2004; Jensen & Spink, 2003; Göker et al 2001). In more detail Jansen, et al. (2000) studied the Excite search engine by examining *how the users search the Web* and *What do they search for*. They did this by examining different parameters such as queries submitted to the search engine and how the users view result pages. This thesis continues their work but within the context of an intranet and by looking at several parameters simultaneously.

Information Seeking Behavior on Intranets

Until early pioneers such as Hawking et al (2000), Göker and He (2000) started the study of information seeking in context of intranets, the area has been pretty much untouched. Hawking et al (2000) migrated a text search engine previously studied in a laboratory setting with the goal to test it in the real world with real users. They adopted the Transaction Log file Analysis (TLA) methodology introduced by Jansen et al (1998) and Silverstein et al (1998) as their method. However they had no intention in further understanding the process of information seeking or trying to find any segments of users. Huang et al. (2004) and Shriver et al. (2002) take the session definition, explained below, further and investigate possible segments of sessions; they however make no attempt in finding segments of users except by looking at the session length. Fagin et al (2003) showed that there are differences in how users search the public web and intranets, but they did not try to understand the process information seeking behavior. Stenmark (2004; 2005a; 2005b) has taken the inputs of Fagin et al (2003) and performed series of studies to test the findings of (Jansen et al 1998; 2001; 2003; 2004) in order to determine if the knowledge acquisitions from Internet can be applied in the context of intranets. His findings speak in two ways; some parameters are more or less equal both on intranets and on the Internet but some of his results point towards great differences between information seeking behavior on intranets and on the Internet (Stenmark 2005c). This supports the initial statement of Fagin et al (2003).

Overall, there has been no research taking an effort of looking at intranet users in segments, except that Stenmark (2005b) suggests an existence of "super users" and Huang et al. (2004) or Shriver et al. (2002) session identification methods.

Parameters of interests

To track down any segments of intranet users I had to take several parameters into consideration. These parameters are well used in different TLA-based studies conducted on both the Internet and on intranets. The parameters are equal in naming, but different researchers sometimes define them differently. Therefore a discussion and explanation of the parameters will follow.

Term, is defined by Spink et al. (2000) as: "... any unbroken string of characters (i.e. a series of characters with no space between any of the characters)" In the cited study Spink et al. counts logical operators² as a term but suggests that in their further research they will interpret them as "commands". I follow their example to the extent of not counting any logical operators.

² Logical parameters are + or – and are used to supply the search engine with information whether a term must (+) or must not (-) exist in the results.

Query, is defined by Spink et al. (2000) as "...consists of one or more search terms, and possible includes logical operators and modifiers...", and in this study I chose not to count the logical operators, making the *Query Length* simply the number of terms found within each query. When calculating the average query lengths I have chosen not consider zero length queries as a query– i.e., queries where the user has submitted nothing. Spink et al. (2000) report that users, in mean, construct their queries by 2.21 terms. Their results and what is presented in this thesis cannot be compared in detail due to the above stated reasons, but could give a hint of validness. Spink et al. (2000) also studied the modification of the queries, i.e. the adding or removing of terms to the query. This study differs since no modification is studied, but would have improved the significance of the findings.

Session, the most simple definition is that all queries sent to the search engine by a user make a session (Spink & Jensen, 2000) these authors later change their definition by adding the concept of interaction: "*A session is the entire series of queries submitted by a user during one interaction with the web search engine.*" (Spink & Jensen, 2003) They do not inform us how they tell if the user has left the search engine, but I can assume they have used a cookie³ which times out when the user closes his or her web browser. Thus making a lot of bias in their results since a user might have the same information need but accessing the site while closing the browser window in between. Spink & Jensen (2003) cannot address any change in information need if the user decides to leave his or her browser window open for several weeks. Still, it is a much better session identification method then the first one. As pointed out by Stenmark (2005b) this session border identification is not optimal since these kinds of sessions can span over several days – and it's fair to assume that the information need has changed.

A solution to this issue has been suggested by He and Göker (2002). They argue that a session is a group of activities performed by a user with a specific information need. A new session begins when the topic of this need changes. They present a method to determine session boundaries and argue that an idle time between 11 and 15 minutes between any actions from a user should indicate such a boundary. In the context of intranets Stenmark (2005b) suggest, an idle time 13 minutes idle time for breaking up the sessions which, also is used in this study. This study takes usage of He and Göker's method but with Stenmark's (2005b) more precise idle time - but, by using this, still yet basic, method of identifying sessions I contaminate the results with errors in the way of handling the users as a homogenous group. A more accurate methodology suggested by either Huang et al. (2004) or Shriver et al. (2002)

³ A cookie is a small pice of data stored locally at the client side, containing user specific information accessible by the server. The cookie can expire in two ways: 1) Timing out. 2) User closes his or her web browser and 3) Never.

should have been used for a better result, but this lies outside the scope of this study.

Viewed result pages is the amount of result pages a user views. After a user has submitted his or her query to the interface, the search engine answers and presents a several results on a result page, usually in groups of ten. On the result pages he or she can usually choose between two types of actions. Either a user can view a hit, or request a resource similar to a presented result. In their study of Excite search engine, Jensen et al (2000) regarded all identical queries submitted to the search engine by a user as a view of result pages, yet they refer to Peters (1993) who states that users quite often retype their queries, which contaminate their findings with errors. They reported that a user in mean view 2.35 number of result pages (including the initial one). In their later study (Jensen & Spink, 2003) they follow the same approach still using identical queries to identify a change of result page creating the same bias. They also tell us that their data get polluted when a user click on a presented result within a result page, view that page, and return to the interface. In this procedure the search engine logs this new entrance with the same identification and the same query, i.e. making this a view of a result page.

All in all, their method and accuracy differ to what is used in this thesis since the Ultraseek engine logs used in this study explicitly log change in result page resulting in a higher precision. In this thesis the initial result page is not counted since it is generated by default and not explicitly asked by the user. Since the parameter *mean number of activities* is calculated by adding all the activity type parameters together – counting the initial view result page would have resulted in polluted data since the submitting of a query and the view of the first result page only requires one action from the user. This would have resulted in a major increase of the parameter *mean number activities*. Anyway, since all viewing of the result pages beyond the initial one *require* the viewing of the first one, adding the number one (1) to the findings in this study will make the findings regarding this parameter comparable with the other stated studies.

Relevance feedback has been studied by Jensen et al. (2000) but they were only able to show results on the maximum number of possible accesses to the relevance feedback function. They were limited because the Excite engine logs requests for relevance feedback as empty queries, which of course could have been generated by users only clicking on the search button without supplying any query. This method seems to hold quite a lot of bias especially since Stenmark (2005c) reported that approximately 5% of all queries are empty ones. His report differs somewhat to the 1.9% that Spink et al (2001) reported. Intranet seeking and Internet seeking differs (Fagin et al. 2003), yet it seems that using empty queries to identify relevance feedback is very

uncertain, and will pollute the results with 1.9% to 5% of the body of queries. This study measure the usage of the relevance feedback much better since the Ultraseek engine in this case also logs any access to the relevance feedback function explicitly, so any problems regarding bias from empty queries or other pollution is nearly non-existent.

Viewed hits, – In the excite study Jensen et al. (2000) report no findings or methodology of identifying the viewing of hits on the result pages, but in another of their studies (Jensen et al., 2003) conducted on the FAST engine they report an approach of capturing the URL of the web page the user clicked on in the result page. They were therefore able to draw conclusions on the time spent on each retrieved document (if the users return to the search engine) and the amount of viewed hits. In this thesis I take a similar approach to study the amount of viewed hits and the time spent on each of the hits, but instead of tracking the web page in question the Ultraseek engine logs all click troughs explicitly.

Activities is the total amount of all the above-mentioned interactions with the search engine including the user's first view of the interface. No information has been found on any studies reported to have taken the interface viewing into consideration. By adding this parameter of study a more accurate session length can be measured. But, since this entry is measured a higher number of activities will be reported.

Chapter 3

METHOD

Since the field of information seeking behavior, especially on intranets, is new and pretty much untouched research field, which lacks solid theories, an explorative approach was chosen. In all studies a decision whether to take a quantitative, qualitative or combined approach must be taken, which is also the case in this study. Qualitative methods (Denzin & Lincoln, 2000) like interviews (Fontana & Frey, 2000), observations (Adler & Adler, 1998), focus groups (Greenbaum, 1993) or even full ethnographical studies (Chambers, 2000) are time consuming and best suited for getting a deeper and holistic understanding why human beings act the way they do (Firestone, 1993), which as stated lies outside this study. In all these tools and methods the researcher itself is also a source of bias, either by directly influencing the statement of the research object or indirectly by disturbing the objects natural environment. And more important, as pointed out by Hawking et al (2000) any naturalistic approaches to study these phenomena would be pointless due to the sporadic nature of information seeking. Also, looking at geographically distribution (185 countries) of the approximately 50,000 users with the possibility to gain access to the intranet search engine would have made it impossible to generate any significant results.

For the above stated reasons any naturalistic approaches were dismissed and therefore different tools used in quantitative research methods were examined. When confining in making this a quantitative study I chose between two main approaches: 1) *Online survey* and 2) *Transaction Log file Analysis*. First, online surveys only cover a part of the population, simply those who take their time filling out the form. And even if they have provided research results in a study within the context of web search engines (Spink et al., 1999), they had a relatively low response rate and thus making their findings not representative for the overall search experience according to Hawking et al (2000). Secondly, surveys cannot address the core question of how the users *really act*; only what they believe they do or want us to believe.

Instead, by adopting the Transaction Log File Analysis (TLA) introduced by Jansen et al (1998) and Silverstein et al (1998) and as pointed out by Hawking et al (2000) will allow me to analyze the whole population of searchers and their behavior, instead of being forced to sampling. The downside of this method is that it gives no information about the context in which the search is performed, the user's purpose, why the search has been initiated; nor does it tell us whether or not the users find what they are seeking for (Hawking

2000) and as pointed out by Stenmark (2005b) optimally a TLA study should be triangulated with qualitative studies.

The method, as follow, can be divided into four main phases: data collection and pre-processing, data- aggregation, visualization and analyzing. I will now proceed and discuss each one the phases starting with a literature study, which initially was done.

Literature search

An extensive literature study was executed with the goal to see if there was any related work done – especially to see if the aid of clustering algorithms and data visualization was utilized in this field of study. The literature study began with first reviewing articles in the field of information seeking to build up a body of conceptual understanding of the area of research. The second stage was to get a wider contextual view of the research area and Google's scholar service was used to get an overview of the academic papers available online. Query terms such as *intranet*, *information seeking behavior*, *self organizing maps*, *clustering*, *users*, and *segmenting* were used. The third stage consisted of accessing the ACM's digital library and browsing trough journals covered by the service. The same query terms were used there, and high-ranking articles were studied for relevance. The choice of relevance was measured by the following criteria: First the article in question should be in the information retrieval or information seeking behavior field. Secondly, any attempts in segmentation of users should be for filled. And finally, if none of the two above criteria was fulfilled, studies in other related fields of research containing clustering of behaviors were examined.

Data Collection and Pre-processing

The raw data was collected between October 14th and 21st, 2004 by the BISON project, at the Volvo Information Technology Corporation, an IT-consultant company in the Volvo Group. The raw data was extracted from their Ultraseek search engine as a transaction log in the combined log format⁴. The log holds entries showing the usage of search engine and carry information such as IP-address, time stamp of access, agent used as well as what kind of request that was made. The request part of the log entry consists of a different number of Ultraseek parameters in-depth explained in appendix 1. The log file consisted of total 61679 activities.

⁴The NCSA Combined log format is an extension of the NCSA Common log format. A more in-depth information can be found at:http://publib.boulder.ibm.com/tividd/td/ITWSA/ITWSA_info45/en_US/HTML/guide/c-logs.html#combined

Parsing the Log Files

The log file was run through a Java application where the previously described parameters were extracted and each log entry was grouped by IP-address – thus, making it possible to track a single user's activities through his or her entire interaction with the search engine. These activities build up a user's behavior since they are all assigned to a specific IP-address. The IP-addresses were sorted and ranked by the number of activities and added to a list with IP-addresses having the most activities at the end of the list and those with least activities at the beginning.

During this process I noticed that some activities were logged twice, resulting in two identical log entries. I have no knowledge from where this contamination has its origin. This, however, was not a major issue since I simply removed one of the multiple log entries to yield a cleaner set of data.

At this stage in the process 8011 IP-addresses were identified as candidates for being users. Another issue that I had to address was the existence of proxies and machine made entries in the log file. To solve this issue I manually examined the 150 IP-addresses containing most activities and removed those candidates that fulfilled one of the following criteria:

- 1) Users that made two entries in the exact same second - but with different queries.
- 2) Users having entries with different queries with different subjects tightly and repeatedly switching between these subjects – i.e., indicating more than one user.
- 3) Users where the entries had a rapid switch of casing of the query -i.e., one query is typed in uppercase and another in lowercase.
- 4) Users with massive amounts of activities consisting of only accessing the search engine's interface doing nothing more.
- 5) Users having a change of user agent or operating system – meaning, it is not very likely that a user have two web browsers or operating systems installed at the same computer and switch between them⁵.

After the examination of the data a total of 109 IP-addresses were removed for the above-mentioned reasons, which left me with a cleaner set of 7902 IP-addresses. From now on these IP-addresses were seen as human users.

⁵ This could happen by people using dual boot systems such as Linux and Windows, but it's not very likely that this happens in this specific company' context.

Self Organizing Maps & MatLab

Each user's parameters mean values were calculated and extracted from the java application and ready as input to a Self Organizing Map (SOM) as an input array of vectors.

The concept of SOMs can be described as a neural network with an unsupervised learning and was first introduced by Kohonen (1995). Unsupervised learning, is a method of machine learning, where the model is fit to the observations (Sarle, 1994), in opposite of the supervised learning, where the data is fit and ordered by a model. Simply put, the algorithm is fed an array of vectors. These vectors are ordered in a map where vectors are visualized as dots those who are most similar (measured by the Euclidean distance) to each other are placed close together on the map. The specific algorithm being used in this study is in-depth presented by Vesanto et al. (1999) and are also similar to methods used in the Information Science field of study for ordering documents (Baeza-Yates et al., 1999).

As Vesanto with colleagues (1999) and Desmet (2001) we use the MatLab software package in this study. The software can be described as a numerical computing environment with own programming language. The software created by The MathWorks, MatLab, provides easy matrix manipulation, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs in other languages. MatLab is specialized in numerical computing, but there are several toolboxes that provides a numerous of calculation and visualization possibilities. More than one million people, in industry and academia use it. The software is also compatible with all major operating systems. For a more complete understanding of the program I recommend a visit to the MathWorks web page⁶.

To be able to compare and add these parameters to a map I needed to normalize them. The technique used here is simply making all vector elements to appear in the interval [0,1]. For example, a vector containing the values (1,5,5,10) and to get each element in the interval [0,1] I need to divide each element with the maximum element value, 10, now getting a vector (0.1, 0.5, 0.5, 1). After normalization the data structure was ready to feed the SOM, which then was trained to order the representation of the users in the aspect of similarity in a 5x5 map-matrix. The choice to make it a 5x5 map-matrix was made because it provides both human readable and easy understandably maps. The size of the map made it also possible to hunt for a maximum of 25 clusters and since this is a rough clustering there would be no point in searching for 100 or 1000 clusters.

⁶ [Http://www.mathworks.com](http://www.mathworks.com)

Visualization and analysis

Since a lot of this study is based on visualization of high dimensional data we now move on to present the visualization technique I used to present the data. The main concept of the visualization is that vectors that are more equal to one another are moved closer together on the resulting cluster map. But first an introduction how to interpret the map representation of the ingoing parameters is presented, later we use these maps to examine the different populations of users in each cluster.

The map representation of parameters

Looking at Illustration 1, each hexagon cell is built up by a population of users in aspects of one parameter. Each cell's coloring represents the mean value of the users' mean values extracted from the log file.

Each cell's position in Illustration 1 corresponds against exactly the same cell in Illustration 2. It is the same population of users but viewed in the aspect of another parameter. For example, the two cells marked $p1$ and $p2$ represent the same population of users in two different dimensions (A and B).

By examining the coloring of the two cells in question we notice that $p1$ is white, $p2$ is gray and $p3$ is black. The white coloring stands for the highest mean value of the populations' mean value on the parameter in question, medium gray stands for the medium value and black for the lowest value – in between the different populations are distributed regarding to the coloring. This gives me the possibility to compare each dimension (parameter) to the others, either visually or by using methods of clustering to identify characteristics of the user populations. For example, a conclusion of studying the two maps could be that the extreme part of the entire users (the whole map) populating cell $p1$ in Illustration 1 (showing the highest values) does not show extreme values in the dimension presented in Illustration 2.

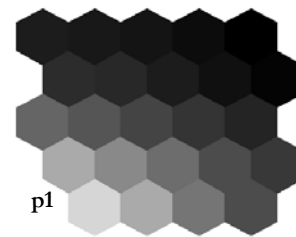


Illustration 1: Example Map A

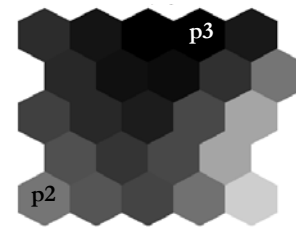


Illustration 2: Example Map B

The k-mean Clustering algorithm

To be able to find population segments one has to find borders, which can divide the entire population of users. This was handled by using the k-mean clustering algorithm. The clustering method used in this study is a method to order objects based on their attributes into k partitions. In this study the different objects are the users and the attributes are the mean values of the

previously stated parameters. The k-mean clustering algorithm is a variant of the expectation-maximization algorithm in which the goal is to determine the k means of data generated from Gaussian distributions. The k-mean clustering algorithm takes the object attributes from the input vector space and tries to minimize the total intra-cluster variance, or, the function

$$V = \sum_{i=1}^k \sum_{j \in S_i} |x_j - \mu_i|^2$$

where there are k clusters S_j , $i = 1, 2, \dots, k$ and μ_j is the centroid or mean point of all the points. $x_j \in S_i$

The algorithm starts by partitioning the input points into k initial sets, in this study set by me after examining the movement of the Davies-Bouldin Index which will be discussed later. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm is repeated by alter clusters (or alternatively centroids are no longer changed).

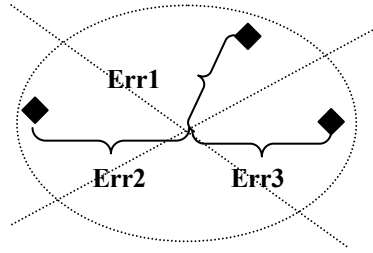


Illustration 3: Intra Cluster variance

Illustration 3 shows a very simple description of the errors that the algorithm is trying to minimize. The square sum of all the diamond dotted observations' Euclidian distance from the cluster centre makes up V.

Since the k-mean clustering algorithm need to be told how many clusters to generate the k-value needed to be chosen carefully – otherwise the clusters would simply be groups of users without any distinct separation. Therefore I needed to measure the quality of the clustering to find a good value of k and still be able to present human readable results. This issue was solved by taking usage of the Davies-Bouldin index, which is presented below.

The Davies-Bouldin Index

The index Davies-Bouldin Index (Davies and Bouldin, 1979) is the function

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{i \neq j} \left\{ \frac{S_n(Q_i) + S_n(Q_j)}{s(Q_i, Q_j)} \right\}$$

of the ratio of the sum of within-cluster scatter to between-cluster separation. Where n , number of clusters, S_n average distance of all objects from the cluster to their cluster centre, $s(Q_i, Q_j)$ - distance between clusters centres. The more compact clusters and the further away from each other they are will result in a smaller index – i.e. the index will have a small value for a good

clustering. This index is used in this study to check the quality of the clustering performed by the k-mean clustering algorithm and the results is shown in the next chapter.

Putting it all together - The analysis

By making maps of all the ingoing parameters, taking the aid of the k-mean clustering algorithm together with analysing the movement of the Davies-Bouldin Index, the different clusters were analyzed in all the aspects of the ingoing parameters.

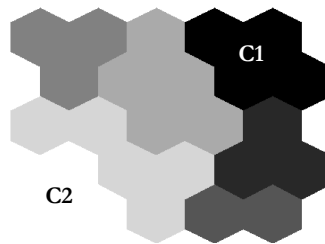


Illustration 4: Example Clusters

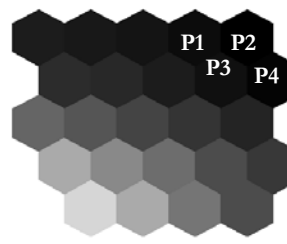


Illustration 5: Example Map Parameter A

For example, the cluster C1 in Illustration 4 in the aspect of parameter A found in Illustration 5 holds the four populations of users, $p1$, $p2$, $p3$ and $p4$ which in the aspects of parameter A show very low mean values. The Cluster C2, which is the opposite of C1, shows on the other hand to hold a population of users with medium to highest mean values on the parameter A.

Table 1: Example statistical values

	Min	Mean	Max	Std
Parameter A	1	5	10	0.5
Map mean	1.5	4	7	-

By applying the statistical and the map mean values of parameter A found in Table 1 the conclusions are that the cluster C1 which consists of the four populations of users having minimum values, here a mean value of 1.5. Cluster C2 on the other hand holds users with values ranging from a mean maximum value 7 to a mean value of 4.

The MatLab Software also provides data on how large each population of users is and the correlation between the different parameters by providing the correlation coefficient for each parameter pair. Note that the parameter maps do *not* provide any information regarding the ingoing sizes of the different populations.

Chapter 4

RESULTS

The results are divided as follows: First the common statistical key values are presented to aid interpreting the graphical map representations of the ingoing parameters. The second section shows the pair-wise correlation between the parameters presented in a matrix. The third section presents the parameters as maps for interpreting the clusters shown in section four.

Common Statistical Key Values

The headers of Table 2 consist of the *Min*, *Mean* and *Max*, which simply is the minimum, mean and maximum value of each parameter. The *Std* stands for the Standard Deviation, which is a more complex key value and can be described as the square root of the sum of the differences of each user's data divided by the number of users minus one. And can be interpreted as the "spread" of the data over the normal distribution.

Table 2: Statistical Key values

		Min	Mean	Max	Std
A	Mean Query Length	1	1.4	10	0.623
B	Mean Relevance feedback Per Session	0	0.00645	3	0.0872
C	Mean Time Examine Hit (s)	0.0769	70.2	779	70.2
D	Mean Time Result Page (s)	1	39	776	54.6
E	Mean Session Length (min)	0	2.2	47.4	3.76
F	Mean Queries Per Session	0	1.45	14.5	1.33
G	Mean Hits Per Session	0	1.12	27	1.44
H	Mean Result Pages Session	0	0.241	22	1.03
I	Mean Activities Session	1	3.16	53	2.97
J	Mean Sessions Per Active Day	1	1.31	10	0.658
K	Active Days	1	1.44	7	0.806

(s) Indicates that the figures are in seconds.

Correlation matrix

Table 3, the correlation matrix, shows each parameters correlation to each other. Some pairs with high correlation are created by the method, for example the *Mean number of Activities per Sessions (I)* is highly correlated with *Mean number of Hits per session (G)* since the number of activities per session is partly built up by the number of hits a user views. In Table 3, parameters with strong correlation are marked with a gray cell.

Table 3: The Correlation Matrix

	B	C	D	E	F	G	H	I	J	K
	Mean Relevance feedback /Session	Mean Time Examine Hit (s)	Mean Time Result Page (s)	Mean Session Length (min)	Mean Queries Per Session	Mean Hits Per Session	Mean Result Pages Session	Mean Activities Session	Sessions Per Active Day	Active Days
(A) Mean Query Length	0.0664	-0.0471	0.0799	0.16	0.156	0.0992	0.101	0.155	0.0296	0.00825
(B) Mean Relevance feedback /Session		-0.0186	0.0167	0.136	0.106	0.116	0.0728	0.154	-0.0062	-0.0195
(C) Mean Time Examine Hit (s)			0.0494	0.124	-0.0731	-0.223	-0.113	-0.185	-0.0561	-0.0671
(D) Mean Time Result Page (s)				0.383	0.0994	-0.032	0.0141	0.0369	0.0218	-0.0064
(E) Mean Session Length (min)					0.595	0.613	0.0442	0.755	0.0941	0.0414
(F) Mean Queries Per Session						0.548	0.34	0.756	-0.0343	-0.0237
(G) Mean Hits Per Session							0.438	0.834	0.0276	0.0293
(H) Mean Result Pages Session								0.701	0.0181	0.00714
(I) Mean Activities Session									0.047	0.0217
(J) Sessions Per Active Day										0.254

The different levels of correlation can be compared to the maps shown in the next section. Strongly correlated parameters are also more likely to have similar graphical representation since parameters with high correlation affects the positioning of each sub population. The difference between the parameter's correlations and the map representations is that the maps are ordered in similarity with *all* parameters in consideration. The correlation figures are on the other hand in aspects of parameter pairs.

Parameters as maps

Below is a listing of all the ingoing variables as maps, black shows low values, white high values and the different grays are values in between spread regarding to the coloring. The map cells with minimum mean values are marked as black and the cells with maximum mean values are marked as white.

Mean Query Length (A)

Queries with the zero terms were disregarded. All values are pretty evenly distributed across the sheet.

	Min	Mean	Max	Std
A Stat	1	1.4	10	0.623
Map means	1.08	1.65	2.22	-



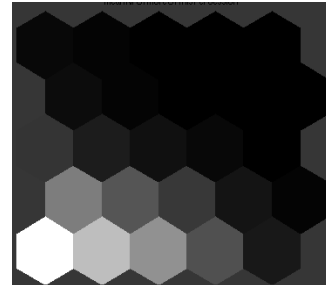
This parameter is correlated with the *mean session length (E)*, *the mean queries per*

session (F) and the mean number of viewed result pages (K). It has weaker correlation with relevance feedback per session (B), mean hits per session (G) and the mean time spent on examining each result page (D).

Mean Relevance Feedback per Session (B)

Only a very few number of users ever used this feature.

	Min	Mean	Max	Std
B	0	0.00645	3	0.0872
Map Means	0	0.02	0.04	-



This parameter is correlated with the mean session length (E) which is created by the method since this parameter is one of the actions that build up a session. It is also correlated with the mean number of examined hits (G).

Mean Time Examined Hit (C)

A population of users spending long time examining hits located in the upper right corner of the map.

	Min	Mean	Max	Std
C	0.0769	70.2	779	70.2
Map means	40	212	384	-

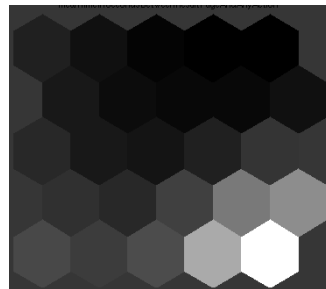


This parameter is correlated with the mean session length (E) and has a very weak correlation with the time spend examining result pages (D). The correlation (C)-(E) is generated by the method.

Mean Time in Seconds on Result page (D)

A concentration of users in the lower right corner spends a lot of time examining result pages. In the top right corner users spends little time on each result page.

	Min	Mean	Max	Std
D	1	39	776	54.6
Map means	22	62	102	-



This parameter is highly correlated with the mean session length (E) which is generated by the method. It has also a weak correlation with the amount of mean queries per session (F)

Mean Session Length Minutes (E)

An evenly distributed sheet with a concentration of users having long session lengths at the bottom left corner. Black indicates users with very short session lengths.

	Min	Mean	Max	Std
E	0	2.2	47.4	3.76
Map means	0.56	4.67	8.78	-

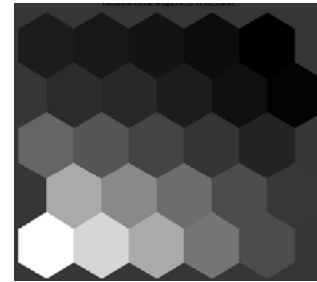


This parameter has very strong correlation with *mean amount of queries per session (F)*, *mean time examining a result page (D)*, *mean number of viewed result pages (H)*. And strong correlations with *mean examined bits per session (G)*, *Mean number of relevance feedback per session (B)*.

Mean New Query per Session (F)

This is an evenly distributed sheet with users submitting many new queries per session at the lower left corner. The users with zero and single query sessions are indicated by black cells.

	Min	Mean	Max	Std
F	0	1.45	14.5	1.33
Map means	0.73	2.2	3.67	-

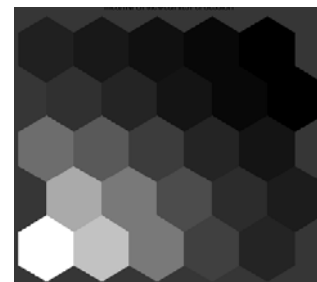


This parameter has a strong correlation with the *mean bits per session (G)* and *mean result pages per session (H)*. It also has a strong correlation with *Mean Session Length (E)*. This parameter also shows a weaker correlation with *mean query length (A)* and *Mean Relevance feedback (B)*.

Mean Viewed Hits per Session (G)

A pretty evenly distributed sheet showing a concentration of users with high number of viewed hits at the bottom left corner – black is users with very few viewed hits.

	Min	Mean	Max	Std
G	0	1.12	27	1.44
Map Means	0.46	2.04	3.61	-

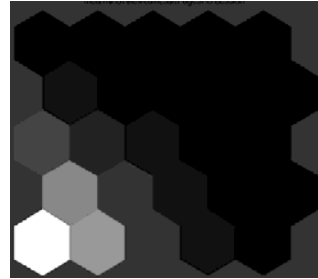


This parameter is highly correlated with *mean number of result pages (H)*. This parameter has also a strongly correlation with *Mean session length (E)*. A weaker correlation with *Mean number of relevance feedback per Session (B)* also exists.

Mean Viewed Result Pages per Session (H)

A concentration of users viewing many result pages at the bottom left corner. Black indicates users with only viewing the initial result page.

	Min	Mean	Max	Std
H	0	0.241	22	1.03
Map means	0.036	0.861	1.69	-

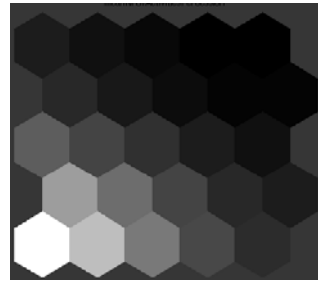


This parameter has no strong correlations with any other parameters except as presented above in (F) and (G). It has a weaker correlation with (A) and (B).

Mean Activities per Session (I)

An evenly distributed sheet with users showing high activity at the bottom left corner and users with little activity at the top right.

	Min	Mean	Max	Std
I	1	3.16	53	2.97
Map means	1.7	5.48	9.29	-

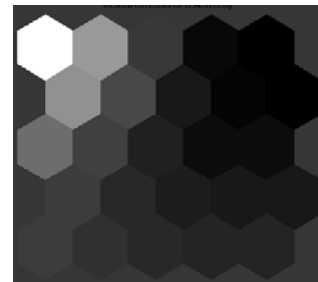


This parameter is built up by the F, G and H and thus creating the correlation between these parameters. A strong correlation between this parameter and the *mean session length (E)* was found. A correlation between this parameter and the *mean query length (A)* and the *mean number of relevance feedback per session (I)* is also shown.

Mean Sessions per Active Day (J)

Users with many sessions per active day at the top left corner and users with little sessions at the top right corner

	Min	Mean	Max	Std
J	1	1.31	10	0.658
Map means	1.05	1.58	2.09	-

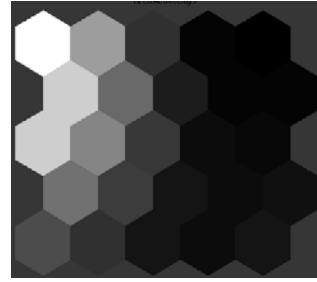


This parameter has none really strong correlation with any of the other parameters except the amount of *active days (K)*.

Nr of Active Days (K)

Users with many active days located at the left right corner and users with few active days at the top right.

	Min	Mean	Max	Std
K	1	1.44	7	0.806
Map means	1.07	1.78	2.49	-



This parameter has no strong correlations except with (J), which is discussed under that paragraph.

Path of finding the clusters

When finding clusters in data with the k-mean clustering algorithm you always have to take a stand between the tradeoff in cluster quality and what is a good visualization and usable for human understanding.

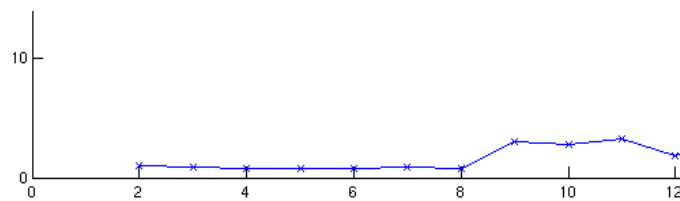


Illustration 6: Davies-Bouldin index with 12 clusters

As stated above, I examined the Davies-Bouldin index for choosing the amount of clusters used in this study. Illustration 6 shows the first run with the K-mean clustering algorithm presenting the index movement up to 12 clusters. The x-axis shows the number of clusters and the Y-axis holds the Davies-Bouldin index. Illustration 6 shows a raise in the Davies-Bouldin index after the eight clustering indicating that there is no idea moving beyond and searching for more then eight clusters. Therefore I reduce the amount of clusters to search for to a maximum of seven clusters. To evaluate the Davies-Bouldin index further a closer look at the index graph was made which is shown in Illustration 7.

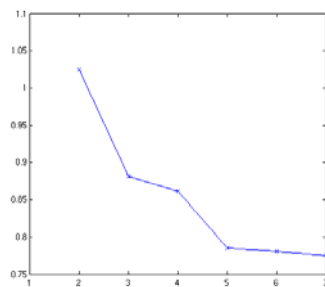


Illustration 7: Davies-Bouldin index

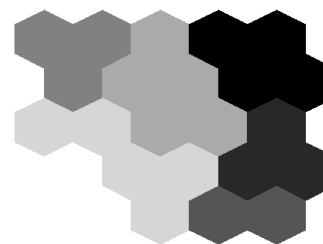


Illustration 8: Seven Clusters

Illustration 8 shows seven clusters, but the movement of Davies-Bouldin index shown in Illustration 7, shows that there is a very little drop between the 7th and 5th clustering. Therefore I decided to stop at five clusters, knowing that a group of six or seven clusters is better but only by fraction. This resulted in an index with a movement presented in Illustration 9.

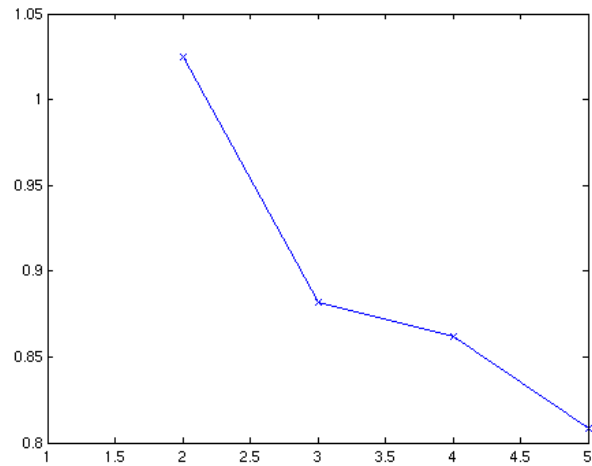


Illustration 9: Five Clusters DB indexes

The Clusters

The final clusters that I from now on will use as is shown in Illustration 10 and their sizes in Illustration 11. By initially looking at the position of the clusters, cluster (2) and (3) are opposites as well as cluster (1) and (5) meaning they are most unlike each other.

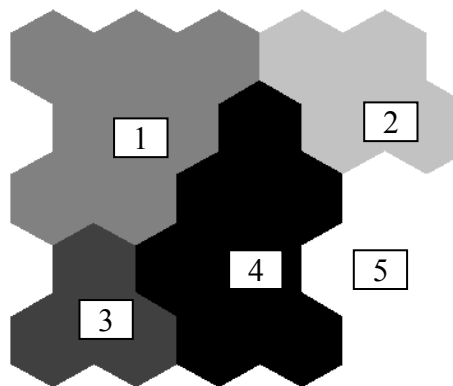


Illustration 10: The Final five Clusters

The total amount of users is 7902 and in cluster (1) there are 2,264 making 29 %, in cluster (2), 2,584 making 33 %, in cluster (3), 725 making 9 %, in cluster (4), 1,249 making 16 % and in cluster (5), 1,077 making 14%

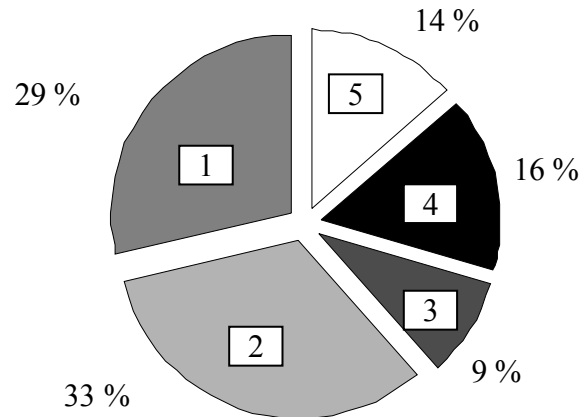


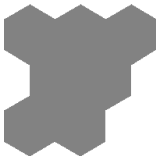
Illustration 11: Cluster Sizes

The five clusters are described below and the characteristics of each cluster are presented.

Cluster 1: 2,264 users, 29 % of the population.

This cluster holds the population of the users that spend little time on each hit, around 40-70 seconds and are in mean active approximately 1.8 days. Almost none used the relevance feedback function but they were the ones having the most number of sessions per day. These users used few terms to construct their queries – however this cluster had a clear division between two groups of users – one who constructed their queries with an average of 1.08 terms and another with 1.4 – both almost equal in size. They showed to be the second most active when viewing result pages.

Their session lengths vary between 0.5 and 5 minutes and they usually submit one to two queries each session. They spent short time on each result page (around 25 seconds) and view some hits (between zero to two). They also showed a mean number of activities ranging from two to five. In size this cluster is the second largest one.



- Spend little time viewing each hit.
- Almost none used relevance feedback
- Active more than one day
- Short queries
- Short session length
- Little time on result page
- View few hits
- Second largest cluster

Cluster 2 – 2,584 users, 33 % of the population

This cluster consists of the users that spent the most time on each hit, in mean, at least 200 seconds, and were only active one day and almost none of them used the relevance feedback feature. They only have one session per active day and almost all of them type single term queries – but there are some users within this cluster that construct their queries with more than one term. They view virtually no result pages except the initial one and have very short session lengths. Their sessions are built up by almost only one query and are the ones that view the least amount of hits with a mean max of 0.5 per session. They are also the ones that spend the least time on each result page⁷



- *Most time on each hit*
- *None used relevance feedback*
- *Single session day*
- *Single term queries*
- *Virtually no result page except first.*
- *Single query session*
- *Only view 0.5 hit per session.*
- *Largest cluster.*

Cluster 3 – 726 users, 9 % of the population

These users are active ones. They spent the second most time of everybody on each hit and are the second most returning users and visit the interface 1.5 times per week. These users are the ones that use the relevance feedback feature the most of all users. They construct their query with medium amount of terms ranging between a 1.6 and 1.8 terms and are also the ones that view the most amounts of result pages almost always going behind the initial page and also viewing the third one. They have the longest session lengths ranging between 5-9 minutes in mean, and are show great activity when viewing hits (2-4). These users spends around 30-40 seconds on each result page and are the ones that has the most activities all in all



- *Second most time on each hit*
- *Second most recurring users, 1,5 per week.*
- *Medium amount of terms in their queries.*
- *View the most result pages.*
- *Longest session, 5-9 mins.*
- *View 2-4 hits per session.*
- *30-40 seconds on each result page.*
- *Had most activities of all.*
- *Smallest cluster*

⁷ If a user submitted a query and did not do anything after that there was impossible to measure how long time these users spent on either viewing a hit or a result page making the duration to zero seconds.

Cluster 4 – 1,249 users, 16 % of the population.

These users spend a short time viewing each hit, between 40-110 seconds and visit the search engine in mean once a week. These users were also the second most active when it comes to using the relevance feedback feature and the viewing of result pages. Most of them construct their queries with more than one term and has the second longest session lengths – but with a small sub-population showing very short session lengths. They are the ones that submit the second most queries per session and also the ones that view the second most hits ranging between 2 to 2.5 viewed hits per session – but with a minority sub population only viewing 0.5 per session. They are also the ones spending the second most time on each result page ranging between 20 to 80 seconds. They're also the number two in the amount of activities per session ranging between two to five.



- *Short time on each hit (40-110 seconds)*
- *Most active using relevance feedback.*
- *More than one term queries.*
- *Second longest session length*
 - *Sub-cluster with very short session length.*
- *Second most queries per session.*
- *View 2-2,5 hits per session.*
 - *Sub-cluster only viewing 0.5*
- *Most time on result page, 20-80 seconds*
- *Second largest amount of activities.*
- *Third cluster in size.*

Cluster 5 – 1,077 users, 14 % of the population.

This cluster holds the population of users that spends the shortest time on each hit and only visits the search engine once. It happens that these users use the relevance feedback feature but very rarely. They almost never have more than one session per day and do not check beyond the first result page. They are the ones that use the most terms constructing their queries with an average of 1.7. They have the third longest session length and usually submit two queries per session. They are the third most active when it comes to viewing hits and are the ones that spend the most time on the (initial) result page they view.



- *Shortest time on each hit*
- *Very rarely using relevance feedback.*
- *Only view initial result page.*
- *Almost always one session days*
- *Most amount of terms used, 1.7*
- *Third longest session length.*
- *Two queries per session.*
- *Third in viewing hits.*
- *Most time on result page*
- *Fourth cluster in size.*

Chapter 5

DISCUSSION

I will now move on to discuss the results of my findings, the implications of the delimitations of the study and the possible generalizations that can be drawn from this study. At the end of this section I will also discuss and provide some reflections about the methodological approach used.

The pair-wise correlation of the parameters

By looking at the correlation matrix shown in Table 3 in the previous chapter the different correlations between the parameters tells us that in overall:

- A user who visits the search engine more days is also more likely to be engaged in more sessions those days than a user who visit fewer days.
- The more activities a user engages in a session, the more likely he or she is to construct their queries with more terms.
- The more terms a users use when constructing the queries the more result pages and hits they view. They also tend to examine the result pages longer, this resulting in longer session length.
- The more queries per session a user submits the longer they seem to spend examining the result pages and also seem to view more of them.
- A user who views many hits is more likely to use the relevance feedback function.

On the other hand the amount of queries a user submits during a session seems to have nothing to do with how long time they spend examining the hits, how many sessions they engage in or the number of days they use the search engine.

The five segments

By moving on in this discussing I choose to rename our clusters to segments because we are now discussing rather homogeneous populations of real human users. The first two segments I will discuss are the opposite ones, Segment 3, *“The Top Seekers”* spawn from cluster 3 and Segment 2; *“The Novices”* spawn from cluster 2. The last three segments are in between these two segments in terms of activity, usage of the search engine and behavior.

The opposites

The two segments discussed below are the opposite ones. The first segment is the ones that take full usage of the search engine; the second take the least usage of it.

Segment 3 – The Top Seekers (9%), spend the second most time of everybody on each hit and are the second most returning users, with a mean of 1.5 visits to the interface the studied week. This segment uses all of the search engines' features and also show the longest session time. They view many result pages and are the ones engaging in the most activities of all segments. It is fair to assume that they know what they are doing, and are skilled with the concept of searching. They are the opposite of the *Novices* (also indicating on the location of the cluster map). This is also the smallest of the segments, representing only 9%.

These users probably does not need further training in seeking, but since they show to be very active in all aspects, one might consider giving them easier access to more advanced search features. Since they are rather small in size it might make sense giving them more sophisticated tools to work with.

In an organizational point of view, these users could teach and encourage the other segments to become better seekers.

Segment 2 – The Novices (33%), are the users that spent the most time on each hit at least 200 seconds, but they were only active one day during this study, and almost none of them used the relevance feedback function. They only have one session per active day and are most common to use single term queries. They view almost no result pages except the initial one and have very short session lengths. Their sessions are almost always built up by only one query and they view the least amount of hits with a mean max of 0.5 per session. They also spend the least time on each result page. Simply put they take least use of the search engine.

An effort to make these users succeed better with their searching could be to provide them the possibility of taking a guided tour of the search engine. This segment is also the largest one, representing a 33% of the population. An effort of educating them in searching would surely make them increase their performance but due to their size any tradition classroom approach would make it very expensive.

The middles

These segments hold users important for organizational implications of intranet seeking. What is common for these three segments are that they all show beginning knowledge and use the search engines functions. They, however, do not reach the level of the *Top Seekers* but they are also the ones, which probably are easiest to push over the edge thus getting the most of any effort in either target education or design of search tools.

Segment 1, The Amateurs (29%), were in mean active approximately 1.8 days showing that these users are not single visit users. It is fair to assume that this is not their first time visiting a search engine since they are familiar with the concept of viewing result pages, navigating through the hits and doing this in a fast manner.

In average, they only spend between 40-70 seconds evaluating their hits and approximately 25 seconds on result pages. Each day they were active they had the most number of sessions per day but the session was very short, between half a minute to five minutes. They examined between zero and two hits and none of them used the relevance feedback function. They constructed their queries with few terms. This segment consist of two sub-segments regarding query construction; one with an average of 1.08 terms and the other with 1.4 – both almost equal in size. I find this segment of users to be similar to behavior presented by Stenmark (2005c) in terms of they believe the answer is “out there” and show the same behavior as web users in terms of lack of patience.

These users should be reminded that they are in fact searching on the intranet, not on the World Wide Web. In according to their size, 29% they are a large group of users, almost a third of the entire studied population.

Segment 4 – The Apprentices (16%) show the same behavior as the *Top Seekers*, but with less density and could learn from them. The Apprentices know the concept of searching on intranets, but have not reach the level of the *Top Seekers* when it comes to activity. This behavior could either be related to work or personal matters and a further investigation is needed to understand this segment more.

Segment 5- The Juniors (14%) are equal in almost all behavior with the *Apprentices* and the *Top Seekers* when it comes to their behavior in the search engine but with even less density then the Apprentices. This segment, however show to hold the population which in mean construct their queries with the most terms.

DELIMITATIONS & FUTURE WORK

The Volvo Group employs approximately 81,000 people of which 50,000 have *access* to the corporate intranet. It is not safe to assume that *all* of these users take usage of the intranet, but it is fair to assume that a majority of them do. The results of this study show that approximately 8,000 visited the search engine during the studied week, but it is impossible to say with what frequency they visited the intranet – a combined log file containing both accesses to the intranet and the search engine would have made this possible.

But still, the 8,000 out of 50,000 possible indicates that roughly 16% of all users with access to the intranet also engaged in some interaction with the search engine during the week of study. However, due to the delimitations previously stated and the choice of method, nothing can be said about their intentions, the information needs that caused their searching to begin or if they succeeded and got the information they needed.

By being a large company with 81,000 employees and active in 180 countries it is possible that the segments identified in this study can be generalized to yield for other organizations. The fitting of the segments can also be expected to be greater within an engineering-heavy organization.

Methodological Reflections

All studies and methods have its bias, including this one. The main bias in this study is the fact that I have worked with mean values when representing the users' behavior.

For example, a user submits a query consisting of very few terms in one interaction with the search engine. Later, in another query, he or she constructs a query with many terms. This leads to that this user will be represented by a mean value between those two searches. This is not optimal since this does not measure how user varies his or her search style.

Yet, this is in this study not a big issue since it is about a rough segmenting, but in a follow-up study a more fine-grained approach for better tracking the variation of behaviors should be used.

When measuring the time a user spent viewing a result page or the time he or she spends examining hits, no contextual parameters has been taken into consideration. For example, I cannot be sure that the time a user spends on those two actions simply consists of examining the results or pages. A user

could simply have left the computer and engaged in some other activity. This however was handled by the session time out heuristics, which removed the extreme bias. And as stated in the delimitations, this study does not take any contextual parameters into consideration.

Data Collection

The data used in this study consists of a one week long log file. This approach has its obvious limitations since it is only a snapshot of that particular week. If the data spanned for several months or perhaps even a year, more reliable conclusions could've been drawn. The only limitations of the methodology used in this thesis, is the amount of available computer resources. The *size* of the log files studied does not affect the time performing the study much.

Data Pre-processing

The pre-processing of the data took usage of human evaluation of IP-addresses to remove proxies and machine made entries in the log file, even though I have been careful when examining the log entries it is more likely than unlikely that some machine made entries have slipped through and contaminated the data. A combined human-machine evaluation and more complex set of criteria would have given a cleaner set.

SOM organization

When deciding the size the SOM I could have experimented further with the different size and shape of the map, instead of relying on pre-defined settings regarding the training of the map and setting the size to 5x5. However, a finer grained map would also have made the results less human readable and more difficult to interpret. Further examination of how to combine the different aspects of the SOM could have provided more precise results.

K-Mean Clustering

The k-mean clustering algorithm does not guarantee to find a global optimum of the best clusters. And a main drawback of the algorithm is that it has to be told the number of clusters to find. These issues were handled by examining the movement of the Davies-Bouldin index to identify the optimal amount of clusters in the current setting. However the Davies-Bouldin index is not the only clustering performance index, but the focus of this study is not to analyze clustering algorithms or their performance. It is to examine segments of users with the *aid* of clustering techniques. Therefore in another study with similar scope, other clustering techniques and indexes should be examined to possibly yield a different result.

Future Work

A first effort of segmenting users of intranet has been done and to fully understand these segments follow-up studies are needed. The delimitations of the research question, the method, the size of the studied data and that only *one* intranet in *one* organization has been studied calls to further investigations of this approach. Below I will present some ideas on what can be seen as next step of this study.

Closer to the raw data

Since this study works with aggregated data, i.e.: mean values, it lacks the precision of in detail track the user behavior. A more fine-grained study using non-linear and using less aggregated data would provide a deeper understanding of the user segments. Therefore I suggest that a user could be represented (as in this study) by different parameters in a time-parameter vector space model. But, instead of working with aggregated mean values, all users' actions are recorded as they are, and then analyzed with clustering techniques in aspects of *similarity between curves* fitted to the different user actions. Such an approach would take each users specific search-style in consideration, thus resulting in an even better segmentation of the users.

Combined Qualitative study

The clusters identified in this study were found by using hardcore statistical tools – a qualitative study could verify these findings. By selecting representative users from each segment, performing either interviews or focus groups would increase the understanding of them.

Other organizations

This study has examined the information seeking behavior in an industrial and production intense corporation, similar studies, as this should be executed in other types of organizations. A study of research intense companies or public sector organizations could examine the full generalization of this study.

Longitudinal study

Since this study only covers a week, a question was raised whether the presented segments are static or change over time. It is almost safe to assume that the clusters are to vary in size over time. Therefore a longitudinal study with the same approach as this could show interesting results. It would also be very interesting to see if users over time show signs of either evolve or degenerate in their seeking. This could be studied by following users across segment borders.

CONCLUSIONS

In the introduction I stated that an identification of user segments would add valuable knowledge in designing new search tools and enriching our understanding of intranet users. Three research questions were to be examined: 1) Can segments of intranet users be identified? 2) Can these segments be described? And 3) what are the implications of these findings? These questions will now be answered as the conclusions and benefits of this study.

Possible to identify segments of intranet users

In this study, I have shown that segments of information seekers with similar behavior can be identified by examining several parameters simultaneously. Five segments were discovered. This has not been done in the context of information seeking and can therefore be seen as new and interesting knowledge.

Possible to describe segments of intranet users

This study has shown that there are differences between segments of information seekers in intranets; they should therefore *not* be treated as a homogenous group. This heterogeneous group can be divided up in rather homogenous segments, which show similar behaviors within the segments, but with differences between the segments. This has not been noticed before and is to be seen as new knowledge.

Implications for different interest groups

Some of the implications of these findings are presented below grouped in three different interest groups: First the research society, second the vendors of search engines and finally, organizations with intranets.

Research society

The finding that users can be seen as segments can aid us in future examining the users. By studying them in segments it is possible to get a clearer view of their behavior. The methodology used in this thesis, Self Organizing Maps, can be used as a tool when examining transaction log files in the same way it has been done in the fields of economics to identify buying patterns. The nature of SOM – i.e. taking discrete parameters in vector representation and the possibility to view the data in terms of maps has shown to be of good usage when it comes to get both an holistic view of the data and at the same

time noticing differences and similarities between segments; Common tools such as excel diagrams can not easily provide such views of data.

Vendors of search engines

Vendors of search engines should bear in mind that when they construct their products that their end users are a heterogeneous group and that each one of the ingoing different segments has shown different behavior. Therefore it is fair to assume they also have different needs when it comes to their interaction with search engines. The result of this study shows that almost a third of the users show little or no knowledge of how to use the search function. Therefore, by adding more educational features with little development costs could improve the quality of service of the vendor's products. By doing this vendors could reduce their customers training costs thus adding extra customer satisfaction and value to their products. On the other hand, the experienced "Top Seekers" take well usage of the current features so additional search functionality can be developed to fit a more complex need

Organizations using intranets

Organizations, which use intranets, and especially the organization of study, should be aware that roughly *one tenth* of their employees using the intranet's search engine show signs of being highly experienced searchers. They should also consider that roughly *one third* of the users show signs of severe lack of skills how to seek successfully on the intranet. This calls for that the organization of study seriously should examine opportunities of educating this segment, preferably with the aid of e-learning facilities.

Approximately half of the population does know how to search, but they still need training and experience to perform better and raise their knowledge to becoming top seekers. The organization of study should really take a look at how to push these users over the edge and turn them into "Top Seekers".

BIBLIOGRAPHY

Adler P. A, Adler P., (1998), Observational Techniques, In Denzin N.K., Kincoln Y.S., (eds), *Handbook of Qualitative Research second edition*, Thousand Oaks, Sage Publication.

Baeza-Yaters, R., et al. (1999) *Modern Information Retrieval*, Addison Wesley, London, p: 46.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O. (2004) Hourly analysis of a very large topically categorized web query log, *Proceedings of the 27th annual international ACM SIGIR Conference on Research and development in information retrieval* pp: 321 – 328, ACM Press.

Chambers, E., 2000, Applied Ethnography, Observational Techniques, In Denzin N.K., Kincoln Y.S., (eds), *Handbook of Qualitative Research second edition*, Thousand Oaks, Sage Publication.

Davies, D.L., Bouldin D.W..(1979) A cluster separation measure. 1979. *IEEE Trans. Pattern Anal. Machine Intell.* 1 (4) pp:224-227.

Denzin N.K., Lincoln Y.S., (eds), (2000) *Handbook of Qualitative Research second edition*, Thousand Oaks, Sage Publication pp 851-869.

Desmet, H., (2001) Buying behavior study with basket analysis: pre-clustering with a Kohonen map, *European Journal of Economic and Social Systems* 15 No2 (2001) pp 17-30.

Fagin, R., Kumar, R., McCurley, K., Novak, J., Sivakumar, D., Tomlin, J. and Williamson, D. (2003) Searching the Corporate Web, *Proceedings of WWW2003*, Budapest, Hungary, 366-375.

Firestone, W.A., (1993), Alternative Arguments for Generalizing From Data as Applied to Qualitative Research, *Educational Researcher*, Vol 22 No.4, pp 16-23.

Fontana A., Frey J. H (2000), The interview: From Structured Questions to Negotiated Text, In Denzin N. K., Lincoln Y. S., (eds), *Handbook of Qualitative Research second edition*, Thousand Oaks, Sage Publication, pp 645-672

Ghani, R., Fano, A., (2002), Building recommender systems using a knowledge base of products semantics. *Proceedings of the Workshop on Recommendation and Personalization in E-Commerce, at the 2nd International on Adaptive Hypermedia and Adaptive Web Based Systems, Malaga, Spain.*

Greenbaum, T. L., (1993), *The handbook for Focus Group Research*, New York Lexington Books.

Göker, A., He, D. (2000) Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning. *Proceedings of Adaptive Hypermedia and Adaptive Web-based Systems*, Trento, Italy, pp:319-322.

Göker, A., He, D. (2001) Web user search pattern analysis for modeling query topic changes, *Presented at User modeling for context-aware applications, a workshop of the 8th International Conference on User Modeling*

Hawking, D., Bailey, P. and Craswell, N. (2000) An intranet reality check for TREC ad hoc, Technical report: CSIRO, Mathematical and Information Sciences.

Hawking, D., (2004) Challenges in enterprise search, *Proceedings of the fifteenth conference on Australasian database - Volume 27*, pp: 15 - 24

Heide, M. (2002) *Intranät – en ny arena för kommunikation och lärande*, Sociologiska institutionen, Lunds universitet, Lund. Avdelningen för medie- och kommunikationsvetenskap.

Huang, X., Fuchun P., An A., Schuurmans, D., (2004) Dynamic Web Log Sessions Identification With Statistical Language Models, *Journal of the American Society for Information Science and Technology*, 55(13):pp:1290-1303.

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

Jansen, B., Spink, A., Bateman, J. and Saracevic, T. (1998) Real life information retrieval: A study of user queries on the web. *ACM SIGIR Forum* , 32, 1, 5-17.

Jansen, B., 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, 207-227.

Jansen, B., Spink, A. (2005) An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management* 41, pp. 361-381

Kotler, P., Armstrong, G., Cunningham, M. H., Warren R. (1996), *Principles of Marketing* 7th edition, Prentice-Hall.

Lempel, R., Moran, S. (2003) Predictive caching and prefetching of query results in search engines, *Proceedings of the 12th international conference on World Wide Web*, pp 19-28, Budapest, Hungary

Peters, T., A., (1993) The history and development of transaction log analysis, *Library High Tech*, 42(11:2), pp 41-66.

Sarle, W.S. (1994), "Neural Networks and Statistical Models" in SAS Institute Inc., *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., pp 1538-1550, <ftp://ftp.sas.com/pub/neural/neural1.ps>.

Shriver, E. Hansen, M., (2002), Search Session Extraction: A User Model of Searching, <http://citeseer.ist.psu.edu/rd/10375466%2C501547%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/25714/httpzSzzSzwwww.bell-labs.comzSzprojectzSzwebsearchzSzuser-model.pdf/shriver02search.pdf> (accessed, 2006-01-03)

Silverstein, C., Henzinger., Marais., H and Moriez, M. (1998) Analysis of a very large AltaVista query log. Technical report 1998-013, Digital Systems Research Center, Palo Alto. <Http://www.research.digital.com/SRC> (accessed, 2006-01-03)

Spink, A and Xu, J., (2000) Selected results from a large study of Web searching: the Excite study, *Information Research*, Vol 6, No. 1, October 2000

Spink, A., Wolfram, D., Jansen, B. J. och Saracevic, T. (2001) *Searching the Web: The Public and Their Queries*, *Journal of the American Society for Information Science and Technology*. Vol. 52, No. 3, pp. 226-234.

Spink, A. (2002.) A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing and Management* 38, pp 401-426

Stenmark, D. (2004), "Intranets and Organisational Culture", *Proceedings of IRIS-27*, Falkenberg, Sweden, August 14-17, 2004.

Stenmark, D. (2005a), "How intranets differ from the web: organisational culture's effect on technology". *Proceedings of ECIS 2005*, Regensburg, Germany, 26-28 May 2005.

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

APPENDIX 1: Ultraseek parameters

<i>Param representation</i>	<i>Exemple entry</i>	<i>Interpreted as</i>	<i>Used as</i>
n	131.97.136.4 - - [14/Oct/2004:03:29:16 +0200] "GET /vhk/cs.html?url=http%3A//violin.nap.volvo.se/&qt=NAP&col=rest&n=1 HTTP/1.1" 302 0 "" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"	Which number of hit the user clicked on.	Identifying which number of hit the user clicks on and determining the time between a user spends on looking on a resultpage.
st	157.171.190.65 - - [14/Oct/2004:03:56:23 +0200] "GET /vhk/query.html?rq=0&col=rest&qp=&qt=the%20volvo%20way&qz=&qc=&pw=565&ws=1&la=en&qm=0&st=1&nh=10&lk=1&rf=0&oq=&rq=0 HTTP/1.1" 200 - "http://violin2.volvo.se/violincgi/wwd_document.cgi?query=the%20volvo%20way" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"	Which resultpage the clicked on. (starting hit)	Identifying which resultpage the user is currently looking at and determining the time between change in resultpages and other actions. Also used as an action aswell as determining session length. <i>An entry of st with the value above 10 indicates that the user is clicked to view a resultpage above the first resultpage.</i> In this manner, the variable is used to calculate the average number of resultpages a user views.
tx1	10.116.0.136 - - [14/Oct/2004:04:56:52 +0200] "GET /query.html?op0=&fl0=&ty0=w&tx0=+goldwing&op1=%2B&fl1=&ty1=w&tx1=+&op2=-&fl2=&ty2=w&tx2=+&dt=an&inthe=604800&ady=8&amo=10&ayr=2004&bdy=15&bmo=10&byr=2004&nh=10&rf=0&lk=1&col=rest&charset=iso-8859-1&qt=&ql=a HTTP/1.1" 200 - "" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; .NET CLR 1.0.3705; .NET CLR 1.1.4322)"	The first advanced query string.	To identifying advanced searches aswell as a number of actions and session length.
qt	157.171.223.95 - - [14/Oct/2004:04:57:40 +0200] "GET /vhk/query.html?rq=0&col=rest&qp=&qt=5023501&qz=&qc=&pw=565&ws=1&la=en&qm=0&st=1&nh=10&lk=1&rf=0&oq=&rq=0 HTTP/1.1" 200 - "http://violin2.volvo.se/violincgi/wwd_document.cgi?query=5023501" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; .NET CLR 1.1.4322)"	The query texty – ie the text the user types as query.	To calculate mean query team length aswell, also gives us the starting time for users who not enter trough the default search interface. Also used as an action to determining session length.

<i>Param representation</i>	<i>Exemple entry</i>	<i>Interpreted as</i>	<i>Used as</i>
fs	10.213.164.105 - - [15/Oct/2004:16:46:18 +0200] "GET /vhk/query.html?pw=565&charset=iso-8859-1&ws=0&fs=http%3A//violin.bus.volvo.se/standards/news/bdaspr%2520list%2520issue%252002%25200306161.pdf HTTP/1.1" 200 - "" "Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; RVIMILE v5.50 SP1)	User use the relevance feedback	Identifying a use of the relevance feedback, and also counts as a activity.
None of the above paramters	157.171.212.119 - - [14/Oct/2004:02:50:25 +0200] "GET /?&ws=1&q=a&nh=10&lk=1&rf=0 HTTP/1.1" 200 8355 "" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; .NET CLR 1.1.4322)"	User entry, where the user enters the site and sees the search engine interface.	Identifying the time when a user starts it's searching session. <i>[note: not all users enter trough this interface.]</i>