NTNU
Innovation and Creativity

Semantic Relations in Yahoo! News Search

Øyvind Arne Evensen

Master of Science in Computer Science
Submission date: June 2007
Supervisor: Jon Atle Gulla, IDI
Co-supervisor: Per Gunnar Auran, Yahoo
Stein L. Tomassen, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

Problem Description

For Yahoo!'s News portal search application, it is important to recognize semantic relations among queries and to understand how people formulate their queries. This may help in the analysis of the users information needs as well as in the retrieval of documents relevant to these needs.

In this project, the candidate will develop a technique for finding semantic relations in raw Yahoo! news query logs. Various strategies for finding semantic relations have to be tested and compared. The log analysis is to be incorporated into the Yahoo! News search application using real news documents. The system will be functionally evaluated against Yahoo!'s existing search application.

Assignment given: 20. January 2007
Supervisor: Jon Atle Gulla, IDI

Abstract

Yahoo News is the world's most popular Internet news site with news search as one of the core pieces of the user experience. Search is by most people a little thought of feature when they are looking for interesting news articles. The common approach of reading news is often by browsing the top news headlines or browsing his or hers favorite news categories. What if search was made intelligent, letting the user browse the search results by subgroups or be given suggestions based on other users interests?

In this thesis we propose a novel approach were 3 days of raw Yahoo! News search query logs are analyzed to find semantic relations among queries. The analysis is based on two independent contributions. The first uses session data extracted from the query logs. By finding the term best describing each session, we get a vocabulary of queries related to that term. Sessions with similar terms are merged to create larger groups of queries with one common term or phrase as group label. The second contribution is the use of temporal correlation to give a measure of frequency variation similarity. Queries that show a similar variation over time have a high chance of either being semantically related or appear in the same situations.

These two contributions are then merged into related term groups, based on their session group label and the most prominent term or phrase of the correlation query. With the use of non-strict parameters for the contribution calculations, a great number of queries are found. With the intersection of the results, this leaves high accuracy groups of related queries with a term or phrase as group label.

A prototype search application was developed to use the created term groups in a search environment. The groups of queries were converted into a tree structure with their group label as the main node. This navigation tree structure let the user navigate up and down in the tree or click directly on a tree node to view its results. When a user's search matches one of the generated groups, he or she is presented with the first search results of the tree's main node together with its children.

The evaluation shows promising results. The tree navigation structure presented to the user was assessed to be of slightly above moderate use. The improved search mode proves to be good at suggesting 'side stories' from the main query. Though number of unique news stories found with both approaches showed similar results. This could either be a news search feature or an implementation problem. It must be stressed that the evaluation is focused on one implementation of many possible uses of the related term groups. It does not directly assess the quality of the term groups.

Preface

This report presents the master thesis of my 5th year in the Computer Science course at NTNU Trondheim. The work was carried out at the Department of Computer and Information Science, Faculty of Information Technology, Mathematics and Electrical Engineering at NTNU, under supervision of Professor Jon Atle Gulla and in collaboration with Yahoo!

I would first like to thank Professor Dr Jon Atle Gulla for his co-operation during the semester. His extensive experience and knowledge in the field of information retrieval has proved invaluable in my work. My co-advisor Stein L. Tomassen has provided me with many helpful suggestions, and has always been available if help or guidance was needed.

I would also like to thank my supervisor from Yahoo! Per Gunnar Auran has helped me form a basis for the assignment and possible progressions. He has also provided me with many helpful insights and tips throughout the semester.

Finally I would like to thank the evaluation test group, for carrying out the evaluation of the search system prototype.

Trondheim June 24, 2007

Øyvind Arne Evensen

Contents

List of Tables

List of Figures

Part I

Introduction

Chapter 1

Introduction

As the Internet grows, more and more advanced search applications have been developed and become one of the main tools of the average user in navigating the World Wide Web. Around the 20th century the Internet was mostly covered by large scale web search engines, which would let the user search in all the information available on the Internet. While web search still is essential, the focus has shifted to so called vertical search. These are highly specialized search engines that typically cover one area of interest. By using domain specific knowledge, the vertical search system can implement techniques that would be highly beneficial for each topic, with unique ways of presenting the information to the user. This is quite different from the field of web search, where the focus is on fast and flexible search, attempting to cover the whole World Wide Web.

Yahoo! News is such a vertical, and is a combined news portal and news search engine. Yahoo! News gather news articles from many of the largest news agencies and let the user search in these documents, much in the same way as regular web search. One of the challenges in Yahoo! News search is that people are not familiar with using a search engine to find interesting news articles, much the way they use web search. What we will focus on in this thesis is to develop a navigation assistant for news search, which will help the user find interesting news stories. This approach will be based on finding semantic relations in raw Yahoo! News search logs. Much information lies hidden in query logs. What are currently the most important topics and how do users combine different terms formulating their queries. By focusing on what the user is interested in, we get a very direct focus on the user information needs and how to satisfy them.

This thesis will combine different approaches within the field of data mining and information retrieval. A typical strategy in query log analysis is to focus on one approach alone. What we propose is to combine different techniques within log analysis to get a combined result better than any technique can achieve on its own. Essential log analysis approaches that will be included in this thesis are: session analysis, temporal correlation and query segmentation. The end result will be a prototype search implementation that will use the semantic relations found to help the user meet his or hers information needs.

1.1 Approach

As already stated, in this master thesis we perform a novel query log analysis resulting in an improved search engine prototype. First the raw Yahoo! News query logs are preprocessed to extract relevant data for further processing. Two key contributions have been selected as the chosen approach.

The first contribution is based on analyzing session data. For each session extracted from they query logs, the most frequent term within the group is found. The queries forming the group are then considered as a vocabulary of related queries for this term. Groups with similar term labels are merged and form session term groups.

The second contribution is using temporal correlation to find queries that have a similar query frequency variation. Queries that show similar variations over time are likely to be semantically related or queries that often appear together.

These two contributions are merged into what we call related term groups. These are groups created by taking the intersection of overlapping groups from the session analysis and the correlation analysis. The implemented segmentation algorithm plays a vital role in finding the most prominent term or phrase, which function as a merge token for similar groups of each contribution.

These related term groups are converted into a tree structure for use in the prototype web application. The web application works as a middle layer between the HTTP server Jetty and the prototype Vespa news search engine from Yahoo!. When users do a search, their queries are mapped against the term group tree structure and if a match is found show this to the user. The user then has the option to navigate in the tree structure to narrow his or her search down different paths.

1.2 Results

The results from the query log analysis show the merits of the approach in this thesis. By combining two independent log mining contributions and intersecting the results, it gives much better results than each approach on its own. This results in a fair number of related term groups with what we judge to have a very high accuracy. A complete listing can be found in appendix A.2.

The results from the evaluation shows promising results. It must be stressed though, that the prototype search application is only one of many possible uses of the related term groups. The evaluation shows that the tree navigation structure was slightly above medium help and the improved system has a good ability to suggest additional 'side stories' to the user. Though the regular system and the improved system showed similar quality in finding unique news stories.

1.3 Outline

In chapter 2 we give an introduction to the Yahoo! News portal and Yahoo! News search. In chapter 3 we give an overview of the technical background which have formed a basis for the assignment. Chapter 4 contains a summary of related work within the field of query log mining. In chapter 5 we explain the approaches taken in this thesis. Chapter 6 provides implementation details. In chapter 7 we present the results of the evaluation of the prototype search system.

In chapter 8 we discuss different choices and difficulties discovered during the work. We end this thesis with a conclusion and further work in chapter 9.

Chapter 2

Yahoo! News

Yahoo! News consists of a combined news portal and a news search engine [1]. The news portal as seen in figure 2.1 features browsing of the most recent news in a series of different categories. The news articles are supplied by Yahoo! partners and the articles are presented as an integrated part of the Yahoo! News web site. News articles are continuously added and presented on the front page, with the opportunity to navigate deeper into the different categories to provide for a more complete news listing.

Yahoo! News feature a search engine that is specialized in news articles search, as seen in figure 2.2. Yahoo! News search is what is called vertical search [2]. Vertical search is similar to regular web search, but is specialized to cover a specific domain. By focusing on one topic, vertical search can adapt to the domain and implement domain specific features to improve search. Web search uses a general approach and will therefore lose some quality to vertical search, but is able to give good results for a wide range of domains.

Yahoo! News search allows for search in a large collection of accumulated news articles, both from partners and external web sites. New articles are continuously added to the document collection and the search index is updated. Documents are ranked with web search techniques with addition to some domain specific features, such as how recent the news is and emphasis on news article headings.

Figure 2.1: Yahoo! News

Figure 2.2: Yahoo! News search

Part II

Theoretical Background

Chapter 3

Theoretical Overview

In this chapter we give an overview of the technological background that has been considered when working with this thesis. First an overview of the information retrieval science is given and then a more in detail look at vertical search. Different approaches for text preprocessing techniques and for query log analysis have been studied. In the final part of the chapter we take a look at data mining and some of its techniques, before including the time aspect of data mining in temporal data mining.

3.1 Information Retrieval

This section gives a short introduction to Information Retrieval, followed by a description of vertical search.

3.1.1 What is Information Retrieval

Information Retrieval (IR) is the science of searching for information in documents, finding a set of relevant documents from a specified query. Information retrieval is closely related to data retrieval [3]. The goal of data retrieval is to retrieve exactly those documents that satisfies defined conditions and no others. Information retrieval on the other hand focuses on interpreting the information, and can often include noise and irrelevant documents. This problem is improved by analyzing and ranking the documents. Attempting to present the user with the most relevant documents to the user query with as few non-relevant documents as possible.

3.1.2 Vertical Search

Vertical Search is a relatively new area in the Internet search industry that focuses on specific domains of interest. By using specific domain knowledge, a vertical search engine can limit its view to a specific area. Thus getting a much higher quality of service than a broad-based approach (web search) that indexes the whole World Wide Web.

The word 'ceramic' is a common material used by dentists in their line of work. A search on Yahoo! yields 36 million pages, most of the high ranking pages evolving around enthusiast with pottery as a hobby. If the user instead did the same query on 'DentalProducts.net' he would get far more relevant results [4]. Vertical search got the advantage that it extracts the information from well defined sources within its domain and by adding domain specific knowledge it can better adapt to the user's needs.

3.2 Text Preprocessing

In this section we present different text preprocessing techniques that are highly relevant to this thesis. Text preprocessing is used to 'control' the vocabulary by aggregating and removing some of the unnecessary words. The techniques being looked at is stemming, lemmatization, and stop word removal.

3.2.1 Stemming

In natural language words follow morphological rules that allow the speaker to explicitly describe some relation or state of the word. These can be inflected forms, singulars, plurals etc. The words 'connects', 'connecting' and 'connection' all share the common base form 'connect'. By removing the prefixes and suffixes of a word one are left with the stem of the word, which in this case is 'connect'. Stemming can greatly increase information retrieval performance, because by using stemming the different variants of the same concept are reduced to one common term [3]. While using stemming seems reasonable and effective, the technique has its drawbacks. By reducing to the stem the word can lose some of its meaning, leading to reduced precision.

Light Stemming

Light stemming is done by reducing only plural and gender form of words to their natural stem [5]. For instance reducing 'cars' to 'car'. This is a more sensible approach for use in web search because this makes little impact on the precision, but gives some reduction in index terms.

3.2.2 Lemmatization

Lemmatization is also a process of finding the normalized form of a word, but with a different approach. While stemming uses defined rules to find the stem, lemmatization is mainly based on doing lookups against a dictionary to find a base form (though attempts have been made to make algorithms for lemmatization [6]). The limitation of lemmatization is that it can only find the lemma of known words, but the advantage is that the base forms found are true normalized words. The words 'computes', 'computing' and 'computed' would all be stemmed to 'comput', with lemmatization each word would be matched to the infinitive verb 'compute' by a dictionary. It has been suggested to use a hybrid system, where words in the dictionary is lemmatized and those not in the dictionary stemmed.

3.2.3 Stop Words

Stop words are words that appear frequently in the document collection. Documents are typically characterized by keywords that are special for a specific topic and rarely found in documents about different topics. Stop words on the other hand appear frequently in a large portion of the document collection and span a wide collection of topics. Other than removing words with low interest, eliminating stop words also gives the benefit of greatly decreasing the search index size, typically by around 40% [3].

Stop words removal is little used in today's web search engines, despite these benefits. This is due to how stop word removal is decreasing recall and can in some cases cause unwanted results. A query for 'The Who' would result in an empty query if stop word removal was implemented. Even though web search engines do not implement a complete stop word removal, search engines like Google seem to use a limited stop word removal for words like 'a', 'by', 'of', 'the' and a few more.

3.3 Query Log Analysis

The popularity of search engines has drastically increased in the last few year, along with the growth of the Internet. Millions of users will daily use search engines in their search for information. From this there has been developed a great interest in computing statistics of search engine query logs and finding user patterns, because it is no longer enough to simply look at the content in the documents to give an acceptable user experience. By using query pattern analysis, the search system can look into how and what the users are looking for, thereby increasing the chance that the system will provide the user with relevant data.

3.3.1 The AltaVista Query Log

As an example of query logs and what information one can expect to find, we give an overview of the query log from AltaVista [7].

Submitted query:

- **Timestamp**; indicating the time the query was submitted.
- **Cookie**; to uniquely identify a user (for e.g. registering a user session).
- **Query terms**; the exact words submitted.
- **Result screen**; requested range of search results.
- **User-Specified modifiers**; such as a language restriction on the results.
- **Submission information**; such as whether the query was simple or advanced.
- **Submitter information**; information such as what type of web browser, ip address and local language.

3.3.2 First-Order Analysis of Query Logs

First-order query log analysis involves counting only single terms, and does not look at the relations between different queries. First-order analysis can mainly be divided into two categories: analysis of individual queries and how queries are modified throughout a user session.

In the analysis of individual queries it is an important property that most user queries consist of very few search terms, compared to other information retrieval contexts. In the analysis of

AltaVista query logs an average number of 2.35 terms represented each user query [7]. Another important issue is the aggregation of identical search queries. Those queries that are often used are topics the users are interested in, and can be used for trend analysis or query term boosting. The analysis of AltaVista query logs showed that the top 25 most frequent queries consisted of 1.5% of the total user queries, despite being only 0.00000016% of the unique queries.

A user session is specified as the queries done by a unique user in a limited duration of time. By analyzing user sessions one can gain knowledge about how the user tries to find information about a certain topic, what links he clicks and how he tries to refine his search. What have been found in analyzing sessions is that sessions are usually very short and that the user will often only look at the top ten results from the result set.

3.3.3 Second-Order Analysis of Query Logs

Second-order query log analysis involves counting pairs of terms or queries. In second-order analysis one are interested in the relations between pair of objects. If terms often appear together in the same queries, it could be deduced that those terms are strongly related. Association rules (see Section 3.4.1) are based on statistical methods, and can say if two or more terms are related by a given threshold.

Correlations can be used to find queries that have similar changes of popularity over time. By comparing the normalized frequency curve of the queries above a certain frequency threshold, one can find queries that show the same variation over time [8]. These will typically be queries that have a high semantical similarity or which are strongly related.

3.4 Data Mining

Data mining in its simplest form refers to extracting or 'mining' knowledge from large amounts of data, also known as Knowledge Discovery from Data(KDD) [9]. Data mining involves an integration from multiple disciplines such as database, data warehouse technology, information retrieval, statistics, high-performance computing, machine learning, pattern recognition, and spatial and temporal analysis. The emphasis in the area of data mining is efficient and scalable mining techniques. Techniques used need to have a close to linear proportion with the data size, utilizing the available amounts of memory and disk space.

A data mining algorithm is largely looking for the following: classifying the data into sets of predefined groups, clustering data in groups based on logical relations, identifying associations between data or finding sequential patterns. We will take a closer look at association rules, temporal data mining and temporal correlations in the following sections.

3.4.1 Association Rules

In data mining association rules are used to discover elements that often co-occur within a data set, and to create rules such as implication and correlation between those elements. A well known example to illustrate association rules is an electronics store transaction database, with single transactions such as buys(PersonX, 'computer') [9]. From mining the transaction database one could create a rule such as:

$$buys(PersonX, `computer') \Rightarrow buys(PersonX, `software')[support = 1\%, confidence = 50\%]$$

A 50% confidence represents if the person buys a computer, it is a 50% chance he will also buy software. A 1% support means that 1% of all of the transactions under the analysis, shows that computer and software were purchased together. This rule is referred to as a single-dimension association rule, due to it containing a single predicate(buys). An association rule including many predicates is referred to as a multidimensional association rule.

3.4.2 Temporal Data Mining

Temporal data mining is closely related to data mining and uses much of the same techniques. The difference is that temporal data mining is concerned with analysis of ordered data streams with temporal interdependencies [10]. While most of the data mining techniques treat temporal data as unordered collection of events, temporal data mining uses this information to gather extra knowledge in the data mining process.

The goal of temporal data mining is to discover relations between sequences and subsequences of events [11]. The discovery of relations can be divided into three parts: representing the data sequence in a suitable form, defining similarity measures between sequences, and the application of the data model to the data mining problem (representation, similarity and operations). Different techniques are chosen depending on the nature of the data material. A sequence composed of a series of symbols from a defined alphabet is usually called a temporal sequence, while a series of continuous real-valued elements is know as a time series. Time series and temporal sequences can be found in numerous domains ranging from engineering, research and finance to medicine. With applications using temporal techniques the user can be assisted in diagnosis, and prediction of future trends and behaviors.

3.4.3 Temporal Correlation

Chein and Immorlica [8] suggest an approach to use temporal correlation to find semantically related queries based on their query frequencies. It is natural to think that items that vary in a similar pattern have a high probability of being related or semantically similar.

$$\frac{1}{d}\Sigma_i\left(\frac{X_{p,i}-\mu\left(X_p\right)}{\sigma\left(X_p\right)}\right)\left(\frac{X_{q,i}-\mu\left(X_q\right)}{\sigma\left(X_q\right)}\right) \tag{3.1}$$

To determine the correlation between two queries, they defined a correlation measure as shown in equation 4.1 where for query q, $X_{q,i}$ is the query frequency function for a given time span, $\mu\left(X_q\right)$ is its average frequency, and $\sigma\left(X_q\right)$ is its standard deviation. The correlation of two queries is then defined as the sum of the correlation factor for each time unit. By applying a normalization factor to all the frequencies, they get a result value between -1 and 1, where 1 means they are strongly correlated, 0 means they are independent, and -1 indicates that they are opposite.

Chapter 4

Related Work

In this chapter we take a look at related work within the domain of web search and more directly at the work done in analyzing query logs. There are many concerns regarding web mining, as described by [12] with the issue of information overload. Not only the technical difficulties with huge amounts of data that are constantly changing, but also on user difficulties with finding relevant information about the topic the person is after. Is the search system able to find the 'correct' pages that are relevant for the user query? And how to keep it up to date with the ever changing World Wide Web.

A taxonomy of web mining given in [12] splits the field into three main directions. Content mining such, as Google News [13] that filter out news stories from different news sites and find relations among the articles. Structured mining is about finding relevant web pages from a given query such as web search, and how to adapt the system to the constant changes of the Web. The last field covers usage mining, how user navigation patterns and query logs can be used to improve the user experience. In this thesis we will explore different aspect of query log mining with regards to entity extraction and news search.

The motivation behind query log mining is that queries are dynamic and will always reflect a current trace of users interests. Who can better know what is most important for the web sites users then they do themselves? By using the information hidden in query logs, we can utilize different strategies to find related queries and associated terms. This can again be used to improve search for users, either by improving ranking with higher precision and recall, or different forms of query recommendation. We will look at different strategies in the following sections.

4.1 Query Log Mining

The amount of information in the Web is continuously changing and growing, thus new areas of search engine technology improvements need to be examined. Several new techniques have emerged to improve the search process, and one of them is based on the analysis of query logs. Query logs register the history of queries submitted to the search engine together with different user ids and behavior information. Baeza-Yates [14] says that analyzing queries has a broad impact in Web search and design in two different aspects: Web findability and information scent. Web findability is a measure of how easy it is to find a Web site, where search engines are the main access tools. Information scent is how good a word represents a notion with respect to words with the same semantic meaning. For example, polysemic words(words with multiple meanings) may have less information scent.

There is limited public work done on analyzing query logs of the popular web search engines such as Google, Yahoo! and MSN due to commercial interests. Stallings et al. [7] performed an analysis over 6 weeks of raw query logs from the AltaVista Search Engine, with a total of over one billion entries. The characteristics found were that most queries were short with 2.35 terms in average and most users only looked at one answer page per query. Similar results were shown for Excite [15] and Fast [16] in table 4.1. About two thirds of the queries were only asked once, the top 25 most frequent queries formed 1.5% of the total number of queries asked, depsite being only 0.00000016% of the unique queries. In [14] it is said that queries follow a biased distribution. The frequency of query words follow Zipf's law [17], with results in a strongly logaritmic curve from the most popular queries to the least popular.

Measure	AltaVista	Excite	Fast
Words per query	2.4	2.6	2.3
Queries per user	2.0	2.3	2.9
Answer pages per query	1.3	1.7	2.2

Table 4.1: Query statistics for three search engines

Various work have been done within the topic of query log mining in the past few years. In the following sections we look at different approaches that will be used in our thesis, or that are closely related. The approaches can mainly be divided into two categories, those that use user behavior data and those that are based on purely statistical methods. Methods that use behavior data typically exploit information such as click-through data (what links the user have visited from a given query) or session information. Statistical methods look at decomposition of queries, frequency analysis and association rules among others. Each method has its merits, but the typical approach is to focus on one technique. What we want to do in this thesis is to combine different approaches to give a better result than each can give when used on its own.

4.2 Query Segmentation

One of the key issues in data mining, and thus query log mining is how to distinguish phrases and compounds from single words that are not related. Performance can be improved drastically with having a method to split up queries into their natural compounds. Take for instance the query 'american idol kelly pickler'. For the human eye it is easy to see that the query should be split into two compounds, 'american idol' and 'kelly pickler', and that treating each word as a single term gives a poor representation of the query. There are many challenges related to recognizing compounds as they are highly numerous, and new compounds are created every day. Keeping a dictionary over all known compounds could be exhaustive on computer resources and creating a dictionary manually is far too time consuming and not feasible. Thus an approach that can dynamically find compounds in the text corpora is highly valuable.

Risvik et al. [18] suggest an approach mining query logs, to produce segment candidates and compute connexity measures. Candidates are considered in context of the whole query and a list of the most likely segmentations is generated. For each segmentation, a segmentation score is computed from the connexity values, which can be used to rank the different segmentations(the higher the value the more probable the segmentation is). It is also suggested to use a proximity penalty to reduce the value of generic words that appear in many different segmentations. It is suggested to use a value between 2 and 4 terms for a valid segmentation. Risvik et al. conclude that proper segmentation brings considerable quality gain for search. By integrating it in the ranking formula, query segmentation and scoring delivers a new valuable block of information in the query answering pipeline.

Medelyan [19] also suggests a similar approach that uses query logs for compound extraction. To show that the techniques presented are interlingual, query logs from English, Dutch and German were used. The initial approach is to include in the compounds dictionary all phrases consisting of two or more words, which exceed a chosen co-occurrence frequency threshold. Two modifications are done to the initial approach: expanding and decompounding. Expanding is when the phrase X is a subphrase of X+1, then the phrase X is replaced by X+1 in the dictionary. Decompounding is to split phrases after the following conditions: phrases in the base form dictionary will not be split, the shortest decompositions will be chosen and in case of several possible decompositions, the one with the highest probability will be chosen. Medelyan exploits a hidden advantage in query logs, some users write highly correlated phrases as one word(e.g. 'matrixreloaded' instead of 'matrix reloaded'), and add these phrases to the base form dictionary. It is also suggested to use electronic dictionaries such as WordNet [20] to add to the base form dictionary to improve the method's quality.

Both Su et al. [21] and Zhang et al. [22] look at compound extraction in large document corpora. Su et al. suggest an approach where the corpus is scanned with a 2 and 3 term window and then calculate the measures' mutual information and relative frequency. Mutual information is a measure of association, comparing the probability of a group of words occurring together to their probabilities of occurring independently. Relative frequency count is the total number of occurrences of the compound over the average number of occurrences of all entries of compounds. This is further improved by incorporating parts of speech information to increase recall and precision. Zhang et al. use a statistical approach to extract Chinese compounds in a large corpus. The approach use mutual information combined with context dependency to extract bi-, tri- and quad-grams.

4.3 User Sessions and Behavior

A user session is specified as the queries done by a unique user in a limited duration of time. The user will often need to modify his query to find what he is looking for. Either because of too many irrelevant results, or the query did not yield what the user had intended. By analyzing user sessions and looking at the user behavior one can capture semantic relations among words appearing in the Web search logs. Users will often refine their queries, adding more terms to the original query to narrow down the search. Extended query logs can also capture what links the user visited from a given query. Work has been done on analyzing click-through data(visited web pages from a given query) to extract related terms from documents the user selected. Even though session information is mostly reliable, there are some sources of unreliability [12]. Several users hidden behind a proxy can be appearing as a single user, cookies can be disabled or many people accessing the Web from the same computer.

Noriaki et al. [23] propose a novel log analysis using user behavior models to extract semantic relations among terms used in user sessions. The approach is a statistical approach to capture semantic relations without manual or natural language processing. The approach analyzes the users' state of information need, to model the users' intentions and thereby be able to draw conclusions about the semantical relations in the user sessions. This approach is based on frequencies of words and co-occurrence, while the other direction, a symbolic approach, tries to find syntactic patters in the form of regular expressions.

Noriaki et al. model the user behavior into five different groups. Paraphrase is when a similar word is substituted. Specialization is when a general word is replaced with a more specific word. Generalization is when a word is used to broaden the topic. Alternative is when a word is replaced within the same topic and lastly interruption which is a change of topic of the session. A selection of characteristics were observed: the more dedicated the user is for a topic the faster he refines with new queries, the latter a word is used in the search sequence the more important it is in specifying the refined information, the closer words appear the more they are related and words with a high frequency should be emphasized.

In the paper by Seco et al. [24] a 6-month query log for tumba!, a Portuguese search engine, is analyzed. A session is defined as all entries that use the same Agent, same IP adresse, the time interval between the sessions last entry and the current is below a defined threshold and at least one term used in the session overlaps with the terms of the entry. Jensen et al. [25] suggest an approach to analyze session logs to find query phrase suggestions. Suggesting query phrases other users have found necessary to add for a given query(mined from session logs), dramatically improves the quality of suggestions over simply using co-occurrence frequency measures.

4.4 Temporal Analysis

Chien et al. [26] suggest an approach to find semantically related web search queries based on their temporal correlation. In other words, one can infer that two queries are possibly related if their popularities behave similarly over time. By taking the frequency functions of the queries in the log and calculate a correlation coefficient, one can above a defined threshold find strongly correlated queries [26]. This approach gives two unique advantages: the approach explicitly 'understands' why a query is interesting at a particular time and it is able to quickly adjust to news events.

The approach suggested by Chien et al. [26] makes use of a set of experimental parameters. The query log is aggregated into discrete intervals, into time units such as e.g. an hour. How long duration for the correlation analysis and a definition threshold for meaningful correlations. Two distinct types of queries are found, event driven queries such as 'national convention' that are best found with long time units, and periodic events such as 'Disney' that perform better with shorter time units. The approach gives promising results but have two distinct weaknesses. The approach only gives results for a fraction of the input queries and discards many queries that are not similar enough in query frequencies. There will also be many false positives. Therefore it is strongly suggested to use temporal correlation in conjunction with other text mining techniques to help filer out many of the false positives.

Maslow et al. [27] have used a similar approach with temporal frequency analysis to extract news relevant queries from a Web search query log. The algorithm is based on calculating query significance, representing the change of query frequency compared to a defined time interval. To improve precision, queries from the previous day(24h before) have been compared to remove queries that have a daily natural variation(such as weather forecast). Very rarely used queries are removed and queries that have a high relevance with a high amount of the total queries.

Beitzel et al. [28] have performed a temporal analysis on a very large topically categorized Web query log. Some notable observations were found. The number of queries are substantially lower during non-peak hours than peak-hours, but the average number of query repetitions does not change significantly on an hourly basis compared to the previous day. Most queries appear no more than several times per hour and the queries received during peak hours are more similar to each other than their non-peak counterparts.

4.5 Statistical Analysis

Medelyan [19] has implemented an approach using statistical extraction of semantically similar terms from Web search query logs. The algorithm takes the 10.000 most frequent terms of the query log and estimates a similarity measure for each term pair, due to their co-occurrence behavior in the query log. The result is a similarity dictionary, a ranked list of pairs above a specified similarity threshold.

The analysis uses a set of different factors. The more query terms a sub-query contains, the higher the contextual information is in the query. The more distinctive terms co-occur with sub-query in the log, the lower its relevance is. The terms of a sub-query and the order of query terms are also considered. The information content of each sub-query can then be calculated used together with the term frequencies to form a weighted value. This value is used to determine the direction of the semantic relation between terms and the quality of the relation.

Lee et al. [29] suggest an approach using association rule mining to create an association thesaurus for interactive query expansion. Based on click-through data, a set of words are extracted from each web page that was visited originating from the query. The algorithm is then used on these term sets for term correlation mining to create a corresponding thesaurus for each query. This thesaurus will then in turn be used to recommend possible query refinements for the user.

4.6 Clustering

Clustering has been used in many different web search query logs mining approaches, either as the main focus or used in conjunction with other techniques. The papers from Wen et al. [30], Beeferman et al. [31] and Baeza-Yates et al. [32] are all based on the use of click-though data, that is information of what links the user visited from his query result page. Wen et al. is based on the fact that if users select the same documents with different queries, they are regarded as similar. If a set of documents are selected from a set of queries they are considered equal to some extent. The algorithm has some key features: no set number of clusters, low frequencies are filtered, its effective on large data and it works in an incremental fashion.

Beeferman et al. use click-through data to discover clusters of similar queries and similar URL's. By viewing the dataset as a bipartite graph, with the vertices on one side corresponding to queries and on the other side to the URL's, one can apply an agglomerative clustering algorithm to the graph's vertices to identify related queries and URL's. The algorithm does not rely on the content of the web pages, but instead uses co-occurrence information across multiple transactions to guide in the clustering.

Baeza-Yates et al. suggest an approach that is ground on clustering the clicked web pages based on extracted keywords of their content. This is again used as a base for clustering the user queries into clusters named with appropriate names.

4.7 Classification

Accurate classification of user queries can in many situations allow for increased effectiveness and efficiency in general-purpose Web search systems. If e.g. a query was recognized as a phrase within a sports topic, the returned results about sports could get a ranking boost. This is specifically useful for search engines using topic-specific back-end indexes.

Beitzel et al. [33] suggest an approach that combines manual matching and supervised learning to increase effectiveness over a single technique used alone. Manual classification requires too much time and resources to be feasible and supervised learning alone does often not provide adequate results. The approach suggested creates a rules based automatic classifier using selectional preferences mined from a very large query log. This is combined with exact matching against a large set of manually classified queries(18 lists of categorized queries classified by a team of editors at AOL) and a weighted automatic classifier trained using supervised learning.

Shen et al. [34] suggest an approach called query enrichment. The algorithm takes a short query and maps it to intermediate objects. Based on the collected intermediate objects, the query is then mapped to the target categories. To build the necessary mapping functions, an ensemble of search engines is used to produce enrichment for the queries.

Part III

Prototype Implementation

Chapter 5

Approach

In this chapter we describe the approach taken to find related term groups from raw news query logs. The techniques chosen are based on the two previous chapters, theoretical overview and related work. We discuss why the different techniques are chosen and how to best apply them to our thesis. The chapter starts out with a comment on the constraints concerning our work, and then the different processing steps that the system go through to produce the final result.

5.1 Constraints

Different constraints concern this thesis. The most important is the fact that a majority of the theory and related work is used in conjunction with web search. Is it justified to use the same techniques for new search? Other important constraints are the data material, the raw news search query logs, and the prototype installation of the Yahoo! Vespa search engine.

5.1.1 News Search

The focus in this thesis is on the extraction of semantically related terms that are valuable in news search, but the techniques used are mainly found within the area of web search. In news search there are important differences such as news articles are added continuously by the minute, while in web search the Web is crawled at regular intervals. News articles also bring in some sense of structure that is much more reliable than what is found in web pages. While time and article title is of high importance for news search, it is not emphasized in the same way in regular web search.

Despite quite some differences between news search and web search, we believe that in large these have little effect on information retrieval techniques. Certain improvements are added to the news search engine to find news articles in a different way, but these are mostly small changes in the search engine configuration and have small effect on the information retrieval theory. This is more prominent in the analysis of query logs, where the log content is quite similar for web search and news search.

5.1.2 Raw News Search Query Logs

The thesis is based on the analysis of raw news search query logs from Yahoo! News search, that span over a duration of 3 days from 2007. The logs are separated into hourly intervals and consist of all requests to the Yahoo! News portal, such as queries from Yahoo! News search and RSS requests [35]. The log entries contain user information, a timestamp, the request string and different control information.

5.1.3 Vespa Search Engine

In addition to the raw query logs from Yahoo! we have been given access to a prototype system with Yahoo!s search engine Vespa. This is a basic installation with a news document feed corresponding to the date of the query logs. Due to this being a regular Vespa installation it will differ some with the live commercial system at Yahoo! News [1], because of news specific improvements that our prototype system does not have.

In the evaluation of our data analysis we implement an 'improved' version of news search that is based on the results found. This will work as a front-end on top of the prototype Vespa installation to provide an alternative search opportunity.

5.2 General Idea

The general idea in this thesis is to combine different approaches within the field of query log mining, to create a new and improved approach to extract information from query logs. The main idea is to combine user session information and temporal correlation on query frequencies. The benefit of combining two independent techniques is to enhance each method's strengths, which will produce better results than each can obtain on its own.

The session analysis evolve around user sessions, which are specified as a series of queries done by a unique user in a limited duration of time. Within a user session the user is often looking for information about a certain topic. Either expanding the query to get more specific results or removing terms or adding more general terms to broaden the search. Noriaki et al. [23] use session analysis to find semantic relations among query terms and Jensen et al. [25] use it for query phrase suggestions.

The temporal correlation analysis of query frequencies is used to find queries that behave similar in popularity over time. Terms that show the same behavior pattern have a high chance of being semantically similar or closely related. Chien et al. [26] use temporal correlation to find semantically similar terms based on hourly and daily search query logs.

Both the session and temporal correlation approaches have a few notable weaknesses. Session analysis is troubled by low frequencies on expanded queries, making it uncertain that terms are in fact related. Temporal correlation has the problem that it requires strong constraints to give good results, thus rejecting much of the data material. Lowering the criteria on the other hand yields many false positives. The key idea in this theses is to use both session analysis and temporal correlation analysis with low parameter criteria and then combine the matching results from both areas. This will result in an approach that gives high accuracy on the results without rejecting a large portion of the data material.

5.2.1 Prototype Overview

The prototype application as seen in figure 5.1 can be divided into three separate phases. The first phase is a preparation phase where the raw news search query logs are processed and the relevant data extracted and stored to disc. The second phase is the query log analysis phase, which is the main part of the application. The first phase is mainly preparation of data and the third phase is presentation and use. The log mining phase consists of two major contributions, the session group analysis and the temporal correlation calculation. Segmentation of queries is used to improve the quality by supporting phrases such as 'president bush' instead of just 'bush'. The results of the two approaches are combined in term group creation where groups of related terms are created. The final phase is a web application that utilizes the created term groups to improve news search, as a front-end on top of the prototype Vespa News installation.

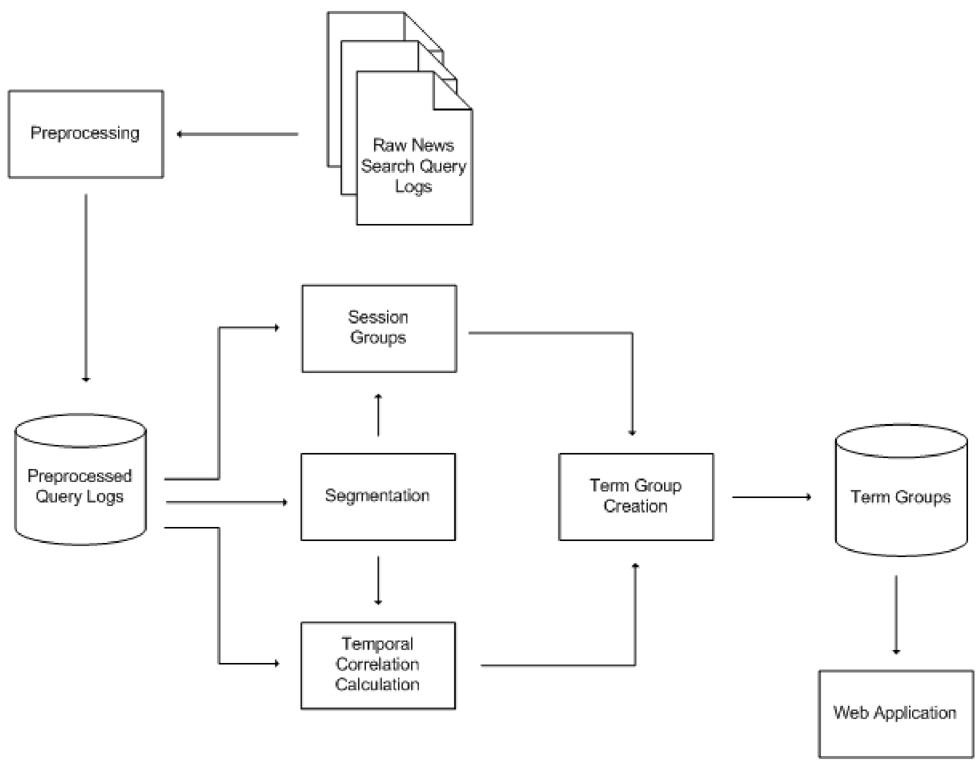

Figure 5.1: System design overview

5.3 Preprocessing

The raw news search query logs are extensive and cover all requests to the Yahoo! News portal. This includes News search, RSS search [35] and different computer generated requests. The interesting entries for our analysis is the News search requests, the rest are removed in the preprocessing. We have decided not to use the RSS entries because these are generated by RSS agents. They are used to poll the Yahoo! News search engine at regular intervals to retrieve new articles about user specified topics. They will thus not contain the query reformulation needed for session analysis and will disrupt the frequency patterns for the temporal correlation calculations instead.

The query log entries contain information about the user, a timestamp, the request to the server, information about the browser, request type, control information and more. What we use is the unique user identification, a timestamp and the search query string. The rest is discarded. Empty queries and queries containing only one term that is a stop word are also removed. This will give drastic reduction in data volume and help speed up the processing time for the log analysis.

5.4 Segmentation

An essential part in any text mining application is how to best handle phrases and words that often appear together. If queries are considered as single terms, information will be lost and the quality on the analysis will be degraded. Take for instance the query:

'american idol kelly pickler'

If this query was to processed as single terms the semantic relations among the words would be lost.

'american', 'idol', 'kelly', 'pickler'

While a good segmentation algorithm would be able to split the query into its natural parts.

'american idol', 'kelly pickler'

Risvik et al. [18] use a simple approach where all the combination of words consisting of 1-4 consecutive terms are added to a database along with their query frequencies. When a query is to be segmented, all combinations of consecutive terms are found and a score calculated based on the database frequencies. The segmentation with the highest score is the most likely segmentation based on query frequencies.

We suggest a slightly altered version to find the most prominent phrase or term within a query string. To give phrases a stronger weight the score is multiplied with the number of terms in the phrase. There is also two hidden features in query logs that can be exploited. Some users write phrases using quotes('president bush') and highly correlating words as one single term('megamillions' -> 'mega millions'). These phrases are added to a phrase dictionary and given extra weight.

5.5 Session Analysis

Different approaches have been used to analyze user sessions to find semantically related terms and terms suitable for query expansions or suggestions, such as those described in section 4.3. Our goal with the session analysis is to find a keyword best describing each session and consider the queries in that group to be a vocabulary describing the keyword.

After all the sessions have been tagged with a keyword, they can be merged with sessions with a similar keyword. This gives a list of keywords and which queries users have used in conjunction with them. To get high quality session groups, it is important to remove unwanted queries. Sessions consisting of only one query and queries that do not contain the session keyword will be removed because they add a lot of noise. To improve the session group merging, the system will use segmentation(explained in 5.4) to allow for phrases as keywords, not only single terms. Such that e.g. 'president bush' is a candidate for a session group and not only 'bush' or 'president'.

5.6 Temporal Correlation Calculation

Temporal correlation is an essential part in our approach in query log analysis. As suggested by Chien et al. [26] one can find semantically related web search queries based on their temporal correlation. When two queries have the same query frequency patterns there is a high probability that they are related in some way. Either semantically or that they appear in the same situations. A correlation measure is calculated after the equation in 3.4.3 and by adjusting the threshold on the correlation value, one can vary the results from few high correlating queries to many medium to high correlating queries.

When calculating temporal correlations the data material needs to be split into discrete intervals, as the query logs are a continuous steam of log entries. The original raw query logs from Yahoo! come in one hour intervals and this structure is kept after the preprocessing. One of the challenges of mining news search logs is that the query frequency is fairly low looking at hourly logs. Most users browse news on the portal page or use RSS search, using the search feature is not so common. Thus it would be beneficial to be able to aggregate the hourly logs into longer intervals. Intervals over a longer duration of time will give more correlating queries, due to the characteristic of the correlation computation. If the coarseness of the data material is increased, the differences in query frequencies will be less thus yielding more results.

5.7 Term Group Creation

The related term group creation process is the combination of the two different query analysis approaches. The session analysis has found groups of related queries with a term or phrase marking the group. The temporal correlation calculations have for each query found a group of correlating queries. By applying the segmentation algorithm on the correlation query, the most prominent phrase or term is found. This phrase or term is then matched against the session groups. If such a session group is found, the list of correlated queries is filtered against the queries in the session group. Queries that are found with both approaches form a group of related terms, with the most prominent phrase or term as group label.

5.8 Web Application

A simple web application has been made to use the related term groups in conjunction with the prototype Vespa news search system. The open source Java based HTTP server Jetty [36] has been integrated into our system to provide for web support. The system provides a search interface that internally executes queries against the prototype Vespa system, receives the document result sets and presents them to the user.

The groups of related terms are converted into a tree structure for use in the web application. This is done by continuously using the segmentation algorithm to find the most prominent phrase in the remaining queries in the group. The queries containing the phrase forms a node and the queries in the new node is segmented recursively until the node contains only one query.

Chapter 6

Implementation

The prototype was developed based on the approaches described in the previous chapter. The implementation details is explained from the first processing steps to the final web application. The following sections mirror those of the previous chapter.

6.1 Overview

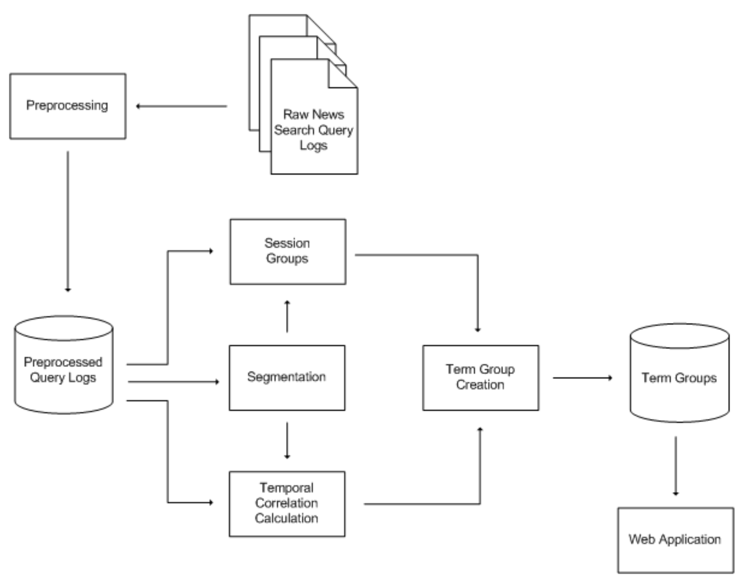

Figure 6.1: System design overview

A system design overview of the implemented prototype is show in figure 6.1. The system consists of 3 different phases. The first, the preprocessing step, extracts valuable information for further processing. The second step is the query analysis step, where the major work in this thesis has been put and is based on the theories presented in chapter 3 and 4. The third and last phase is the web application phase, where the output of the data analysis is put to work in a prototype search application.

The preprocessing and the query analysis is implemented in Sun Java 1.4.2 [37] due to the query logs being on a FreeBSD system with this Java version. The web application is implemented in Sun Java 1.6.0 [38] for Jetty HTTP server [36] compatibility and is no longer restricted to run on the Yahoo! system.

6.2 Preprocessing

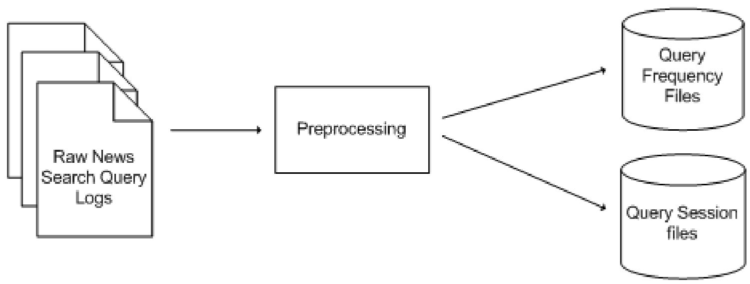

Figure 6.2: Preprocessing

The preprocessing phase is the process of extracting the information needed from the raw news search query logs for further analysis. As seen in figure 6.2 two new sets of corresponding files are created, query frequency files and query session files. Query frequency files are hourly aggregated query frequencies for the correlation computation, containing the fields query and query frequency. Query session files keep all log entries from the original raw logs, but remove excessive information. Query session files contain the query, a unique user identification and the timestamp for the log entry.

Before the log entries are added to the preprocessed files, they are filtered to remove some unwanted entries. The raw query logs include all requests to the Yahoo! News portal, thus we discard entries that are not generated from Yahoo! News search. Queries that are empty, consist of only one term that is a stop word and queries with more than 100 characters are removed. Characters that are not a letter, number, quote or punctuation are removed from the query strings. Quotes are used to extract phrases and is removed in a later step in the log analysis.

6.3 Segmentation

Segmentation is the process of splitting strings of words into their natural parts. As explained in section 5.4 this is mainly based on the co-occurrence of terms in the query logs.

During the preprocessing phase, a file with the total query frequencies during the 3 days period was created. The algorithm iterate over each query and create multi grams over consecutive terms in each query that contains 1 to 4 terms. These multi grams are added to a hash map along with a score. For single terms this is the query frequency. For 2-4 grams the number of terms in the multi gram is multiplied with the query frequency. This is to give more weight to phrases as single terms will always have the highest frequency.

Two features in the query logs are exploited to give extra weight to phrases. Some users write highly correlating terms as one term, e.g. ‘thematrixreloaded’ instead of ‘the matrix reloaded’.

The second is that some users use quotes around a phrase, these are words that often are very closely related or represent one unique unit. Compounds are weighted with 2 times the normal phrase score and phrases in quotes are weighted with 3 times the score. An example of the results of the segmentation algorithm is shown in table 6.1

Segment	Score
american idol	1.000
antonella barba	0.808
american	0.269
idol	0.269
antonella barba pictures	0.226
american idol antonella barba	0.206
antonella	0.174
barba	0.174
barba pictures	0.167
american idol antonella	0.161
idol antonella barba	0.161
idol antonella	0.116
american idol antonella barba pictures	0.095
idol antonella barba pictures	0.079
pictures	0.044

Table 6.1: Segmentation results of the string ‘american idol antonella barba pictures’

6.4 Session Analysis

Figure 6.3: Session group creation

A user session is a series of queries from a single user in a limited duration of time. These are found by iterating over the query session files and grouping queries with the same unique identifier. If a query is within a 5 minute window of the previous query it belongs to the same session. Sessions containing only one query are discarded as they do not contain any semantic relation among queries. Sessions containing more than 10 queries are also removed. This is to remove sessions where several people are behind the same proxy, thus getting the same user identification, but also to remove computer generated sessions that can contain numerous queries.

As seen in figure 6.3 after the sessions are found, they are analyzed and merged into groups. This is done by finding the most frequent term within each session. Queries in the session not containing this term are removed. This is important to keep precision high, else the groups would contain a considerable amount of irrelevant queries. After all the sessions have been tagged and filtered, they are merged into groups with a similar tag. To get the best term or phrase marking the group, all queries are sent to the segmentation algorithm and the score of possible segmentations are aggregated. The most prominent term or phrase is then chosen as a new group tag. Lastly the session groups are merged again in case of duplicate group tags.

An example of a session group is shown in table 6.2. This low frequency group is quite small compared to the more popular topics.

Frequency	Query
9	atkins diet
6	atkins diet plan
6	atkins
1	atkins diet guidelines
1	what do you eat on atkins diet
1	what is atkins diet like
1	atkins diet menu
1	atkins diet food list
1	atkins wsj
1	atkins diet recipes
1	free atkins diet plan
1	how much do you lose with atkins
1	atkins diet menus
1	atkins diet carbohydrates
1	atkins diet plan menu
1	what is the atkins diet

Table 6.2: Session group for the phrase ‘atkins diet’

6.5 Temporal Correlation Calculation

The temporal correlation calculations are based on the query frequency files created during the preprocessing stage. These are hourly files, that gives the option to easy adjust the total time duration of the correlation analysis and how many hours for each time unit. One of the downsides of using correlations is that they work poorly with low frequency queries. With the fairly low search count on Yahoo! News search, aggregating the query logs into 3h for a 24 time period and 5h for a 70h time period, gave the best data material.

A simple approach was chosen for the correlation calculation because of the relative small data material. Using the equation in 3.4.3, a correlation coefficient is calculated based on the variation in query frequency. This coefficient measure is a value between -1 and 1. The score 1 represents a perfect match in query frequency variation, 0 says the queries are independent and -1 for total mismatch. The correlation computations are quite processor intensive($O(N^2 * M)$ where N is the number of queries and M the number of time units) and are therefore filtered before the calculations as seen in figure 6.4.

For the correlation calculations over a 70 hour period, queries with a frequency lower than 25 in the total frequency file are removed. Queries with the frequency of 0 in more than 6 of

Figure 6.4: Temporal correlation calculation

the time intervals are also removed. This speeds up the computation time considerably. Due to the time frame of the query logs being only 3 days and the intention to find all relevant queries ignoring false positives, the correlation coefficient value is set as low as 0.5. This can be interpreted as queries that have a moderate to high similarity are added to the correlation list. An example of a correlation calculation result is shown in table 6.3 where the 10 highest matching correlations for the query 'american idol' are shown. A complete listing of the correlation group for 'american idol' can be found in appendix A.1.

Correlation Value	Query
1.0000001	american idol
0.9659458	american idle
0.95593166	idol
0.9530445	american idol.com
0.95264685	american idol fox
0.9489394	fox tv
0.93390244	american idol antonella barba photos
0.9063261	american idol 6
0.90311986	americanidol.com
0.8759125	american idol 2007

Table 6.3: Top 10 correlating results for the query 'american idol'

6.6 Term Group Creation

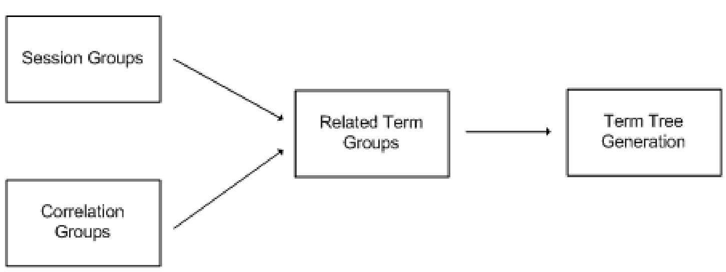

Figure 6.5: Term group creation

An overview of the term group creation is shown in figure 6.5, as the process of merging the results from the session and temporal correlation analysis. This is a fairly simple process where the segmentation algorithm is used on the correlating query to find the most prominent phrase or term. That term or phrase is then matched against a session group with the same label, if it exists. Queries that are found in both groups form a new related term group with their common term or phrase as label.

This process has been implemented in a simple straightforward way. First all the session groups are calculated and next the correlation groups, with all data stored in memory. Then they are matched against each other and the new groups found and written to disk. A complete listing of the related term groups can be found in appendix A.2. For larger amounts of data this is not possible, as the memory requirement increases almost quadratically with data volume. An approach where session and temporal calculations were continuously flushed to disk and then read selectively to carry through the matching of groups, would have to be implemented to handle larger data volumes.

After the related term groups are found they are further processed to create tree structures of the groups. A list of the tree structure of the 50 most popular groups can be found in appendix A.3(the rest of the groups are omitted to limit the number of pages of results).

The related term tree group creation process use a recursive algorithm, where all queries within a group are initially added to a candidate leaf a list. The segmentation algorithm is used on all queries in the candidate leaf list and the score for all segments for all queries are added together. The segment with the highest score is chosen as a new leaf and all queries containing one or all terms of the segment are added to that leaf's candidate leaf list. The segment forms the name of the new node and the terms of the segment are removed from the queries added to the node. This is done recursively until all candidate leaf lists are empty. With the use of segmentation the tree generation gives good merging of overlapping terms, but this also means that the original order of words in the query is lost. It gives a good ranking as the best segments are picked first and the worst last.

6.7 Web Application

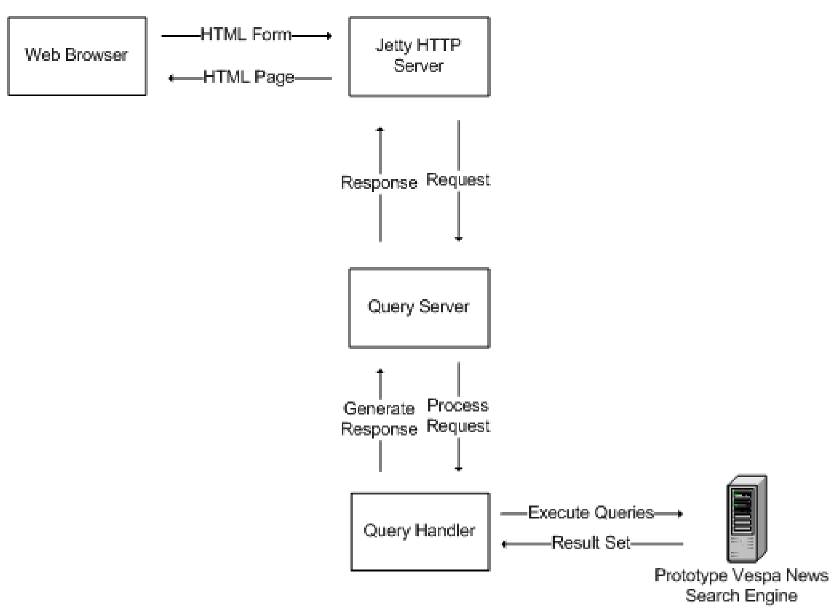

Figure 6.6: Preprocessing

To be able to view the results of our work in a news search environment and to do a relevant evaluation of the results, a simple web application was implemented. As seen in figure 6.6, the application is based on the HTTP server Jetty and function as a middle layer between the web interface and the prototype Vespa search system. Jetty is a free open-source web server implemented in Java [36] and works as an embedded component within the web application. The Query Server process creates an instance of a Jetty server and creates new instances of Query Handler for each incoming request. The Query Handler processes the request, executes appropriate queries against the prototype Vespa news search engine and generates a response for the user.

In figure 6.7 one can see an example of the implemented web application. On the left side regular search is shown, which is the query executed against the Vespa system without any modifications. On the right side is the 'improved system'. Queries are sent through the segmentation algorithm and matched against the created related term groups. If multiple groups are found, the user is able to choose from a list of matching groups. As seen in the figure, the tree structure is presented to the user together with a few hits from each of the main tree node's children. The user can either click on the query group labels to navigate around in the tree structure or click directly on the tree navigation structure to jump to a node in the tree.

Figure 6.7: Web application example

Part IV

Evaluation and Conclusions

Chapter 7

Evaluation and Results

This chapter describes the evaluation of the prototype search application. The main focus on the evaluation is to get a measure of the ability to find related news stories and how helpful the navigation tree structure is to the user. The evaluation is carried out by comparing an unmodified version of news search against the prototype search application.

Several aspects of the data analysis leading up to the final results used in the search application, have only been assessed through experimentation with the prototype code. These are choices that have both affected the quantity of term groups and their quality. The evaluation represents one way of judging the data found and does not give a full assessment of the log analysis process. This will further be discussed in chapter 8.

7.1 Evaluation Strategy

In this section we describe the evaluation test set up, what the evaluation focus is and what the expected results are.

7.1.1 Query Selection

In the related term groups there are some generic words and Internet community sites that are not very news related. This is most notable in the highest frequency groups. Therefore the 10 most news relevant groups were selected from the 20 related term groups with the highest popularity. The group title which is also the tree navigation main node was chosen as the query term or phrase.

The focus on the prototype search system is to evaluate the value of the system used in a news search context. We therefore think it is a valid approach to pick out a selected number of the most news relevant queries for the evaluation. The selected queries can be seen in table 7.1.

7.1.2 Test Setup and Focus

The prototype system was set up according to the approach described in chapter 5 and the implementation details in chapter 6. The news article feed used in the prototype Vespa search

Query Number	Query
1	antonella barba
2	american idol
3	britney spears
4	mega millions
5	baseball
6	immigration
7	grand canyon
8	president bush
9	tournament
10	anna nicole

Table 7.1: The 10 selected queries for the evaluation

system is from 2007. The query log analysis is based on 3 days of raw query logs from the corresponding time period. A screen shot of the prototype search system is shown in figure 6.7.

An evaluation group consisting of 6 people were asked to carry out the evaluation. They were given a simple description of the system and instructions on how to fill out the forms. The evaluation instructions can be found in appendix B.1. They were given a list of 10 links, each for one query to be evaluated after a set of measures. The three different aspects of the system to be valued were: the tree navigation structure, number of unique news stories and an evaluation of the improved search versus regular search.

7.1.3 Tree Navigation Structure

The tree navigation structure gives the user an overview of different subgroups of the query. The nodes that form the tree structure are terms that earlier users have written in conjuncture with the search query. With the tree structure the user can click directly on a tree node to view the search results of the node and its children.

In the evaluation we are interested in the usefulness this tree structure gives the user. The evaluation group is to judge if the tree structure gives the user poor, medium or good overview and navigation help.

7.1.4 Unique News Stories

The number of unique news stories is used to give a measure of how good the regular and the improved approach is to bring news stories to the user. The evaluation groups is to count the number of unique news stories each approach gives. A unique news story is in the evaluation defined to be: a news article about a selected set of people at a specific event in time. Two news articles about the same people connected to the same event, would thus be counted as one unique news story.

7.1.5 Improved News Search

The improved news search prototype is compared against the regular search approach. For each news article in the result set of improved search, the evaluation group is to judge if the news article is not relevant to the original query, if it matches one of the news stories in regular search or if it is a new news story. This is to give a measure of how good the improved approach is to broaden the search. By adding keywords to the original query, the search is pointed in different directions to give other interesting results about the topic.

7.2 Evaluation Results

A listing of the results of the evaluation is given in this section. The evaluation is split into three different measures: tree navigation structure, number of unique news stories and improved news search evaluation.

7.2.1 Tree Navigation Structure

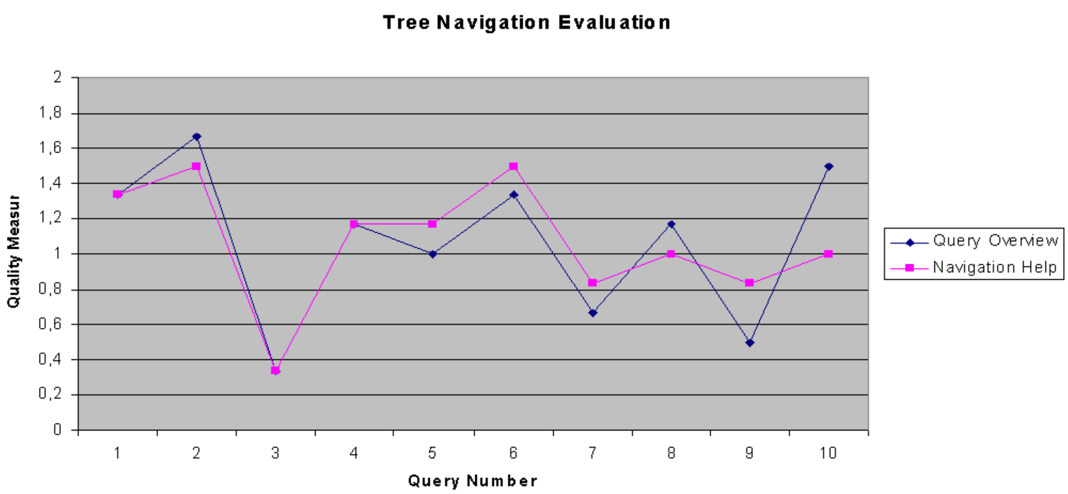

Figure 7.1: Results of the tree navigation structure evaluation

In figure 7.1 the results from the tree navigation structure evaluation are shown. A value of 0 represents poor quality, 1 for medium and 2 for good. The results vary from query to query, but the results are centered slightly above 1. This shows that the evaluation group found the navigation structure to give a partly good overview of the query and functioned partly useful as navigation help.

7.2.2 Number of Unique News Stories

Figure 7.2 shows the average number of unique news stories for each query. The results show a slight increase in number of news stories for the improved mode, but the values do not account for the result set of improved search being larger. In figure 7.3 the result set frequencies for

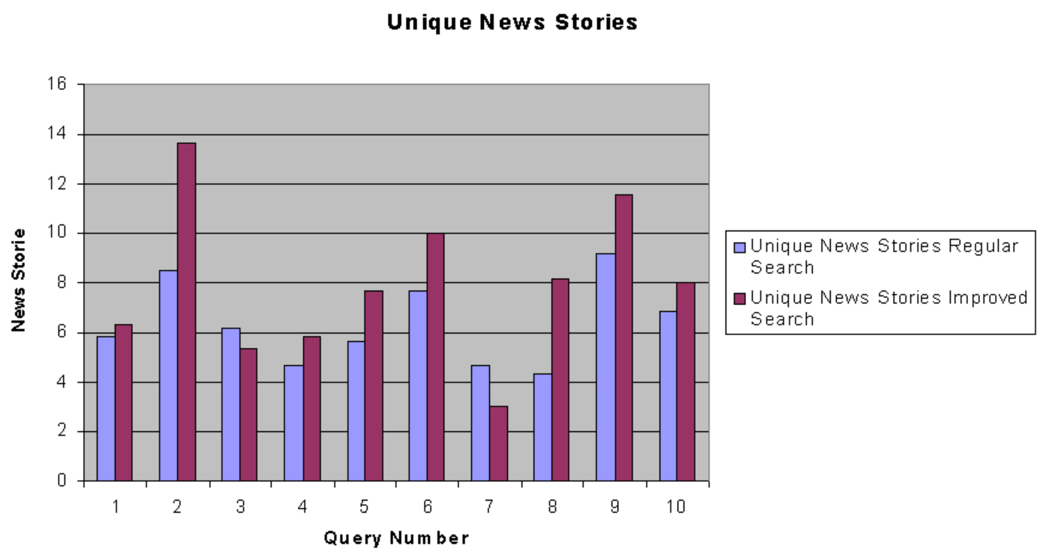

Figure 7.2: Number of unique news stories

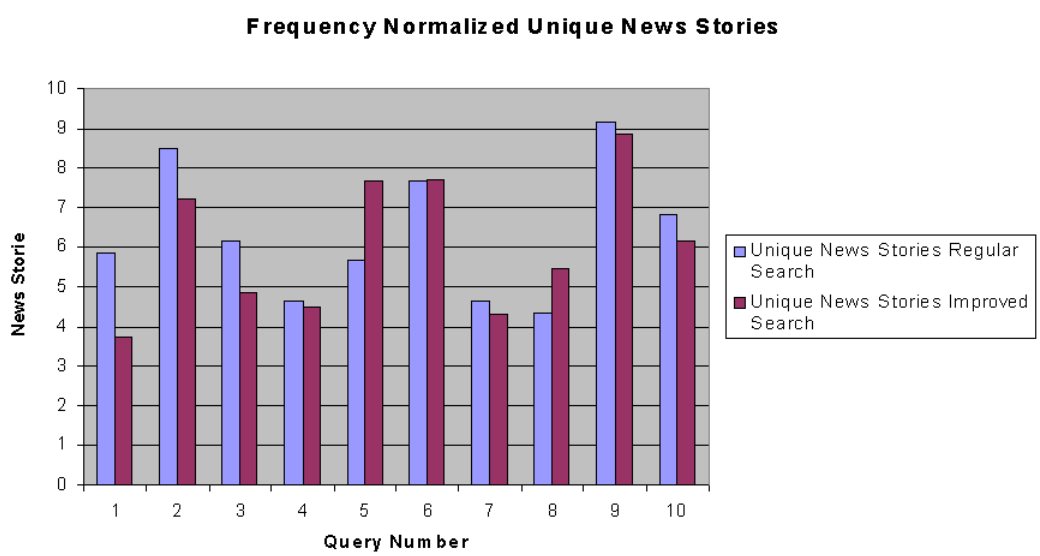

Figure 7.3: Normalized number of unique news stories

improved search are adjusted to accommodate the regular search approach. The adjusted graph shows similar values for improved mode and the regular approach, with a slight advantage to regular search.

Two issues arise when evaluating the number of unique news stories. One is that a larger result set will produce more duplicate news articles than a query consisting of few results. The second is the way the improved mode run the additional queries. To widen the search, a term or phrase is added to the original query. If the original query is very strong compared to the added term or phrase, it may ‘overrun’ the new query and produce results similar to the original query.

7.2.3 Evaluation of Improved News Search

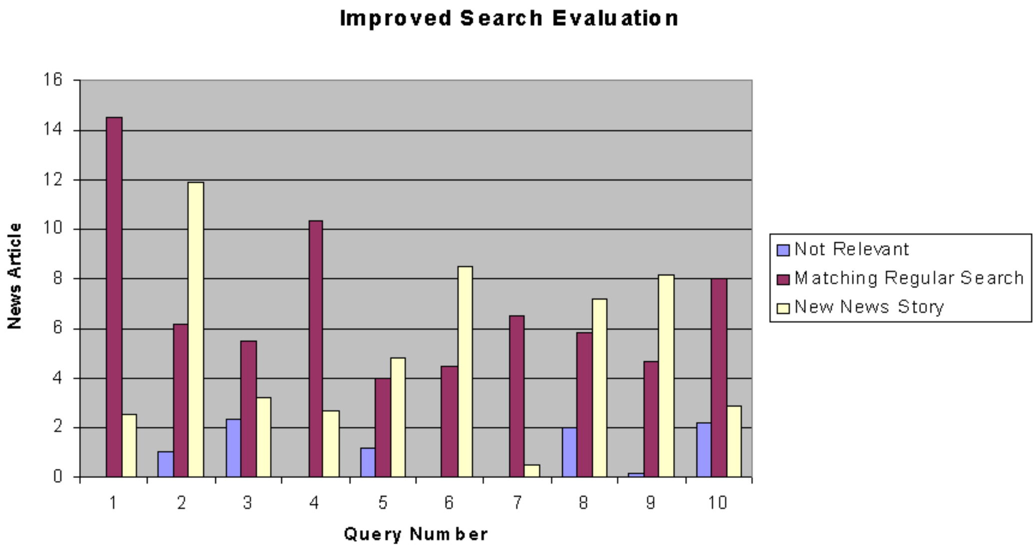

Figure 7.4: Evaluation results of the improved search mode

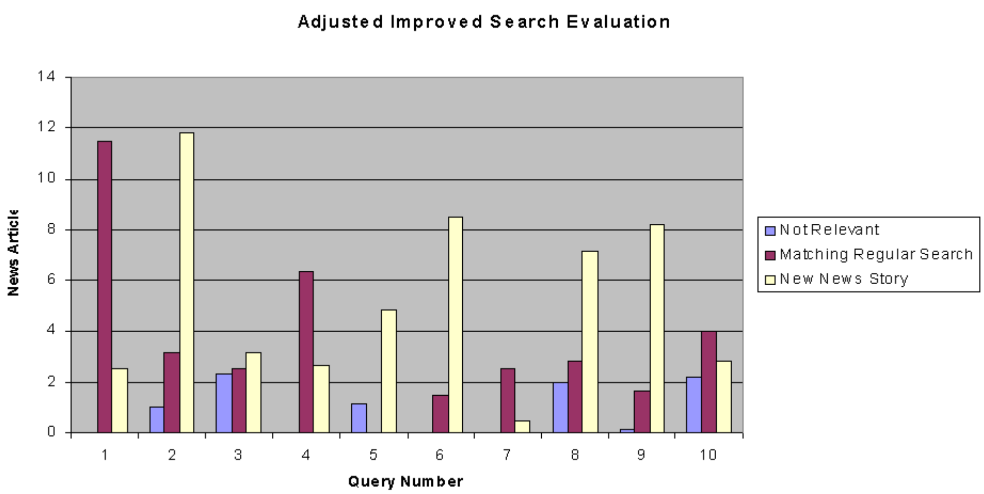

Figure 7.5: Adjusted evaluation results of the improved search mode

Figure 7.4 shows the results of the improved search mode evaluated against the regular search approach. The results show that in a little of over half the queries, the improved system performs well and gives relevant 'side stories' to the query. For a few queries the system does little more than duplicate the news articles of regular search.

Figure 7.5 shows the evaluation results where the result set of the duplicate main node search in the improved approach is removed. As seen in figure 6.7 the improved search mode does a search with the main node as a query followed by adding a term or phrase matching its children. Thus the first results of the improved search mode is exactly the same as the first results in regular search.

The adjusted graph gives a different overview of the evaluation results of the improved mode. The improved approach is good at suggesting 'side stories' that the original query would normally miss. By broadening the search the improved mode does contain some noise, but this is fairly low. The improved mode produces very similar results to the regular search for some queries, e.g query 1, 4 and 7. These are typical topics that have been focused on in media at one particular event and is little known for anything else. Antonella Barba's picture scandal, Mega Million's grand prize and the Grand Canyon skywalk. A complete listing of the answers from the evaluation group can be found in appendix B.2.

7.3 Evaluation Summary

After completing the evaluation there were some notable observations:

- The tree navigation structure help was found to be slightly better than medium in average. This means that the evaluation group found the navigation help partly useful and the overview of the query as partly useful. For an initial approach this shows promising results, with different improvements this could provide to be of good value to the users.
- Both the regular search and the improved search approach show similar ability to be able to find unique news stories. This can either be related to the way the improved mode uses the related term tree structures or that finding 'side story' topics have little effect on the number of unique news stories.
- The evaluation of the improved search mode proves that it is good at finding additional related news articles compared to regular search. For a majority of the queries, improved mode finds a considerable amount of 'side story' articles. In the cases where the phrase in the evaluation query has been focused on in solely one news event, the results are poor.

The evaluation of the prototype improved search application shows promising results. The navigation tree structure was found to give medium help to the user and can be improved additionally. The number of unique news stories were similar for both regular and improved. This can either be an implementation issue or simply that a system that broadens the search for more 'side topics', does not improve the amount of unique news stories. The evaluation of the improved search mode shows that it is good at suggesting additional news stories, but it falls through and produces similar results to regular search, when the news focus has been very one sided.

Chapter 8

Discussion

In this chapter we discuss the different design choices that were made and possible improvements to them. What parameter settings did we choose to get the best results and why. The discussion is in relation to the approach, implementation and evaluation chapters.

8.1 Segmentation

The segmentation algorithm gave overall very good results. As seen in table 6.1 it performs well in finding the natural parts of the strings and to give them an appropriate score. Though one weakness was found, which was apparent in the related term tree structures that was created. The problem is when one have a phrase with a low frequency and one term within the phrase has a much higher relative frequency. The algorithm would then split the high frequency term and the remaining phrase in two.

An example is the string 'antonella barba smoking gun'. First the segment 'antonella barba' is found, but then 'smoking' and 'gun' is split into two segments. This is because smoking has been frequently used in the connection of teenagers smoking. To counteract this effect a measure of 'uniqueness' could be calculated for single terms. Terms that typically are in many different phrases would be given less weight.

8.2 Session Group Analysis

The session group analysis is centered on extracting user unique sessions and finding the most popular term within the session. The queries that do not contain the most frequent term in the group are discarded. If there are more than one term with the highest frequency the first term in the query is chosen. This seems a fair choice as most people write the most prominent words first in a query.

This approach gives good results and give many good describing queries for the group term. Session groups with a similar label are merged and this produces quite large session groups, maybe too large. The simplest approach would be to remove the low frequency queries. This would be particularly important with larger data volumes.

The session analysis has one major weakness. If a user in the same session search for e.g. ‘jennifer lopez’ and ‘jennifer anniston’. Totally different queries that use a common term, can result in a session group with the label ‘jennifer’ and the two queries as its content. This problem has no easy solution when the group label is too general and gives little indication of what the group should contain, other than people that have Jennifer in their names. One could analyze the session groups and remove groups without a clear subject.

8.3 Temporal Correlation Calculation

Temporal correlation gives a good indication if two queries are semantically related or often appear in the same situation. The drawback with using correlation is that it needs good support in the data material to perform well. For instance calculating the correlation for a query with a total frequency of 5 over a 3 day time period would be futile. To give meaningful results the temporal correlation needs few 0 frequency occurrences in the time interval.

With the fairly low query frequencies found in only 3 days of Yahoo! news query logs using temporal correlation is not optimal. Better results would have been found with a longer time interval to get a more extensive data material. To get the best results from the logs, the correlations were calculated over the 3 days available with a 5 hour long time unit.

8.4 Related Term Group Generation

The process of creating the related term groups was quite straightforward, extracting the overlapping queries of the session and temporal calculation analysis. The prototype implementation was created to be only run once to produce the final results, but it can easily be adapted to an incremental approach. By running the analysis every time new query logs were available, new related term groups could be created and current groups added with new queries.

The process of merging the results of the session and correlation analysis seems to work well. Both analysis have been implemented with the idea of having low threshold values to get all relevant results in mind. Both approaches include a considerable amount of noise. The session groups quality is quite good, but there are occasions with generic words and too general words such as Jennifer. In the temporal correlation analysis the threshold value for frequency similarity is set very low and thus generate mostly noise. As seen in appendix A.1 for the query ‘american idol’. This is of little consequence, as when the results are combined the approach shows its strength. Two independent analysis with considerable amount of noise combined together, gives groups of related terms with very high accuracy.

The groups are mostly news related, but other groups are found as well. Internet community sites are popular even though it is news search. Some generic words like videos, photos, etc form group labels and general words e.g. jennier. The sizes of the related term groups are directly related to the query frequency of the group. Among the top 10 most popular term groups the number of queries in each group is high, while the least frequent groups often only consist of 2 queries. This is typical for news stories not related to a popular person or organization, which often are very one sided.

The tree navigation structure generation is not working optimal. Some duplicates within the tree are formed when the segmentation sequence is unfavorable. In the britney spears group there are queries with the strings ‘pictures bald’ and ‘bald’. With pictures having a much higher frequency than bald it is split into pictures and bald, with bald both as a leafnode of britney spears and prictures

```
britney spears
        pictures
                bald
        bald
```

There is also some nodes in the tree that is misspellings or variations of the same word. A technique such as the Levenstein distance could have been used to find words or phrases that are very close to each other and then keep the most frequent of them.

8.5 Web Application

The web application was implemented as a navigation help for people searching for news articles. The tree navigation structure gives a simple overview of what other people have searched for before and can help them find 'side stories'. It was not a given that the related term groups were to be used for query navigation. For instance a popular use of semantic related terms is query recommendation, where the user is suggested terms or phrases to be added to the query to narrow the search.

The main work in this thesis was finding semantic relations in Yahoo! News query logs. This lead to some unexpected problems when implementing a simple prototype search system. Simply setting up a web application, utilizing the created term groups and connecting it to the prototype Vespa search system, was a naive approach. After implementing the simple search prototype, it was evident that the system needed different adjustments and added functionality to perform optimally.

The first issue was that the news feed added to the prototype Vespa system contained many duplicates. Either the same story was published at different web sites and subcategories or these were minor updates of previous articles. News articles that had a perfect match on both title and abstract were removed, but this had a minor effect. Many duplicates that were minor changes from the previous articles did not get removed among others.

The second issue is how the improved system handle the suggested subgroups of the original query. The way it was implemented was simply to add the name of the tree node to the original query. This gave too little weight to the added term or phrase. If the original query was too dominant compared to the added term or phrase, the result of the query variation was very similar to the original query.

8.6 Evaluation Approach

Evaluating a search application is not an easy task, because the results are linked to the document material and the users themselves. The initial idea was to evaluate the system by evaluating the result set after the measures: not relevant, exact match and 'side story'. The problem with search is that in most cases all results are relevant. And defining a measure to separate the results of the main query from the term tree node's queries provided to be close to impossible. If the news article contains the original query, it is an exact match and what was needed for the article to be defined as a 'side story' was a very personal characteristic.

The idea of evaluation the system after variations of the content of documents returned were discarded. Instead we defined evaluation measures for the number of unique documents and the overlapping of results for improved search versus regular search. This gave well defined quality measures with a less sense of person individuality. The intention with this approach was to get a measure for the usability of the tree navigation structures and the ability of the improved search approach to suggest additional 'side stories' to the original query.

Chapter 9

Conclusion and Further Work

In this thesis we have analyzed raw Yahoo! News search query logs to find semantic relations among queries. Yahoo! News is an interactive news service and it is important to help the user navigate and suggest new topics. This is particular important for news search, which is not a familiar tool for most people.

Before choosing the approach in this thesis, an in-debt theoretical study was undertaken in the fields of data mining and information retrieval. Related work in query log mining has been analyzed and taken into consideration in the forming of our chosen approach.

The log analysis approach suggested in this thesis is based around two main contributions, session analysis and temporal correlation. By extracting user sessions and finding the most frequent term within the session, one gets a vocabulary of queries describing that term. In the session group creation process, similar groups are merged and given a label with the most prominent term or phrase describing the group. Temporal correlation is used to find semantically related queries based on their query frequency variation. Queries that have shown the same popularity changes over time are often strongly related.

Taking the intersection of the results of these two contributions, we get what we call related term groups. These are high accuracy groups of related queries, with a term or phrase as group label. These groups were further processed into tree structures and incorporated in our prototype search system. When a user query match one of the created groups, he is presented with a navigation tree structure. Giving the user the possibility of narrowing down his or hers query down different paths.

9.1 Conclusion

The approach formed in this master thesis seems to function well. The approach of using two independent log analysis techniques shows its strength when the results are combined. On its own the session analysis gives good results, but it is prone for noise and often too general words form the groups. The correlation analysis need strict parameter settings to give high accuracy, but this results in most of the data material being rejected. By intersecting the queries found from both approaches, we get a two-way confirmation that the queries are semantically related. False positives from one approach are seldom found within the other one and are removed by combining the results.

The quality of the groups seem to be good in general, and the accuracy within each group is very high. Some of the weaker groups are often formed around one general term. These can probably be removed by analyzing the content of each group, checking if they form around one specific subject within the group. The quantity of the related term groups are directly connected to the amount of data material. With only 3 days of query logs from Yahoo! News search, we believe our analysis is somewhat hindered by the data material. Given a longer time period, the analysis would yield more and larger related term groups generated from the analysis. A complete listing of the groups can be found in appendix A.2;

There are many possibilities for the further use of the related term groups. Either as a part of a larger analysis or implemented as an alternative user navigation and suggestion tool. To generate tree structures for navigation help is just one of many uses. Thus the performed evaluation is concerned directly with the result of one possible prototype implementation and only indirectly the quality of the generated related term groups.

The results from the evaluation of the prototype search application show fairly good results. The navigation help and overview from the tree structures were valued to be slightly above medium. The generation of tree structures can easily be improved with some adjustments and added functionality. The improved search mode shows a good ability to suggest many 'side stories' from the original query, compared to regular search. Though a number of unique news stories show similar results for both the regular approach and the improved. This could either be an implementation issue or the nature of news article search.

9.2 Further Work

There are two main areas of interest concerning further work. The first is the query log analysis approach, which is the main body of this thesis. The second is the prototype search application, implemented to provide an example application using the log analysis results.

Improvements to the query log analysis:

- Perform the analysis on query logs or over a longer duration of time. The query log analysis is somewhat limited with only 3 days of raw Yahoo! News search query logs. Though this would require some modifications to the log analyzer code for better processor and memory usage.
- Different improvements can be added to the contributions forming the approach in this thesis. The segmentation could incorporate negative weights for general words, a filtering of queries in the session groups and performance improvements to the temporal correlation calculations.
- Some of the groups generated are based on a very general term, resulting in different subgroups. By analyzing the group content, groups not centered on one subject could be removed.
- Adapt the log analyzer to work in a sequential manner. When new query logs are available, the analyzer could be run to create new groups and add queries to the existing ones. This way the related term groups can easily be kept up-to-date.

Improvements to the prototype search application:

- The tree node navigation structure generation is not optimal as discussed in chapter 8. Duplicates within the tree and variations of the same term should be removed.
- Running queries against the Vespa news search system from Yahoo! returns many duplicates. Better functionality to remove duplicates should be incorporated.
- The term or phrases that are added to the original query to create 'side stories' should be weighted in some manner.

There is also the possibility of using the related term groups for other purposes. Either as a part of a larger data analysis or to use the groups in a different approach in a search application. Such as incorporating the document collection in the log analysis or implement a different prototype application such as a query recommendation system.

Bibliography

[1] Y. News, Dec. 2006. http://news.yahoo.com/.

[2] C. S. W. Verne Kopytoff, "New search engines narrowing their focus," 2005.

[3] B. R.-N. Ricardo Baeza-Yates, *Modern Information Retrieval*. Addison-Wesley, 1999.

[4] S. Barshinger, May 2006. The Emerging Opportunity in Vertical Search.

[5] C. K. Base, Dec. 2006. http://www.coveo.com/en/Support/articles/Information%20-%20CES4-060330-3%20-%20Understanding%20Stemming.htm.

[6] D. M. Joel Plisson, Nada Lavrac, "A rule based approach to word lemmatization,"

[7] H. M. M. M. C. Silverstein, M. Henzinger, "Analysis of a very large altavista query log," 1998.

[8] N. I. Steve Chien, "Semantic similarity between search engine queries using temporal correlation,"

[9] M. K. Jiawei Han, *Data Mining - Concepts and Techniques*. Morgan Kaufmann, 2006.

[10] P. S. S. Srivatsan Laxman, "A survey of temporal data mining," *Sadhana*, vol. 31, pp. 173–198, April 2006.

[11] A. L. O. Claudia M. Antunes, "Temporal data mining: an overview,"

[12] G. Dupret, 2006. Web Mining: a review of some applications.

[13] G. News, 2007. http://news.google.com/.

[14] R. Baeza-Yates, "Applications of web query mining," 2005.

[15] B. J. J. Amanda Spink, Dietmar Wolfram and T. Saracevic, "Searching the web," 2001.

[16] H. C. O. Amanda Spink, Seda Ozmutlu and B. J. Jansen, "U.s. versus european web searching trends," 2002.

[17] Z. M. F. Saraiva, de Moura and Ribeiro-Neto, "Rank-preserving two-level caching for scalable search engines," 2001.

[18] Q. S. for Web Search, 2003. http://www2003.org/cdrom/papers/poster/p052/xhtml/querysegmentation.html.

[19] O. Medelyan, "Why not use query logs as corpora?," 2004.

[20] WordNet, May 2007. http://wordnet.princeton.edu/.

[21] J.-S. C. Key-Yih Su, Ming-Wen Wu, "A corpus-based approach to automatic compound extraction,"

[22] M. Z. Jian Zhang, Jianfeng Gao, "Extraction of chinese compound words,"

[23] H. M. Kawamae Noriaki, Mukaigaito Takeya, "Semantic log analysis based on a user query behavior model," 2003.

[24] N. Seco and N. Cardoso, "Detecting user sessions in the tumba! query log," 2006.

[25] A. C. Eric C. Jensen, Seven M. Beitzel and O. Frieder, "Query phrase suggestion from topically tagged session logs," 2006.

[26] N. I. Steve Chien, "Semantic similarity between search engine queries using temporal correlation," May 2005.

[27] S. Maslow, Golovko and Braslavski, "Extracting news-related queries from web query log," May 2006.

[28] C. G. Beitzel, Jensen and Frieder, "Hourly analysis of a very large topically categorized web query log," 2004.

[29] C.-C. H. Hahn-Ming Lee and C.-Y. Chao, "Association thesaurus construction for interactive query expansion based on assoiation rule mining," 2004.

[30] J.-Y. N. Ji-Rong Wen and H.-J. Zhang, "Clustering user queries of a search engine," May 2001.

[31] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log,"

[32] C. H. Ricardo Baeza-Yates and M. Mendoza, "Mining search engines logs: Clustering query traces," 2005.

[33] S. M. Beitzel and D. D. Lewis, "Improving automatic query classification via semi-supervised learning,"

[34] S. J. P. W. Y. Shen, Pan and Yang, "Query enrichment for web-query classification,"

[35] Wikipedia, 2007. http://en.wikipedia.org/wiki/RSS_(file_format).

[36] Jetty, 2007. http://www.mortbay.org/.

[37] S. J. 1.4.2, 2007. http://java.sun.com/j2se/1.4.2/docs/api/.

[38] S. J. 1.6.0, 2007. http://java.sun.com/javase/6/docs/api/.

Part V

Appendices

Appendix A

Query Log Analysis Results

A.1 Temporal Correlation Group: American Idol

```
american idol
1.0000001       american idol
0.9659458       american idle
0.95593166       idol
0.9530445       american idol.com
0.95264685       american idol fox
0.9489394       fox tv
0.93390244       american idol antonella barba photos
0.9063261       american idol 6
0.90311986       americanidol.com
0.8759125       american idol 2007
0.86926955       american idol news
0.84978884       american idol contestants
0.8394189       america idol
0.838453       american idol antonella barba photos blog
0.82663584       chinese
0.8069563       spanish
0.80672234       american idol results 2007
0.8005493       grand canyon skywalk
0.7924204       fox
0.7923774       bluffton bus crash
0.7848175       van halen
0.78386694       grand canyon
0.77613074       polar bears
0.7736698       bush brazil
0.77175504       canada
0.7707447       sri lanka
0.77028906       panama city
0.7651176       duke
0.7639794       american idol results
0.7544426       you tube videos
0.7538369       ebay auctions
0.7522186       leonardo dicaprio
0.7444869       carrie underwood
0.742336       agriculture
0.7369451       ebay website usa
0.7345353       pokey chapman
0.7331871       sopranos
0.724866       antonella barba smoking gun
0.7235123       antonella barba american idol
```

```
0.717852        ebay website usa auctions
0.7107933        antonella barba nude photos
0.7107318        nas
0.70944977        fox news channel
0.7078475        ethanol
0.702109        antonella barba racy photos
0.6972331        forbes
0.6913597        star jones weight loss
0.6877319        jimmy carter
0.68645227        bush
0.68261105        chicago weather
0.6806665        bahamas
0.6788887        hotmail
0.67526025        sanjaya malakar
0.66581285        ebay
0.66136104        acc
0.6612433        panama
0.6603632        ebay website
0.65637505        star jones
0.6526355        mcdonalds
0.6491786        alcatel lucent
0.6459486        antonella
0.64546126        newspapers
0.6440501        bush protest
0.6438835        vonage
0.64223844        antonella barba pics
0.6416299        big 12 tournament 2007
0.6405483        la la vasquez
0.63526726        jessica lunsford
0.6278284        sundance head
0.6247783        algeria
0.62343025        barba photos
0.6206921        iditarod 2007
0.6176571        view american idol antonella barba pictures
0.61539954        50 cent
0.61519116        bill gates
0.6141849        basketball
0.61371577        brazil
0.6120568        atlanta bus crash
0.60851544        eddie van halen
0.6056542        airtran
0.60495806        new century
0.60442764        ebay motors
0.6028442        charleston sc
0.5930633        american idol antonella barba pictures
0.5887108        laura bush
0.5872938        lil wayne
0.58603424        antonella barba photos
0.5805769        acc tournament
0.5791877        antonella barba photo
0.5748706        iditarod
0.5714735        world news
0.5711438        smoking gun
0.5697232        big 12
0.56642216        oprah
0.55902237        miami fl
0.55879605        porn
0.5489611        antonella barba
0.54756707        yahoo finance
0.54552525        msn home page hotmail
0.5455147        map quest
0.54317003        katie couric
0.5428065        sec basketball tournament
0.540257        europe
```

```
0.5402383        britney spears pictures bald
0.5398666        2007 acc tournament
0.53963083        car
0.5345999        president bush
0.5325229        american idol antonella barba pics
0.5311338        george w bush
0.52950406        uganda
0.5262617        ronaldo
0.5235531        toronto
0.5228266        energy
0.52010614        san antonio tx
0.51767766        oakland raiders
0.5176561        coffee
0.51326686        pokey chatman
0.5125246        miss usa
0.51157796        oil prices
0.51113015        movies
0.5098021        dear abby
0.50883514        hawaii
0.50716347        antonella barba pictures
0.50707525        new york yankees
0.50456077        daytona
0.50373        2007 big ten tournament
Number of correlating queries found        126
```

A.2 Related Term Groups

Complete list of related term groups found over a 70 hour interval with 5 hour time units. The values have been normalized and multiplied with 1000 for easy reading.

```
group name:         google search
1000.0        google
808.2981        google search
369.50662        google search web site
278.7495        google home page
174.39929        google search engine
133.76447        google search web site home page
56.24855        google.com
54.915043        google home page search
34.986774        google search engine home page
30.107813        google home page site
21.566284        www.google.com
18.68564        google search web
14.998318        google search web site home page car insurance
11.893635        google home page search earth
7.380858        google search engine home page web site
4.643651        google .com
2.659129        google maps
2.0109901        google.
1.5736659        people search

group name:         antonella barba
431.2776        antonella barba
407.90707        antonella barba pictures
183.75435        american idol antonella barba pictures
60.060734        antonella barba smoking gun
39.898777        antonella barba american idol
38.2621        antonella barba photos
22.144068        american idol antonella barba photos
17.727404        antonella barba pics
13.226929        barba
11.287567        antonella barba nude
11.164677        view american idol antonella barba pictures
8.5704        antonella barba photo
6.19937        american idol antonella barba pics
6.179472        antonella barba nude pictures
6.1684456        pictures of antonella barba
4.73297        antonella barba nude photos
4.024619        barba photos
3.8632922        american idol antonella barba photos blog
3.5830274        antonelle barba
1.5240321        barba pictures
1.4774853        antonella barba racy photos

group name:         american idol
163.49951        american idol
104.451256        american idol antonella barba photos
11.5966215        american idol.com
8.986275        american idol results
7.665904        american idol antonella barba photos blog
6.184282        idol
5.918117        antonella barba american idol
4.6520057        american idol contestants
3.2181296        american idol results 2007
2.5725424        american idol 2007
2.0370095        americanidol.com
1.8005396        america idol
1.5643495        american idle
```

```
1.467066        american idol antonella barba pictures
1.4594592        american idol 6
1.4475055        american idol news
0.39518747        american idol fox

group name:        ebay website
150.1971        ebay
95.575806        ebay website
41.77916        ebay website usa
27.063017        ebay auctions
7.186785        ebay website usa auctions
4.9416714        ebay motors
2.1488822        ebay.com
1.9164755        ebay website uk

group name:        youtube
118.90648        youtube
79.71727        youtube videos
44.558064        youtube website
28.308207        youtube videos web site
9.635554        youtube.com
3.3877447        youtube website homepage
3.1409218        youtube website home
2.7107148        youtube videos web site paris hilton

group name:        britney spears
107.355125        britney spears
64.68107        britney spears pictures
23.611563        britney spears bald
15.393676        britney spears pictures bald
4.6440997        brittany spears
3.939804        britney
2.9931712        britney spears bald pictures
2.472464        britney spears new
2.4044056        britany spears
1.3546494        brittney spears
0.9685643        britney spears news

group name:        mega millions
106.87777        mega millions
60.643913        mega millions lottery
41.20423        mega millions lottery numbers
11.019735        mega millions results
6.4432397        mega million
2.2014823        mega millions lottery numbers new jersey
0.8844629        megamillions
0.6167305        mega lottery

group name:        baseball
70.43048        baseball
55.489403        baseball minor league
50.502174        baseball team

group name:        home page
65.8175        google home page
39.036846        google home page search
25.393309        google home page site
9.054762        google home page search earth
2.4421086        google search web site home page
2.3189611        myspace home page
2.0197182        myspace home page login
1.916476        google home page site parker
1.5407228        msn home page
1.2552664        home depot
```

```
1.0463856        msn home page hotmail
0.47911888        google search engine home page

group name:         immigration
40.70849        immigration
31.736422        immigration bill
12.936212        immigration news
5.237757        president bush immigration bill
4.7811365        immigration reform
3.3538332        us immigration
2.9934645        new immigration bill
1.3246049        illegal immigration

group name:         grand canyon
64.7888        grand canyon
59.889874        grand canyon skywalk

group name:         president bush
28.504625        laura bush
22.99771        president bush
19.03831        bush
5.8848557        jenna bush
5.1799893        george bush
3.1438637        bush brazil
2.9089181        george w. bush
2.8690908        bush daughter
1.4373568        president bush immigration bill
1.3292998        george w bush
0.74812984        president george w. bush

group name:         tournament
11.26287        big ten tournament brackets
9.045923        2007 acc tournament
7.871152        big ten tournament
7.5123315        sec tournament
4.473619        big 12 tournament schedule
4.31207        acc basketball tournament
3.6688497        acc tournament
3.5401716        2007 big ten tournament
3.5169353        big east tournament schedule
3.198155        acc tournament brackets
2.941215        big east tournament
2.6505485        big 12 tournament 2007
2.3447387        big 10 tournament
1.916476        big east tournament 2007
1.7984629        sec basketball tournament
1.5843209        big ten basketball tournament
1.4719816        big ten tournament 2007
1.400909        big 12 tournament
1.2623652        2007 sec tournament
0.7167298        big 12 basketball tournament

group name:         photos
27.97673        american idol antonella barba photos
24.534546        antonella barba photos
2.8747134        barba photos
2.5250921        antonella barba nude photos
1.4702052        antonella barba racy photos

group name:         wikipedia
31.107399        wikipedia
28.019405        wikipedia encyclopedia
13.41533        wikipedia encyclopedia free
```

```
group name:        anna nicole
29.895329       anna nicole smith
20.471254       anna nicole
5.2406955       anna nicole cause of death
4.312071       anna nicole smith death
1.2444454       anna nicole smith cause of death
0.24538848       nicole kidman

group name:        videos
25.160503       youtube videos
24.480835       youtube videos web site
2.3955946       you tube videos
0.95600414       videos
0.24312767       music videos

group name:        wal mart
24.435068       wal mart online
18.65618       walmart
14.130391       wal mart
6.106189       wal mart online store

group name:        news
8.536193       cnn news
5.450132       news
4.217883       odd news
3.832952       world news
3.738131       cnn news headlines
3.586852       bbc news
3.3538332       local news
3.2586074       yahoo odd news
2.764491       immigration news
2.395595       fox news
2.3955946       philippine news
1.916476       abc news
1.9164757       science news
0.95823777       yahoo news
0.3134242       britney spears news

group name:        lottery
16.769165       lottery
12.697678       mega millions lottery
8.026861       mega millions lottery numbers
2.7157753       california lottery
2.4832015       nj lottery
2.4783309       ohio lottery
1.780262       texas lottery
1.578114       illinois lottery
1.5708534       mass lottery
0.958238       lottery winners
0.9582378       lottery numbers
0.32754585       mega millions lottery numbers new jersey
0.2529202       mega lottery

group name:        beyonce
14.541309       beyonce knowles
14.370169       beyonce
6.2277284       beyonce jay z
4.312071       beyonce shakira

group name:        iran
38.329514       iran
8.62414       iran israel
4.312071       iran war
4.2283044       iranian
```

```
1.4373568        iran nuclear

group name:         real time traffic
9.1887045        predictive traffic
9.009565        maptuit traffic
8.768457        cobra traffic
8.722928        traffic.com
8.1450205        real time traffic probe vehicle
7.6212473        msn direct traffic

group name:         myspace
27.788904        myspace
13.10606        myspace home page
2.7764041        myspace home page login
0.6292646        myspace.com
0.4233133        www.myspace.com

group name:         weather forecast
16.770437        weather
6.2285457        las vegas weather forecast
5.226881        las vegas weather
5.1905446        weather channel forecast
0.6097921        washington dc weather forecast

group name:         las vegas
10.828395        las vegas
9.967692        las vegas weather
3.8330462        las vegas weather forecast
3.832953        las vegas nv
0.47911888        las vegas nevada

group name:         south park
29.70538        south park
14.030072        south park n word

group name:         cnn
21.560352        cnn
12.424637        cnn news
4.8496857        cnn news headlines
1.1796829        cnn.com

group name:         tube
21.560354        you tube
7.6544485        youtube
2.6136744        youtube videos
1.7098889        youtube.com
0.6127082        youtube videos web site
0.31369278        youtube website
0.31118265        utube

group name:         china
14.2242365        china
4.767217        china military
3.832952        china india

group name:         msn
22.03947        msn
9.145124        msn home page
1.6580516        msn.com

group name:         antonella
8.189241        antonella barba
8.145023        antonella
7.5251913        antonella barba pictures
```

```
1.6420653        antonella barba smoking gun
0.7903613        american idol antonella barba pictures
0.41523874        american idol antonella barba pics
0.35558853        american idol antonella barba photos
0.3169704        antonella barba photo
0.2710136        view american idol antonella barba pictures

group name:         mapquest
18.685637        mapquest
8.652107        mapquest driving directions
3.0145514        yahoo mapquest driving directions

group name:         green tea
23.955948        green tea
8.145021        green tea benefits

group name:         pokey chatman
26.351543        pokey chatman
2.3481677        pokey chapman

group name:         brain
12.936212        brain man
6.569808        brain man 60 minutes
3.5669563        brain man 60 minutes daniel
2.874714        brain

group name:         immigration bill
12.457093        immigration bill
7.665904        new immigration bill
2.586572        president bush immigration bill
1.8447989        bill clinton
1.1895579        bill gates

group name:         60 minutes
5.3256006        60 minutes
4.491577        brain man 60 minutes
3.3538327        savant 60 minutes
3.1046464        brain man 60 minutes daniel

group name:         barack obama
15.331804        barack obama
6.6949825        obama

group name:         johnny depp
15.676213        johnny depp
5.2703085        johnny depp daughter
0.86539394        depp

group name:         hillary clinton
10.061497        hillary clinton
4.2732234        clinton
3.672789        bill clinton

group name:         myrtle
7.6659026        myrtle beach weather forecast
7.439583        myrtle beach
3.832952        myrtle beach sc

group name:         daylight savings
10.061499        daylight savings
4.4667635        daylight savings time
2.515009        daylight savings time change
1.9755734        daylight saving
1.2016644        day light savings
```

```
group name:        maps
6.3823953       yahoo maps
6.097653       maps
3.3538337       google maps

group name:        jennifer
8.044158       jennifer hudson
3.832952       jennifer aniston
2.2291281       jennifer lopez

group name:        orlando
10.061499       orlando
2.8680327       orlando fl
2.5557024       orlando florida weather
0.77547896       orlando bloom
0.27513468       orlando florida

group name:        games
11.977974       games
4.976691       yahoo games

group name:        time
4.882131       daylight savings time
4.1828365       daylight saving time
2.4317112       daylight savings time change
2.3955946       time
1.9164757       time warner cable
0.958238       day light savings time

group name:        top model
13.415333       america s next top model
4.374102       americas next top model
1.9164758       america next top model

group name:        basketball
2.4090948       ncaa basketball
2.3955948       basketball
2.235979       college basketball
1.916476       ohio state basketball
1.4373568       syracuse basketball
1.2544851       florida basketball
1.1396952       sec basketball tournament
0.6485041       big 12 basketball tournament
0.40301862       acc basketball tournament
0.3670614       big ten basketball tournament

group name:        fox
8.624141       fox news channel
3.6328022       fox tv
2.0640259       fox
0.55965793       american idol fox

group name:        hotmail
11.4988575       hotmail
4.2799077       msn home page hotmail

group name:        florida
4.7911897       orlando florida weather
4.593603       florida
4.312071       orlando florida
0.47911906       florida basketball

group name:        daniel
```

```
10.061498        daniel radcliffe
7.665904        daniel radcliffe equus

group name:         iditarod
7.665903        iditarod
6.228546        iditarod 2007
3.832952        iditarod race

group name:         music
4.3753557        music
4.31207       yahoo music
2.6981494        music videos

group name:         chicago
11.977976        chicago
5.2703094        chicago weather

group name:         driving directions
4.3120713        driving directions
3.9554458        mapquest driving directions
2.9217012        yahoo mapquest driving directions

group name:         elizabeth hurley
5.538761        elizabeth hurley
3.3538322        elizabeth hurley wedding
1.6232505        hurley
1.511069        liz hurley

group name:         espn
8.145021        espn
5.4938693        espn sports

group name:         rosie o
7.212161        rosie odonnell
2.395595        rosie o donnell
1.4124893        rosie

group name:         angelina jolie
10.061498        angelina jolie
1.806226        jolie
0.9582379        brad pitt angelina jolie

group name:         grey s anatomy
9.103259        grey s anatomy
4.212567        greys anatomy

group name:         in iraq
5.246471        iraq
3.8329525        war in iraq

group name:         san
5.27031       san antonio
4.3120704        san francisco
1.4034945        san juan
0.797732        san antonio tx
0.4791189        san diego

group name:         2davweb
4.31207       boeing 2davweb
3.4035814        airbus 2davweb
3.1013134        eads 2davweb

group name:         jones
7.11491       star jones
```

```
2.8747144        star jones weight loss
1.916476        star jones reynolds

group name:         exploration
6.2285476        diamond exploration
5.078663        uranium exploration

group name:         giant squid
8.1450205        giant squid
2.1521256        squid
1.4951763        colossal squid

group name:         new york
7.186785        new york city
3.8329525        new york
0.479119        new york yankees
0.346709        new york times
0.26702514        new york giants

group name:         dear
5.749428        dear margo
4.044871        dear abby

group name:         bbc news
4.9580765        bbc
4.791189        bbc news

group name:         verizon
7.6659026        verizon
2.363098        verizon wireless

group name:         finance
6.7076654        yahoo finance
4.79119        finance

group name:         boeing
8.145023        boeing
1.9855306        boeing 787

group name:         walter reed
8.145022        walter reed
3.3538322        walter reed hospital

group name:         liberal arts
4.1628985        liberal arts
3.3538327        liberal arts degree

group name:         ncaa basketball
4.312071        ncaa basketball
3.9676282        ncaa

group name:         dictionary
7.186784        dictionary
2.4676356        online dictionary

group name:         coulter
5.156809        ann coulter
3.8329513        coulter
1.043092        anne coulter

group name:         smoking
3.8329515        smoking gun
3.8329515        smoking
1.4553651        antonella barba smoking gun
```

```
group name:        brazil
7.5283923       brazil
1.4373572       bush brazil

group name:        spiderman 3
7.1867843       spiderman 3
3.2215168       spiderman

group name:        airlines
3.8329515       american airlines
1.916476       delta airlines
1.4373568       united airlines
1.4241333       northwest airlines
0.9582379       continental airlines

group name:        ray
4.420139       rachel ray
4.312071       rachael ray

group name:        lotto
5.006502       lotto
2.8747137       california lotto
1.1891958       texas lotto

group name:        frenchie
5.13357       frenchie davis
3.3538322       frenchie davis pics

group name:        van halen
5.270309       eddie van halen
3.3850946       van halen

group name:        savant
5.749427       savant
3.2281785       savant 60 minutes

group name:        boston
6.7076664       boston
2.3955948       boston ma

group name:        scooter libby
4.312069       scooter libby
3.7208848       libby

group name:        acc tournament
2.2911148       acc
1.6835912       acc tourney
1.4373565       2007 acc tournament
1.0792238       acc tournament brackets
0.8770705       acc tournament

group name:        nigeria
3.0232866       nigeria
2.8747137       nigeria news

group name:        haiti
6.228546       haiti
2.395595       haiti news

group name:        jet man
7.1867843       jet man
0.645008       jet pack
```

```
group name:        south
1.916476       south africa
1.9164758       south padre island
1.9164758       south korea
0.74468386       south park
0.71097046       south park n word
0.47911906       south carolina

group name:        mexico
6.2285476       mexico
1.9164757       cancun mexico

group name:        italy
5.7494287       italy
1.9164757       rome italy

group name:        lebanon
5.270309       lebanon
1.7108986       beirut lebanon

group name:        yahoo
2.874714       yahoo
1.9164757       yahoo mail
1.4373567       yahoo at
0.958238       yahoo maps
0.25787416       yahoo news

group name:        prince
2.3955948       prince charles
1.9164758       prince
1.5691917       prince william
0.34789535       princess

group name:        atkins diet
4.3120704       atkins diet
2.1731339       atkins

group name:        sex
3.3538327       sex
1.4373572       teen sex
1.437357       sex offender
0.9582381       teacher sex

group name:        gas
3.3538325       gas
2.3955946       gas prices
1.4373568       natural gas

group name:        jay z
4.7911887       jay z
1.3709472       beyonce jay z

group name:        paris
4.312071       paris
2.8747134       paris france

group name:        indonesia
3.3350239       indonesia
2.3955946       indonesia earthquake

group name:        oil
3.785936       oil
1.9164757       oil prices
0.25326008       crude oil
```

```
group name:         bay
2.8747137        e bay
1.3126535        ebay
0.9582379        green bay packers
0.26964843        ebay website

group name:         fema
5.2703094        fema
0.96709585        fema trailers

group name:         harry
5.7494287        harry potter
0.958238        prince harry

group name:         people
3.832952        people
1.5325315        people magazine
0.47911894        people search

group name:         big 12
3.4480627        big 12
1.4373568        big 12 tournament 2007
0.6969068        big 12 tournament

group name:         mets
2.395595        mets
1.9164762        david wright mets
1.9164757        new york mets

group name:         nowak
4.373598        lisa nowak
1.4373567        nowak

group name:         smith
3.7098312        anna nicole smith
1.916476        anna nichole smith

group name:         philadelphia
4.312071        philadelphia
1.916476        philadelphia pa

group name:         sports
3.8329506        sports
2.395595        espn sports

group name:         stephen grant
4.79119        stephen grant
0.7584248        tara grant

group name:         bus
1.0650465        atlanta bus crash
1.0554582        bluffton bus crash
1.015376        bus crash
0.958238        bus accident
0.2975913        airbus

group name:         new orleans
4.7911897        new orleans
0.9582379        new orleans saints

group name:         de
2.395595        siebel or peoplesoft or grilles de calcul
1.3407345        rsa or novell or sun microsystems identity
```

```
0.9582378        oscar de la hoya

group name:         nfl
2.8747144        nfl combine
2.3955946        nfl
0.4791189        nfl draft

group name:         uranium
3.3538325        uranium
1.916476        uranium mining
0.3014391        uranium exploration

group name:         john edwards
3.3538332        john edwards
1.5165716        john mccain

group name:         gay
3.832951        gay
1.1756845        gay marriage

group name:         kobe bryant
4.312071        kobe bryant
1.0380743        kobe

group name:         digital
2.8747144        digital photography
2.8747137        digital camera

group name:         teacher
5.2703094        teacher
0.25213298        teacher sex

group name:         texas
3.3538322        texas youth commission
1.916476        texas

group name:         nintendo wii
3.0549028        wii
1.437357        nintendo wii

group name:         simpson
4.31207        jessica simpson
0.6570289        oj simpson

group name:         duke
2.874714        duke basketball
1.5856843        duke

group name:         james
3.3538322        james brown
1.9164758        lebron james

group name:         big ten
2.6828563        big ten tourney
0.9196384        big ten tournament
0.479119        big ten basketball tournament
0.44900823        big ten tournament brackets
0.3936059        2007 big ten tournament

group name:         navigation system
2.3955946        navigation system
2.1899157        dash navigation

group name:         cheney
```

```
3.8329515        cheney
0.74695563        dick cheney

group name:         illinois
1.9772232        illinois lottery
1.4373572        illinois

group name:         hilton
4.31207        paris hilton
0.37456015        youtube videos web site paris hilton

group name:         aol
2.3955948        aol
1.4189944        aol.com

group name:         tom cruise
2.3955946        tom cruise
0.9523264        cruise
0.37829006        tom brady

group name:         breast cancer
3.3538332        breast cancer
0.5368831        lung cancer

group name:         bluffton
2.395595        bluffton
1.198862        bluffton university
0.47911894        bluffton bus crash

group name:         crime
2.3955948        crime
1.0428481        organized crime

group name:         mega million
3.8329515        mega million
0.40764925        megamillions

group name:         airtran
1.916476        airtran airways
1.378948        airtran

group name:         barcelona
2.3955948        fc barcelona
1.9164758        barcelona

group name:         houston
2.395595        houston
1.4373567        houston tx

group name:         daylight saving
1.9164758        daylight saving
1.3421676        daylight saving time

group name:         hugo chavez
2.2205844        chavez
1.4373568        hugo chavez

group name:         mcdonald s
1.916476        mcdonald s
1.3538296        mcdonalds

group name:         miami
2.874714        miami
0.9582379        miami fl
```

```
group name:         credit card
2.395595        credit card
0.90187645        credit cards

group name:         50 cent
1.9164758        rapper 50 cent
1.9164755        50 cent

group name:         saudi
1.916476        saudi
1.9164757        saudi arabia

group name:         yellow pages
1.8318566        white pages
0.958238        yellow pages

group name:         spider man
1.4373572        spider man
1.1671326        spider man 3

group name:         nbc
2.8747144        nbc
0.24703541        msnbc

group name:         panama
1.916476        panama city
1.4373572        panama

group name:         syracuse
2.395595        syracuse
0.49576554        syracuse basketball

group name:         dallas
2.395595        dallas
0.9582378        dallas tx

group name:         real estate
2.8747146        real estate
0.31489465        real time traffic probe vehicle

group name:         gallo
1.4373572        gallo
1.295549        ernest gallo

group name:         nashville
2.3955944        nashville star
0.47911906        nashville tn

group name:         michael
2.3955948        michael jackson
0.47911906        michael jordan

group name:         dolphins
1.9164757        miami dolphins
0.9582379        dolphins

group name:         sec tournament
1.4197975        sec
0.9582379        sec tournament

group name:         man
0.9390251        jet man
0.479119        smartest man
```

```
0.3437214        spider man 3

group name:         odd news
1.4373567        odd news
0.655057        yahoo odd news

group name:         google.com
1.4373572        google.com
0.83506125        www.google.com

group name:         airbus
1.437357        airbus
0.9582379        boeing airbus

group name:         big east
1.916476        big east tournament
0.4791189        big east tournament schedule

group name:         new century
1.354878        new century
0.47911894        new century mortgage

group name:         washington
1.437357        washington dc
0.2604501        washington dc weather forecast

group name:         myspace.com
0.47911888        myspace.com
0.2844638        www.myspace.com

Number of correlation groups found: 175
```

A.3 Related Term Tree Navigation Structure

A list of the 50 first tree navigation structures found over a 70 hour interval with 5 hour time units. The values have been normalized and multiplied with 1000 for easy reading.

```
google search        1000.0
  home page        187.28476
    web site        61.542473
      engine        2.4388335
      car        4.9558463
        insurance        4.9558463
    engine        11.56057
    earth        3.929976
  web        128.26915
    site        122.09491
  www        7.126079
  engine        57.6262
  people        0.51998144
  maps        0.8786476
antonella barba        423.94894
  american idol        88.23206
    pictures        64.40649
      view        3.6891084
    photos        8.59353
      blog        1.2765354
    pics        2.048438
  pictures        139.3671
    nude        2.0418632
  photos        16.024769
    nude        1.5639004
    racy        0.48820078
  smoking        19.845678
    gun        19.845678
  nude        3.729715
  pics        5.8576097
  photo        2.83189
  antonelle        1.1839284
american idol        108.68259
  antonella barba        39.486782
    pictures        0.484758
    photos        37.046516
      blog        2.5330205
  news        0.4782946
  2007        1.9133935
    results        1.0633564
  america        0.59494656
  results        2.9693065
  fox        0.13058054
  contestants        1.5371475
  idle        0.51690304
ebay website        109.30813
  usa        16.179663
    auctions        2.3747067
  uk        0.63325495
  auctions        8.942347
  motors        1.6328607
youtube        95.94438
  web site        27.12991
    home        2.1572466
      page        1.1194018
    videos        10.249483
      paris        0.8956928
        hilton        0.8956928
```

```
  videos        26.340725
britney spears        75.938225
  pictures        27.447868
    bald        6.075508
  news        0.32003963
  bald        7.801894
  brittany        1.534535
  brittney        0.44761252
  britany        0.7944801
mega millions        75.96234
  lottery        34.58457
    numbers        14.342412
      jersey        0.72742873
  results        3.641216
  million        2.1290193
baseball        58.29458
  league        18.335188
    minor        18.335188
  team        16.687275
home page        50.331
  google search        47.62776
    site        9.830824
      web        0.80693823
      parker        0.6332551
    engine        0.15831374
    earth        2.9919364
  myspace        1.433616
    login        0.66736907
  msn        0.85485
    hotmail        0.34575388
  depot        0.41477373
immigration        34.057724
  bush        1.7306956
    bill        1.7306956
      president        1.7306956
  news        4.2744713
  bill        11.475687
  illegal        0.43768498
  reform        1.5798157
grand canyon        41.19718
  skywalk        19.78922
president bush        31.074049
  immigration        0.47494125
    bill        0.47494125
  jenna        1.9445143
  george        3.3592305
  brazil        1.038817
  laura        9.418693
  daughter        0.9480245
tournament        25.274288
  big ten        15.0977545
    2007        3.1652172
      east        0.6332551
      12        0.8758123
    basketball        0.7603289
      12        0.23682678
    east        2.1339462
      schedule        1.1620897
    12        1.9411018
      schedule        1.4782038
    10        0.77476466
    brackets        3.7215543
  2007        3.4061348
    sec        0.41711935
```

```
    acc        2.9890156
  basketball        2.019084
    sec        0.59426033
    acc        1.4248236
  acc        2.2690427
    brackets        1.0567563
  sec        2.482276
photos        19.62117
  antonella barba        19.621172
    american        9.244263
      idol        9.244263
    nude        0.8343582
    racy        0.48579523
wikipedia        23.96987
  free        4.432785
    encyclopedia        4.432785
  encyclopedia        9.258364
anna nicole        20.29124
  smith        11.71424
    death        1.8360221
      cause of        0.41119817
  death        1.7316664
    cause of        1.7316664
  kidman        0.08108294
videos        17.590624
  youtube        16.40283
    web        8.089125
      site        8.089125
  music        0.08033591
  tube        0.7915687
wal mart        20.925213
  online        10.09165
    store        2.0176487
news        16.070269
  spears        0.10356377
    britney        0.10356377
  immigration        0.91346204
  yahoo        1.3933588
    odd        1.0767313
  world        1.2665102
  fox        0.7915689
  cnn        4.055766
    headlines        1.2351788
  bbc        1.1851921
  local        1.1081965
  abc        0.6332551
  science        0.633255
  odd        1.3937018
  philippine        0.7915687
lottery        17.379519
  mega millions        7.039746
    numbers        2.7605205
      jersey        0.10822994
  texas        0.58824635
  ohio        0.81890696
  numbers        0.3166275
  california        0.897365
  nj        0.82051635
  illinois        0.5214512
  winners        0.31662756
  mass        0.5190521
beyonce        13.03576
  knowles        4.804839
  jay        2.0578086
```

```
    shakira        1.424824
  iran        18.811659
    israel        2.8496473
    war        1.424824
    nuclear        0.47494125
    ian        1.3971453
  real time traffic        17.002417
    msn        2.5182645
      direct        2.5182645
    vehicle        2.6913335
      probe        2.6913335
    cobra        2.8973334
    predictive        3.0361946
    maptuit        2.9770021
  myspace        14.777991
    home page        5.247993
      login        0.91739845
    www        0.13987407
  weather forecast        11.24317
    washington        0.20149167
      dc        0.20149167
    las        3.7851806
      vegas        3.7851806
    channel        1.7150953
  las vegas        9.56295
    weather        4.5601344
      forecast        1.2665414
    nv        1.2665106
    nevada        0.15831374
  south park        14.451366
    word        4.6359124
  cnn        13.221818
    news        5.7079
      headlines        1.6024663
  tube        11.490906
    web        0.30610773
      site        0.30610773
        videos        0.20245522
    videos        0.8636281
    you com        0.5649932
  china        7.541796
    india        1.2665102
    military        1.5752164
  msn        10.852091
    home        3.0217946
      page        3.0217946
  antonella        9.136532
    american idol        0.6054088
      barba pictures        0.6054088
        photos        0.11749599
        pics        0.13720602
        view        0.08955016
    barba        5.839789
      pictures        2.486525
      smoking        0.5425824
        gun        0.5425824
      photo        0.10473552
  mapquest        10.029214
    yahoo        0.99608874
      driving        0.99608874
        directions        0.99608874
    driving        2.8588884
      directions        2.8588884
  green tea        10.607022
```

```
  benefits        2.6913335
pokey chatman        9.483155
  chapman        0.77589756
brain        8.573813
  man        7.62393
    60 minutes        3.349459
      daniel        1.1786182
immigration bill        8.506484
  bush        0.85467273
    president        0.85467273
  clinton        0.60957104
  gates        0.39306188
60 minutes        5.3779135
  man        2.509996
    daniel        1.0258584
      brain        1.0258584
    brain        1.4841375
  savant        1.1081964
barack obama        7.2782416
johnny depp        7.207243
  daughter        1.7414513
hillary clinton        5.9501643
  bill        1.213588
myrtle        6.257768
  weather        2.53302
    beach        2.53302
      forecast        2.53302
  beach        3.7247484
    sc        1.2665102
daylight savings        6.681399
  time        2.3069649
    change        0.8310265
  day        0.39706218
    light        0.39706218
  saving        0.65278244
maps        5.2319393
  google        1.1081967
  yahoo        2.1089146
jennifer        4.66108
  hudson        2.6580057
  lopez        0.7365637
  aniston        1.2665102
orlando        5.463888
  weather        0.84447265
    florida        0.84447265
  florida        0.09091188
  fl        0.94767505
  bloom        0.25623906
games        5.6022763
  yahoo        1.6444322
time        5.540262
  daylight savings        4.1154385
    day        0.31662756
      light        0.31662756
    saving        1.3821213
    change        0.8035026
  cable        0.633255
    warner        0.633255
top model        6.5113616
  america        6.5113616
    s next        5.8781066
basketball        4.694462
  tournament        0.8453242
    big        0.33557007
```

```
    ten       0.12128693
    12       0.21428315
  sec       0.37658584
  acc       0.13316818
florida       0.41451555
state       0.6332551
  ohio       0.6332551
college       0.7388275
ncaa       0.7960295
syracuse       0.47494125
```

Appendix B

Evaluation

B.1 Evaluation Instructions

Evalueringsinstruksjon

Systemet som skal evalueres er en søkehjelper for nyhetssøk. Basert på 3 dager med querylogger har systemet funnet grupper av lignende queries, og omformet disse gruppene til trestrukturer som brukeren kan navigere i. Dette systemet er koblet opp mot en testinstallasjon av søkekjernen Vespa, til Yahoo. Nyhetsdokumenter og query logger er på Engelsk/Amerikansk

Evalueringen gjennomføres ved å jobbe seg gjennom en liste av 10 forskjellige queries. Disse har blitt valgt ut av de 20 mest populære gruppene som mest relevante for news. For hvert query er det en link som starter evalueringen, for å forsikre at riktige parametre blir satt. På venstre side av resultatsettet vises vanlig yahoo nyhetssøk(regular search) mens på høyre side vises det forbedrede systemet(improved search). For hvert query skal improved search resultatsettet, navigasjonstreet og antall unike nyhetssaker for de to forskjellige søketypene bestemmes. NB Bare første nivået av treet i forbedret søk skal evalueres, dvs den resultatsiden man får opp ved å klikke på linkene.

Så litt veiledning til evalueringen Not Related Story : nyheteten er ikke relatert til orginalqueriet Matching Regular Search : nyheten omhandler samme personer og er knyttet til en bestemt begivenhet som en eller flere artikler i Regular Search New News Story : en nyhet som det ikke finns lignende til i Regular Search (ny i forhold til regular search, men trenger ikke å være ny i forhold til andre nyheter i improved search)

Evalueringen av trestrukturen går på nytteverdien av å ha en trestruktur å navigere i, husk at dette er basert på 3 dager med logger så en kan ikke forvente et komplett komplekst tre.

Det skal bestemmes hvor mange unike nyheter er det i resultatsettet for både regular og improved search. Typisk hvis en nyhet omhandler samme personer og knyttet til en bestemt begivenhet så er det den sammen nyhetssaken.

Skriv inn ditt brukernavn for epost i username feltet, dette er meget viktig da forskjellige personer som bruker samme brukernavn vil skrive over den andres resultater. Det er mulig å svare på samme query flere ganger, da blir det forrige resultatet overskrevet. Alle felter må være fyllt ut for å kunne sende inn svar.
På linken http://129.241.152.222:12022/?results_for_user_name=brukernavn kan du se hva du har svart, og fint om du sjekker at alt har blitt svart på når evalueringen er gjennomført.

Evalueringen er rimelig kjapp og gjennomføres fint på godt under en time.

Liste over queries som skal evalueres:

antonella barbra http://129.241.152.222:12022/?query=antonella%20barba&mode=improved &evaluation=true&navigation=true&tree_navigation_node_id=18&query_number=1

american idol http://129.241.152.222:12022/?query=american+idol&search=Search&mode=improved &navigation=true&evaluation=true&query_number=2

britney spears http://129.241.152.222:12022/?query=britney+spears &search=Search &mode=improved&navigation=true&evaluation=true&query_number=3

mega millions http://129.241.152.222:12022/?query=mega%20millions &mode=improved &evaluation=true&navigation=true&tree_navigation_node_id=76&query_number=4

baseball http://129.241.152.222:12022/?query=baseball&search=Search &mode=improved &navigation=true&evaluation=true&query_number=5

immigration http://129.241.152.222:12022/?query=immigration &mode=improved &evaluation=true&navigation=true&tree_navigation_node_id=101&query_number=6

grand canyon http://129.241.152.222:12022/?query=grand+canyon&search=Search &mode=improved &navigation=true&evaluation=true&query_number=7

president bush http://129.241.152.222:12022/?query=president+bush&search=Search &mode=improved &navigation=true&evaluation=true&query_number=8

tournament http://129.241.152.222:12022/?query=tournament &mode=improved &evaluation=true&navigation=true&tree_navigation_node_id=122&query_number=9

anna nicole http://129.241.152.222:12022/?query=anna+nicole&search=Search &mode=improved &navigation=true&evaluation=true&query_number=10

B.2 Evaluation Data

```
User name : user1
query number 1
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : new article 12 : new article 13 : new article 14 : match 15 : match
 16 : match 17 : match
unique queries in regular serach: 7
unique queries in improved serach: 8

query number 2
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : match 7 : new article 8 : not relevant 9 : new article 10 : new article
 11 : new article 12 : match 13 : new article 14 : new article 15 : new article
 16 : new article 17 : new article 18 : new article 19 : new article
unique queries in regular serach: 9
unique queries in improved serach: 15

query number 3
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : match 7 : match 8 : new article 9 : match 10 : not relevant
 11 : not relevant
unique queries in regular serach: 7
unique queries in improved serach: 9

query number 4
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : new article 8 : new article 9 : match 10 : new article
 11 : match 12 : match 13 : match
unique queries in regular serach: 6
unique queries in improved serach: 8

query number 5
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : new article
 6 : not relevant 7 : not relevant 8 : not relevant 9 : new article 10 : not relevant
unique queries in regular serach: 6
unique queries in improved serach: 7

query number 6
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : match 13 : new article
unique queries in regular serach: 7
```

```
unique queries in improved serach: 11

query number 7
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : new article
unique queries in regular serach: 5
unique queries in improved serach: 3

query number 8
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : match
 6 : not relevant 7 : not relevant 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article 14 : not relevant 15 : not relevant
unique queries in regular serach: 4
unique queries in improved serach: 8

query number 9
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : not relevant 7 : new article 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article
unique queries in regular serach: 10
unique queries in improved serach: 11

query number 10
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : new article 9 : new article 10 : new article
 11 : new article 12 : not relevant 13 : not relevant
unique queries in regular serach: 8
unique queries in improved serach: 10

User name : user2
query number 1
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : new article 12 : match 13 : match 14 : match 15 : match
 16 : match 17 : match
unique queries in regular serach: 1
unique queries in improved serach: 2

query number 2
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : match 7 : new article 8 : new article 9 : match 10 : match
 11 : match 12 : new article 13 : new article 14 : match 15 : match
 16 : match 17 : match 18 : match 19 : match
unique queries in regular serach: 6
unique queries in improved serach: 8
```

```
query number 3
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : match 7 : match 8 : match 9 : match 10 : not relevant
 11 : not relevant
unique queries in regular serach: 1
unique queries in improved serach: 3

query number 4
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : match 12 : match 13 : match
unique queries in regular serach: 2
unique queries in improved serach: 2

query number 5
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
unique queries in regular serach: 6
unique queries in improved serach: 8

query number 6
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : match 9 : match 10 : new article
 11 : new article 12 : match 13 : match
unique queries in regular serach: 8
unique queries in improved serach: 9

query number 7
navigate help : medium
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match
unique queries in regular serach: 1
unique queries in improved serach: 1

query number 8
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : match
 6 : new article 7 : new article 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article 14 : new article 15 : new article
unique queries in regular serach: 1
unique queries in improved serach: 7

query number 9
navigate help : good
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : match
```

```
 11 : match 12 : new article 13 : new article
unique queries in regular serach: 10
unique queries in improved serach: 13

query number 10
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : not relevant 12 : not relevant 13 : not relevant
unique queries in regular serach: 4
unique queries in improved serach: 5

User name : user3
query number 1
navigate help : good
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : new article 12 : new article 13 : new article 14 : match 15 : match
 16 : match 17 : match
unique queries in regular serach: 7
unique queries in improved serach: 7

query number 2
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : match 13 : new article 14 : new article 15 : new article
 16 : new article 17 : new article 18 : new article 19 : new article
unique queries in regular serach: 10
unique queries in improved serach: 16

query number 3
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : not relevant 5 : not relevant
 6 : match 7 : new article 8 : new article 9 : match 10 : not relevant
 11 : not relevant
unique queries in regular serach: 6
unique queries in improved serach: 2

query number 4
navigate help : medium
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : new article 8 : new article 9 : match 10 : new article
 11 : match 12 : match 13 : match
unique queries in regular serach: 6
unique queries in improved serach: 7

query number 5
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
unique queries in regular serach: 6
```

```
unique queries in improved serach: 9

query number 6
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : match 9 : match 10 : new article
 11 : new article 12 : new article 13 : new article
unique queries in regular serach: 10
unique queries in improved serach: 11

query number 7
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match
unique queries in regular serach: 8
unique queries in improved serach: 4

query number 8
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : not relevant 7 : not relevant 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article 14 : not relevant 15 : not relevant
unique queries in regular serach: 5
unique queries in improved serach: 8

query number 9
navigate help : good
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article
unique queries in regular serach: 10
unique queries in improved serach: 13

query number 10
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : new article 9 : new article 10 : match
 11 : match 12 : not relevant 13 : not relevant
unique queries in regular serach: 7
unique queries in improved serach: 7

User name : user4
query number 1
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : new article 12 : new article 13 : new article 14 : match 15 : match
 16 : match 17 : match
unique queries in regular serach: 8
unique queries in improved serach: 8

query number 2
```

```
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : not relevant 9 : new article 10 : new article
 11 : new article 12 : match 13 : new article 14 : match 15 : new article
 16 : new article 17 : new article 18 : new article 19 : new article
unique queries in regular serach: 10
unique queries in improved serach: 18

query number 3
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : match 10 : not relevant
 11 : not relevant
unique queries in regular serach: 10
unique queries in improved serach: 8

query number 4
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : new article 8 : new article 9 : new article 10 : new article
 11 : match 12 : match 13 : match
unique queries in regular serach: 7
unique queries in improved serach: 8

query number 5
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : new article
 6 : new article 7 : not relevant 8 : new article 9 : new article 10 : new article
unique queries in regular serach: 6
unique queries in improved serach: 8

query number 6
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : match 13 : new article
unique queries in regular serach: 10
unique queries in improved serach: 12

query number 7
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : new article
unique queries in regular serach: 6
unique queries in improved serach: 4

query number 8
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : match
 6 : new article 7 : new article 8 : new article 9 : new article 10 : match
```

```
 11 : match 12 : new article 13 : new article 14 : new article 15 : new article
unique queries in regular serach: 8
unique queries in improved serach: 12

query number 9
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article
unique queries in regular serach: 10
unique queries in improved serach: 12

query number 10
navigate help : medium
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : new article 9 : new article 10 : new article
 11 : new article 12 : not relevant 13 : not relevant
unique queries in regular serach: 8
unique queries in improved serach: 9

User name : user5
query number 1
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : new article 12 : match 13 : match 14 : match 15 : match
 16 : match 17 : match
unique queries in regular serach: 7
unique queries in improved serach: 7

query number 2
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : match 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : new article 13 : new article 14 : match 15 : match
 16 : new article 17 : new article 18 : new article 19 : new article
unique queries in regular serach: 8
unique queries in improved serach: 14

query number 3
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : match 7 : match 8 : new article 9 : match 10 : not relevant
 11 : not relevant
unique queries in regular serach: 6
unique queries in improved serach: 5

query number 4
navigate help : good
navigation overview : good
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : new article 8 : new article 9 : match 10 : new article
 11 : match 12 : match 13 : match
```

```
unique queries in regular serach: 2
unique queries in improved serach: 4

query number 5
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
unique queries in regular serach: 5
unique queries in improved serach: 8

query number 6
navigate help : good
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : match 13 : new article
unique queries in regular serach: 6
unique queries in improved serach: 10

query number 7
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match
unique queries in regular serach: 1
unique queries in improved serach: 1

query number 8
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : match
 6 : new article 7 : new article 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article 14 : new article 15 : new article
unique queries in regular serach: 4
unique queries in improved serach: 9

query number 9
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article
unique queries in regular serach: 7
unique queries in improved serach: 10

query number 10
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : new article 12 : not relevant 13 : not relevant
unique queries in regular serach: 7
unique queries in improved serach: 9

User name : user6
query number 1
navigate help : good
```

```
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : match 8 : match 9 : match 10 : match
 11 : match 12 : new article 13 : new article 14 : match 15 : match
 16 : new article 17 : new article
unique queries in regular serach: 5
unique queries in improved serach: 6

query number 2
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : not relevant 8 : not relevant 9 : new article 10 : new article
 11 : new article 12 : match 13 : new article 14 : new article 15 : new article
 16 : new article 17 : new article 18 : not relevant 19 : not relevant
unique queries in regular serach: 8
unique queries in improved serach: 11

query number 3
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : match 7 : new article 8 : new article 9 : match 10 : not relevant
 11 : not relevant
unique queries in regular serach: 7
unique queries in improved serach: 5

query number 4
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : match 7 : new article 8 : new article 9 : match 10 : new article
 11 : match 12 : match 13 : match
unique queries in regular serach: 5
unique queries in improved serach: 6

query number 5
navigate help : medium
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : match 5 : new article
 6 : not relevant 7 : not relevant 8 : new article 9 : new article 10 : new article
unique queries in regular serach: 5
unique queries in improved serach: 6

query number 6
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : new article 13 : new article
unique queries in regular serach: 5
unique queries in improved serach: 7

query number 7
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
```

```
 6 : match 7 : new article
unique queries in regular serach: 7
unique queries in improved serach: 5

query number 8
navigate help : poor
navigation overview : poor
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : match
 6 : not relevant 7 : not relevant 8 : new article 9 : new article 10 : match
 11 : match 12 : new article 13 : new article 14 : not relevant 15 : not relevant
unique queries in regular serach: 4
unique queries in improved serach: 5

query number 9
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : new article 5 : new article
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : new article 13 : new article
unique queries in regular serach: 8
unique queries in improved serach: 10

query number 10
navigate help : medium
navigation overview : medium
improved search
 1 : match 2 : match 3 : match 4 : match 5 : match
 6 : new article 7 : new article 8 : new article 9 : new article 10 : new article
 11 : new article 12 : not relevant 13 : not relevant
unique queries in regular serach: 7
unique queries in improved serach: 8
```