DEPARTMENT OF INFORMATION SCIENCE AND MEDIA STUDIES

MASTER THESIS

Use of Wikipedia content, structural connections and usage statistics to generate context aware query augmentation in a topical search engine

Øyvind Døskeland

December 22, 2012

Table of content

Abstract

This thesis presents the TCSearch2, a Master's project. The thesis studies different approaches to bridging the gap between user expectations and existing search engine result and their impact on the quality of the results. Four search engines were developed to evaluate the methods proposed by this thesis. This was achieved by using publicly available data from the online encyclopedia - Wikipedia. Content, structure, such as links, and usage statistics from Wikipedia were extracted and applied in the process of creating the general knowledge base for topic identification. The knowledge base is used for the query augmentation process. To bridge the mentioned gap, the search engines developed needed some intelligent capabilities; those intelligent capabilities are contextual topic identification of user input. Users have access to directly work with the augmented query terms and weight of the terms. An online public prototype of the TCSearch2 project will be deployed by 2013.

Two types of studies have been conducted to evaluate the developed search engines: a qualitative study with seven test subjects in a laboratory evaluation, with a total duration of 21 hours, and a quantitative search simulation, with a total of 30 different queries. In the qualitative study, the subjects’ usage data and feedback were analyzed. In the quantitative evaluation, the developed search engines were compared to existing search engines, including Google and Wikipedia's search engine.

The studies show that the proposed methods of this thesis reduce the gap between users’ expectations and search engine results.

Keywords

Search Engine, Wikipedia, Topic Identification, Query Enhancement, Query Augmentation, Query Expansion, Intelligent Search, Semantic Search, Big Data, Self-Organized Map

Acknowledgments

I would like to thank my supervisor Weiqin Chen for the help and motivation. Especially for believing in me even when the motivation was low. A large part of the motivation to work on this thesis was the work Josef Pihera and I did on the TCSearch project in 2011. I would therefore call this search engine core TCSearch2. This project would not have been possible without all the contributors and the donors of Wikipedia. During this process, Wikipedia’s quality and range of topics have surprised me in a very positive way almost every day. But most of all I would like to thank my significant other Ingunn Rose with all the help over the past year, with all my ups and downs.

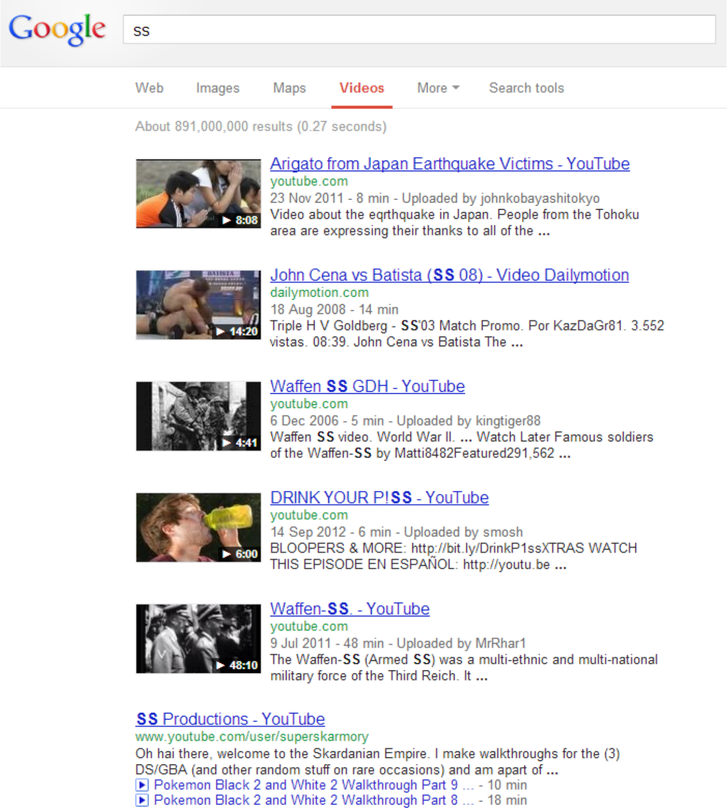
Chapter 1 – Introduction

1.1 Background

From the start of civilization and throughout history, people have looked for new ways to acquire information, new ways to preserve the knowledge and new ways to share the knowledge with others. Throughout time, improvements have been made for storing and accessing information. Often, improvements have been made in times when the amount of information available or needed made traditional ways of storage and retrieval difficult to use or maintain. In the last few decades, this discipline of information storage and retrieval has to a large extent been transformed into the sub-field of Computer Science, Information Retrieval. Today, every single minute, users all over the world are contributing to the growth of the internet, with 2,4 billion users at the beginning of 2012 and growing in ever larger numbers (Internet World Stats, 2012). This ever growing tidal wave of data may result in the need to have a new approach to Information Retrieval. Today's dominant web search provider Google alone has over four billion queries per day, and over one billion unique visitors per month (Experian, 2012; Efrati, 2011; Statistic Brain, 2012). This can indicate that searching is a part of online users' everyday activity.

The current dominating web search engine design uses input term to index term matching (Manning, Raghavan & Schütze, 2008, page 100). Term to term matching can produce a search result that is different from the expectation of the user. This difference occurs due to how a term to term search engine assesses if a document should be included in the search result. In a term to term search engine a document is based to be relevant solely based on the on the input terms provided by the user and not the wider content of the document in the search result. The gap between the user's expectations and the search results is in some cases considered to be unintelligent from the point of view of the user. One of the theories proposed for improving the results is semantic processing of the search input, to identify the topic of the query or search input and perform a topical search (Witten & Nichols, 2007).

One example of this gap can be seen in Figure 1, the search result of Google's video search engine with the query input "SS". The first hit was a video from the 2011 earthquake in Japan. With no mention of "SS" in the document or any "SS" topical mentions other than in the auto generated URL of the video http://www.youtube.com/watch?v=SS-sWdAQsYg. Retrieving results that have no semantic connection to the search input can be seen as unintelligent behavior of the search engine from a user standpoint.



About 891,000,000 results (0.27 seconds)

Arigato from Japan Earthquake Victims - YouTube
youtube.com
23 Nov 2011 - 8 min - Uploaded by johnkobayashitokyo
Video about the eqrthquake in Japan. People from the Tohoku area are expressing their thanks to all of the ...

John Cena vs Batista (**SS** 08) - Video Dailymotion
dailymotion.com
18 Aug 2008 - 14 min
Triple H V Goldberg - **SS**'03 Match Promo. Por KazDaGr81. 3.552 vistas. 08:39. John Cena vs Batista The ...

Waffen **SS** GDH - YouTube
youtube.com
6 Dec 2006 - 5 min - Uploaded by kingtiger88
Waffen **SS** video. World War II. ... Watch Later Famous soldiers of the Waffen-**SS** by Matti8482Featured291,562 ...

DRINK YOUR P!**SS** - YouTube
youtube.com
14 Sep 2012 - 6 min - Uploaded by smosh
BLOOPERS & MORE: http://bit.ly/DrinkP1ssXTRAS WATCH THIS EPISODE EN ESPAÑOL: http://youtu.be ...

Waffen-**SS**. - YouTube
youtube.com
9 Jul 2011 - 48 min - Uploaded by MrRhar1
The Waffen-**SS** (Armed **SS**) was a multi-ethnic and multi-national military force of the Third Reich. It ...

SS Productions - YouTube
www.youtube.com/user/superskarmory
Oh hai there, welcome to the Skardanian Empire. I make walkthroughs for the (3) DS/GBA (and other random stuff on rare occasions) and am apart of ...
Pokemon Black 2 and White 2 Walkthrough Part 9 ... - 10 min
Pokemon Black 2 and White 2 Walkthrough Part 8 ... - 18 min

Figure 1 Google search result for "SS" (Google Search Result, 2012)

A different example to illustrate the gap is when a user searches for "flowers". If a document contains all types of flowers such as rose, orchid or magnolia, but not the term flower, it will not show in the search results. From a user's standpoint, this can be considered to be a gap between the expectations and the results.

Popular existing search engines use several methods to improve term to term search such as synonym search, stemming, TF-IDF, term zone weighting and PageRank to optimize the results. But still there are several examples where these methods are not able to bridge the gap between the users' expectation and the results of the search engines.

While a second alternative is search engines enhanced by semantic web technologies, those search engines have had minor success due to their demand for a total shift in the existing search engine storage structure and design (Finin, Peng, Cost,Sachs, Joshi, Reddivari, Pan, Doshi & Ding, 2004). In addition, ontology creation for the semantic processing is a time consuming and expensive task (Gruninger & Lee, 2002). SPARQL, the most widely used Semantic Web query engine, have had limited success in scalability with large data sets (Huang, Abadi & Ren, 2011), thus making it too slow for a full web search engine.

This project will create an additional alternative that uses the same query augmentation methods in existing popular search engines and incorporates intelligent input query processing capabilities. This is achieved by using a general knowledge base created from data in Wikipedia.

1.2 Aim

The aim of this project is to show how an already existing technology query augmentation, combined with an automatically created knowledge base using public information from the online encyclopedia - Wikipedia, can improve search results compared to term to term search engines. This project is based on several existing studies as further described in Chapter 2. This project will contribute with new ways of processing the usage of Wikipedia to further improve Wikipedia as a knowledge base. This new approach aims to decrease the distance of the users' expectation of the result and the search engine's actual result.

Since the first web search engines were created, several studies have been carried out on the combinations of using knowledge as the base for query augmentation by extending the original query (Gauch & Smith, 1993). Knowledge bases used in query augmentation by query expansion were typically private closed scientific or commercial knowledge bases that were maintained manually by experts. Most knowledge were domain specific, expensive to maintain and often not up to date. Thus, not usable for a full web search engine (Manning et al, 2008, page 174-175).

One of the websites that have experienced large increase of data the last couple of years is the online encyclopedia - Wikipedia. The amount of unique articles made publicly accessible on Wikipedia is over four million, and a monthly growth of up to 60 000 new Wikipedia articles (Wikipedia G, 2012). Wikipedia is updated constantly by a horde of contributors, and a new database dump of the English Wikipedia is made public in a less then weekly interval (Wikipedia H, 2012). As Wikipedia increases in size and quality, the quality of the result from search engines that uses Wikipedia as a knowledge base, could also improve.

This project will process the usage of Wikipedia articles as a method to further improve Wikipedia as an auto generated general knowledge base. This is achieved by looking at the language dispersal of the Wikipedia articles' readers. A Wikipedia article is often written in several languages. For example, the Wikipedia article of the finance minister in Estonia, Andris Vilks, http://en.wikipedia.org/wiki/Andris_Vilks is written in five different languages: English, Estonian, Russian, French and Polish. The amount of readers of the different languages is listed below:

- English 473 times in the last 90 days.

- Estonian 54 times in the last 90 days.
- Russian one time in the last 90 days.
- French 92 times in the last 90 days.
- Polish 286 times in the last 90 days.

The idea is to use the dispersal of the Wikipedia readers’ languages to impact the strength of the connection between Wikipedia articles. Wikipedia articles that have a similar dispersal of languages it is read in are believed to be closer related than Wikipedia articles that have a dissimilar dispersal of languages. This concept is referred to as language dispersal in this thesis.

In addition, a web search interface was constructed for testing and evaluation purposes so users could interact with the search engines. The frontend construction of the prototype could also give other researchers in this research field the possibility to compare results with this project. But only minimal resources were put into the frontend development due to the nature of the intended application design of this project.

As a whole-Internet search is expensive and impossible to implement with the limited resources at my disposal, a subset of the Internet was used. With such restrictions, it was then possible to work with real data. The desire to use free publicly available data sets, so that future studies could do a comparative evaluation with this project, limited the choice of the corpus to be used. With the existing restrictions, Wikipedia was chosen as the search corpus for this project. Aspects that make Wikipedia the ideal choice for a knowledge base can make Wikipedia a less than ideal choice for a search corpus, but Wikipedia has a considerable size and a wide range of topics. In addition, Wikipedia nonetheless had to be processed when using it as a knowledge base for the query augmentation process.

The design of this project is to be an additional module in an existing search engine system design. An example of a simplified standard search engine system design is shown in Figure 2.

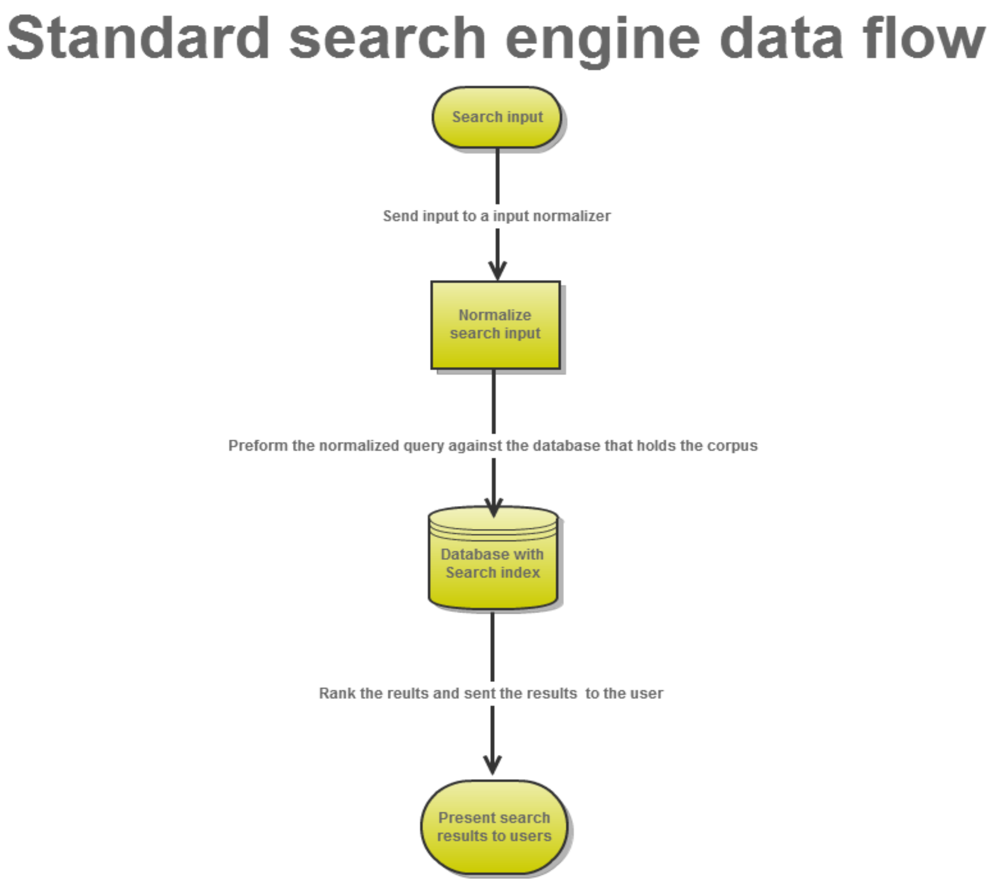

Standard search engine data flow
Search input
Send input to a input normalizer
Normalize search input
Preform the normalized query against the database that holds the corpus
Database with Search index
Rank the reults and sent the results to the user
Present search results to users

Figure 2 Simplified standard search engine system design (Manning et al, 2008, page 135)

In this project a module is added to the standard search engine design. This module is the query augmentation module. The proposed search engine system design for the query augmentation search engine for this project is shown in Figure 3.

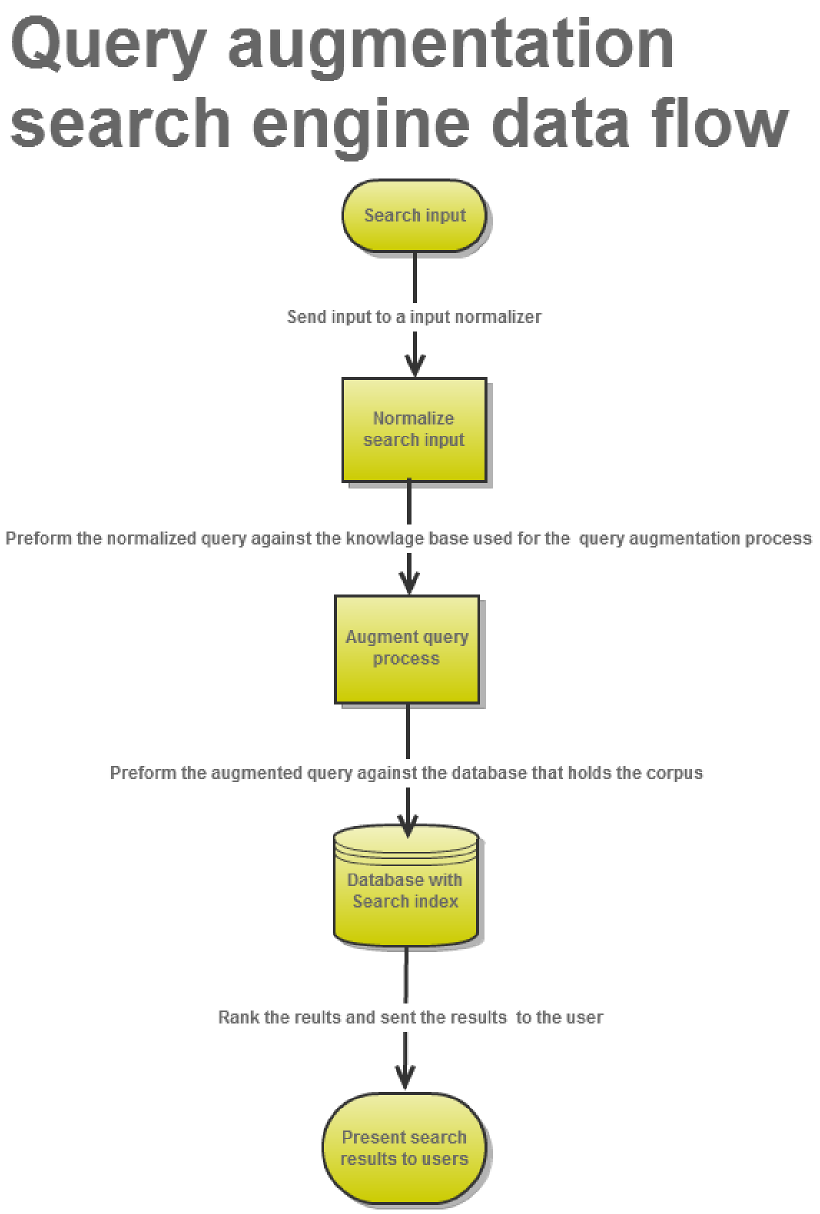

Query augmentation
search engine data flow
Search input
Send input to a input normalizer
Normalize search input
Preform the normalized query against the knowlage base used for the query augmentation process
Augment query process
Preform the augmented query against the database that holds the corpus
Database with Search index
Rank the reults and sent the results to the user
Present search results to users

Figure 3 proposed search engine system design for this project

Most existing search engines can with minimal effort add the query augmentation module to their existing search engine as a plug-in, without the risks and costs associated with any major system redesign.

Commercial search engines have strict performance criteria. Those performance criteria put a limit on to what degree a search algorithm can be computationally demanding. Google uses a great amount of resources to optimize down to the millisecond (Google Forum, 2012). The query augmentation module developed for this project was solely for research purposes, several optimizations would be needed before using this type of module in a commercial grade search engine. While this project was not aimed at being a commercial grade product, efforts have been made on optimizing the implementation of the query augmentation process, such as optimizing MySQL memory caching usage and using a multithreaded text parsing algorithm. This was done to highlight the potential commercial viability of this project.

The query processing in this project happens after the user enters his/her query and before the engine starts searching through the search index. A search index is a structure where the terms are connected to the associated value the term within in a particular document in the corpus. This is also known as query augmentation or query enhancement which is a well-known principle. Unlike some query augmentation implementations where there is no guarantee that the original query will be the most important part of the augmented query, in this project the original query is to remain the most important part of the augmented query. The user might want to have a control of such a process. The augmented query is visible to the users and the users may remove, add and change the terms in the augmented query.

This project is a continuation of the TCSearch (Topical Contextual search) project by Josef Pihera and Øyvind Døskeland developed in the spring of 2011. The original TCSearch was considerable smaller in size and used less than 3000 Wikipedia articles from the sub-category health and lifestyle in Wikipedia. The 3000 Wikipedia articles were processed into the knowledge base that was used for the query augmentation process. This knowledge base was not a general knowledge base, but a health and lifestyle limited knowledge base. The duration of the development of the original TCSearch project was only a couple of months, thus none of the methods implemented in the TCSearch was usable for a large scale project such as the TCSearch2. The TCSearch2 project was one of the proposed future work proposals of the TCSearch project, the implementation of a general knowledge base by processing all Wikipedia articles.

To sum it up, in this project a web interface has been developed, similar to what the user is probably familiar with, which enables him/her to perform the standard search easily, while also providing a way to work with the whole complexity of resulting query augmentation. TCSearch2 has an interface which is the front-end layer that is connected to the several fully implemented search engine server logics. Nevertheless, it must be noted, that the main aim of this project is not to produce a full-fledged search engine, but to enrich search with intelligent capabilities gained from knowledge acquired from Wikipedia.

1.3 Research questions

The research questions in this thesis can be divided into two groups. The first group consists of two research questions regarding the result quality of different types of search engines with different features added. The second group consists of a single research question relating to the differences in the result quality between existing search engines and the different search engines from this project.

1) Can the search engine developed using query augmentation based on a knowledge base created from Wikipedia content and structure reproduce the positive result of previous studies?

2) Can usage statistics from Wikipedia be processed in a manner that creates a positive impact on the query augmentation process, and reduces the gap between the expectations of the users and the search result?

3) How do the results from the different TCSearch2 search engines compare to popular existing web and domain search engines?

1. 4 Organization of the thesis

This thesis has 7 distinct chapters. Every chapter will present a different topic or aspect of this thesis. The first chapter is the introduction of the thesis. It describes the motivation and the aim behind the thesis and explains some of the shortcomings of today's existing solutions.

The second chapter will put this thesis in the context of similar work regarding the use of Wikipedia as a source of improving search. Several researchers have used Wikipedia in different ways, with focus on different parts of Wikipedia. This will be presented in greater detail in the second chapter.

The third chapter will focus on the fundamental research this thesis is based on, including Information Research and Artificial Intelligence aspects.

Chapter four will give an introduction to Wikipedia, the architecture, purpose and history. This chapter explains why Wikipedia is currently a good candidate to be used as a general knowledge base for query augmentation.

Chapter five will present the TCSearch2 system data flow of the query augmentation module and development aspects of this project are explained. The different steps, tools and methods used in this project will be explained.

Chapter six will explain the different evaluations preformed in this project. This chapter will also present the data from evaluations that were conducted during development, and evaluating data from the final evaluations.

The last chapter presents the conclusions may be drawn from this project. Also, some proposals to future work are presented.

Chapter 2 – Similar work

In this chapter this thesis will be put in context of the existing large field of research in Information Retrieval that uses knowledge bases automatically generated from Wikipedia. Wikipedia with its vast amount of data has been the center of several studies. Several aspects of Wikipedia have been used in research for improving the result quality from search engines.

The study “On improving Wikipedia search using article quality” from the University of Singapore focused on Wikipedia contributors’ edit history (Hu, Lim, Sun, Lauw & Vuong, 2007). Contributors grouped as high contributors would by editing an article boost that Wikipedia articles value. Using the contributors edit history to rerank search results from the Wikipedia search engine, this study was able to achieve search result accuracy comparable to Google.

The findings of the study “Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge” indicated that there exists a potential in using Wikipedia as a semantic network to look at semantic connection strength between terms (Han & Zhao, 2009). An example used in the paper was “IBM”. Processing the Wikipedia article dump, this study was able to calculate with accuracy the semantic relationship between the terms “IBM”, “Big Blue” and “International Business Machine Corporation”. The conclusion of the study indicated that query augmentation through co-occurrence based on Wikipedia content and structure was achievable.

“Extracting Semantic Relationships between Wikipedia Categories” is a study on the possibility of using the links between Wikipedia articles and using the Wikipedia categories to calculate the semantic connection strength between Wikipedia articles (Chernov, Iofciu, Nejdl & Zhouz, 2006). The findings of the study showed that links between Wikipedia articles correlates with the semantic connection strength between the Wikipedia articles. A larger study of semantic relatedness between articles based on links between Wikipedia articles was performed by Michael Strube and Simone Paolo Ponzetto (Strube & Ponzettoz, 2006). This study was called WikiRelate! and it compared WordNet with Wikipedia in computing semantic relatedness. One of the interesting findings was the quality of the automatically created taxonomies when using Wikipedia.

> “What is most interesting about our results is that they indicate that a collaboratively created folksonomy can actually be used in AI and NLP applications with the same effect as hand-crafted taxonomies or ontologies” (Strube & Ponzetto, 2006)

While there are hundreds of millions of links between Wikipedia articles, some links may be missing. The study “Discovering Missing Links in Wikipedia”, conducted in 2005, proposed a simple method to find missing links by the following algorithm.

> “First, we compute a cluster of highly similar pages around a given page, and then we identify candidate links from those similar pages that might be missing on the given page.“ (Adafre & Rijke, 2005)

Discovering missing links may further improve the possibility to correctly extract semantic relationships between Wikipedia articles, thus improve the query augmentation process for this project.

The study “Improving Web Search Ranking by Incorporating User Behavior Information“ performed at Microsoft Research in 2006 was interesting for this project with regard to the use of usage statistics to improve the result quality (Agichtein, Brill & Dumais, 2006). The study finds that usage information can drastically improve ranking.

> “We show that incorporating user behavior data can significantly improve ordering of top results in real web search setting” (Agichtein et al, 2006)

Finally, Koru is a study and prototype search engine that uses Wikipedia as a knowledge base for query augmentation (Witten & Nichols, 2007). It is currently at version 2.0 and it is publicly available at http://www.greenstone.org/greenstone3/koru2.0/. The Koru search engine uses Wikipedia to create a knowledge base and uses this knowledge base to perform query augmentation by query expansion.

> “Koru use only the link structure and basic statistics for articles, which consume 500 MB” (Witten & Nichols, 2007)

With minimal amount of data from Wikipedia the results from the Koru study were very promising. Thus the Koru study was a great motivation for this project. Several aspects of this master project were based on the Koru project; one of the aims of this project was to confirm the findings of the Koru study. In addition, the Koru public prototype web interface was helpful for this project to establish its aims and goal. When using the Koru search engine it became evident that there were several parts of Wikipedia not being used in the Koru project. One example was the lack of processing alternative titles of Wikipedia articles, thus when searching for the alternative Wikipedia article title “Floral”, the Koru search engine was unable to correctly process the input, but when searching for the main Wikipedia article title “Flower” the Koru search engine was able to correctly process the input from the user. Those lacking properties became an aim to fill in the TCSearch2 project by including substantially larger amounts of data from Wikipedia, such as the alternative titles of Wikipedia articles.

Chapter 3 – Information Retrieval, Artificial Intelligence and related concepts

In this chapter I will present the main research fields this thesis is based on. The Information Retrieval concepts and Artificial Intelligence methods used in the project, such as neural networks and particularly Self Organized Maps will be explained in detail.

3.1 Information Retrieval

There are several definitions of the field of Information Retrieval. The following definition is taken from one of the textbooks in Information Retrieval widely used at university level.

> “Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually documents) that satisfies the information need from within a large collections (usually stored on computers)” (Manning et al, 2008, page 1).

About 30-40 years ago, this definition would only apply to a handful of professions such as librarians and some researchers. Now, billions of people are involved in the Information Retrieval process. Most people would now describe the Information Retrieval process just as searching. One of the reasons for the large increase in searching is the overtaking of the complicated database lookup. There is no longer a need to know the ID number of a product to find information regarding that particular product, nor the need to have any computer engineering training to perform a search (Manning et al, 2008, page 1).

The definition stated above is a strict one, but this definition would fit the large majority of Information Retrieval Systems. In the strict definition of the term very few documents are truly unstructured, most documents normally have some structure, such as title, footnote, size and font of text (Manning et al, 2008, page 1).

The modern history of Information Retrieval shifted gear in the beginning of the 1990s with web search engines. The paradigm shift from simple Boolean retrieval to advanced ranking methods was introduced. The dominance of Google can be in part traced back to its link analysis algorithm PageRank. The basic idea behind link analysis is that documents that are often linked to from several other documents are more important than documents with were few other documents links to the document (Ma, 2008). This project will use a simplified version of PageRank algorithm..

3.1.1 Query Augmentation

Query Augmentation is the process of improving the users’ query input to achieve better result quality. Below is a list of some of the most used query augmentation methods:

- The most used query augmentation process is query reformulation, with spelling correction being the most common query reformulation scenario. Query reformulation is often based on query log mining, by looking at the manual query reformulation of past queries to suggest queries for new users. Query log reformulation is used in adding query terms with auto complete writing or suggesting alternative queries (Manning et al, 2008, page 173-175).

- A controlled vocabulary that is manually maintained is a different form of query augmentation. This approach has been used since the 1990s, but has predominantly been domain specific. An Information Retrieval system that uses such an approach is the Unified Medical Language System used with Medline. This is a specific system for biomedical literature. Where the most used query terms have been manually extended to improve result quality (Manning et al, 2008, page 173-175).

- Co-occurrence query augmentation is the process of extracting terms that have a high co-occurrence value from a collection of documents. The augmented query is extended by terms that have a high co-occurrence value, with an attached co-occurrence value. The idea behind co-occurrence in query augmentation is that terms with a high co-occurrence value are likely to have a semantic connection (Manning et al, 2008, page 173-175; Kraft, Chang, Maghoul & Kumam, 2006).

Statistical co-occurrence of terms in by a knowledge base is used in this projects query augmentation process. In addition the links between Wikipedia articles is used to further improve the co-occurrence correctness. This projects query augmentation process is described in detail in Chapter 5.

Below is an example of an augmented query, when input query was the term “html”. The original query was extended by 10 terms. The augmented query shows how terms that have a high co-occurrence to the original term are added with a corresponding value.

```
{html=0.4335192853375962,
markup=0.08038535362289888,
xhtml=0.07392004855327027,
browser=0.07106274987536501,
xml=0.0650348553776306,
web=0.06285991715721546,
w3c=0.06126501843999697,
css=0.05341810017667767,
document=0.05099356249632967,
html5=0.04754037898734221}
```

The general idea behind query augmentation by query extension is that documents containing several of the terms in the extended query is more likely to be relevant for the user than documents solely containing the original query input. For a query augmentation process to be successful, the extended terms must have a strong semantic connection to the original input.

3.1.2 Zone Weighting

Zone Weighting is the method of differentiating the weight given to a term or link based on structural information in a document. This process try to capture how humans normally communicate what is the most central information in a document. A common example of zone weighting is to value terms in the title in a document higher than terms in the sub-subsections of a document (Manning et al, 2008, page 101-104).

3.1.3 TF-IDF

TF-IDF is the process of using the term frequency(TF) with the combination of the inverse document frequency(IDF) to set a value of a term's weight in a document. The terms importance is to what degree a term can be used to differentiate between documents and this is the IDF value. A high IDF value is given to terms that only occurs in a few documents, while a low IDF value is given to a terms that most documents contain, often none-descriptive terms such as; "or", "are", "a". To calculate the IDF value of a term, the document frequency(DF) have to be counted. DF is the amount of documents in the corpus containing the term. With the DF value and the total amount of documents in the corpus**,** the IDF value can be calculate. The formula for the IDF value is as following:

> log(Amounts of documents in a corpus / amount of documents in the corpus containing the term.) (Manning et al, 2008, page 110)

TF is the term frequency of a term in a specific document. Standard TF-IDF value is given by the frequency of the term in a document multiplied by the IDF value of the term. This system is using the structural information to improve the TF value with structural information of the Wikipedia article (Manning et al, 2008, page 111).

3.1.4 Character Normalization

To improve the recall of the search engine result, all terms were character normalized. Character normalization is the process that changes all the characters in the input to their base characters. This process used in this project consisted of the three most common character normalization.

Step 1) Normalizing all special letters to the normal form
Step 2) Removing all non-letters or numbers from the input
Step 3) Convert all characters to lowercase.

An example of input "Führers!"
Result step 1) "Fuhrers!"

Result step 2) “Fuhrers”
Result step 3) “fuhrers”

While character normalization is a process intended to improve the result quality, it can have some negative side effects. Terms that have the same character base, but do not share any semantic connection the result quality is lowered. An example of this is “WHO” as the abbreviation of in the World Health Organization and the term “who”. (Manning et al, 2008, page 21-25)

3.1.5 Stemming

Stemming is the process of removing or changing the suffix of a term to bring the term to its grammatical root form. This project integrated an updated version of the widely used stemming algorithm, the Porter stemmer. Porter stemmer uses several steps that aim to result in a grammatical root form of a word by using suffix stripping (Porter, 1997). Contrary to character normalization, grammatical normalization is language dependent and there is a need to have an understanding of the language to create a stemming tool due to the different grammatical structures of different languages (Manning et al, 2008, page 30-33).
Stemming is usually performed after character normalization. We continue with the example from the character normalization, where we started with the term “Führers!” and ended up with “fuhrers”. Using stemming on “fuhrers” will remove the the plural suffix which in this case is the “s”. The grammatical root which will be returned is “fuhrer”.

This process can improve recall, but can have a negative impact on precision when the grammatical root form is shared by several different terms with no semantically relations. Examples are “animal” and “anime”, both would be stemmed to “anim”.

3.1.6 Tokenization

One of the most challenging aspects of text parsing is splitting a text into its terms or tokens. Knowing when a group of words are one term or several individual terms, are even for human experts at times a challenging task. (Manning et al. 2008, page 21)

In this project a text is split into different terms when there is a whitespace between terms or a non-character or number. The following example will show how a text D1 would be divided into its separate tokens using this projects tokenization algorithm.
D1 “O'Neill is a genius, too bad he is in South-Africa now”’

List of tokens from D1:

1 O
2 Neill
3 is
4 a
5 genius

6 too
7 bad
8 he
9 is
10 in
11 South
12 Africa
13 now

The tokenization algorithm used in this project has some limitation since it does not allow for multi term tokens. This process would not recognize “O’Neill” correctly as a term. A benefit of such primitive tokenization process is the limited computational needs, thus the process is fast compared to the language analyzing tokenization algorithms.

3.1.7 Inverse Index

Using pointers from terms to document IDs is known as an inverted search index or a posting list. In its simplest form an inverted index is a table that holds the term and a pointer to the documents thus connecting the term and the document together (Manning et al, 2008, page 9). An example of an inverted index constructed by the documents D1 and D2 is shown in Table 1.
D1:”the high wall is 5 meters high”
D2:”the man is 2 meters”

Table 1 Simple inverse index with a one to many pointer

Term	Document ID
The	D1,D2
Wall	D1
Man	D2
Is	D1,D2
5	D1
2	D2
Meters	D1,D2
High	D1

Table 1 shows a basic inverted index. But table 1 is expensive to maintain with removing documents from the corpus. A slightly different way of storing an inverted index as in Table 2, has a lower penalty for removing documents from the corpus.

Table 2 Simple inverse index with a one to one pointer

Term	Document ID
The	D1
The	D2
Wall	D1
Wall	D2
Man	D1
Is	D1
Is	D2
5	D1
2	D2
Meters	D1
Meters	D1
High	D1

The index in Table 1 and 2 it is a Boolean index with no value to the different terms. The term “high” had a TF of two in D1. Including value and/ or term positions in the inverse index, is also often done as seen in Table 3.

Table 3 Inverse index with term frequency and term position

Term	Document ID	Term Frequency	Term position
the	D1	1	1
the	D2	1	1
Wall	D1	1	3
Wall	D2	1	2
Is	D1	1	4
Is	D2	1	3
5	D1	1	5

2	D2	1	4
meters	D1	1	6
meters	D2	1	5
High	D1	2	2,7

3.1.8 Hash table

Hash table is a well known method that has been used in databases and programming languages for decades to speed up retrieval in a collection or table. Using an inverse index with a billion entries can be a slow process without using a hash table. A hash is the method of converting a key value to a corresponding index position of the entry (Manning et al, 2008, page 46). In this project hash tables was used to speed up the search process.

3.1.9 Bag of Words

Bag of Words is a term model that ignores the exact ordering of terms. Only the terms are stored, or the terms and corresponding values are stored. Not keeping the ordering in a document in the search index have some advantages such as drastically reduces storage requirements.
D1: “five is bigger than two”
D2: “two is bigger than five”

In a bag of words model document D1 and D2 are identical, but semantically they have very different meanings. While D1 and D2 have a different meaning, it is intuitive that documents that have a very similar bag of words representation probably contain similar content.
(Manning et al, 2008, page 107)

3.2 Artificial Intelligence

Artificial Intelligence, commonly written only as AI, is a combination of Computer Science, philosophy, logic, linguistics and math. AI is centered on creating programs that enable computers to display behaviors that can broadly be characterized as intelligent, or resemble human like behavior (Russell & Norvig, 2003).

The field of AI has existed since antiquity with the idea of intelligent robots dating back to ancient Greece (McCorduck, 2004). From 1950s the field of AI was proposed as a separate subfield of Computer Science by Alan Turing in "Computing Machinery and Intelligence". Artificial intelligence has since grown and exists both in academia and in commercially available products such as toys, mobile phones, cars and computer games to name a few.

3.2.1 AI in search

The field of AI has been used to improve search engines for decades. The use of a manually constructed knowledge base to improve search was used a few years after full web search engines were first developed (Lovic, Lu & Zhang, 2006). A newer addition is the search engine Siri. Siri is Apple Computers search engines for mobiles that takes speech as input and uses a combination of AI methods such as natural language processing (Strauss, 2012).

3.2.2 Artificial Neural Networks

Artificial Neural Networks(ANN) is a mathematical model inspired by the biological neural network found in the brain. Several interconnected artificial neurons form an ANN. The most common form of ANN is the multilayer perceptron(MLP) model. In MLP input nodes sends data forward to the hidden layers which sends the result to the output layer. Figure 4 is an example of an ANN MLP mode.

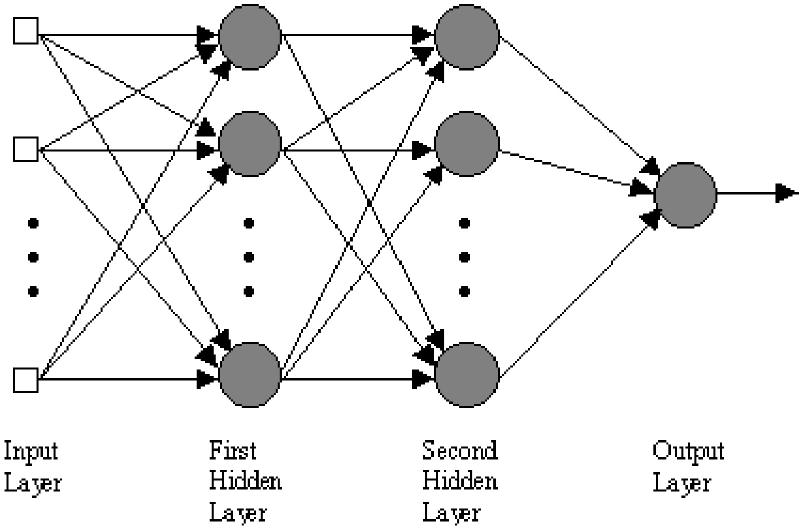


Figure 4 A multilayer perceptron ANN (NeuroSolutions, 2012)

MPL and several other ANN implementations uses back propagation to train the hidden layers in the ANN. This is done by repeatedly sending data through the ANN. The output data is compared to the desired output data and the error is computed. The error value will then be used as feedback to adjust the weights in the ANN to minimize the distance between the actual output and the desired output. Repeating this process over several iterations will train the ANN. Back propagation uses a training set of wanted output. For several tasks it is not possible to create a training set, such as finding new patterns and clusters.

3.2.3 Self-organizing map

Self-Organizing Map(SOM) is a popular non-parametric AAN algorithm based on unsupervised learning. Being non-parametric means that a SOM does not rely on any assumptions regarding the structure of the function it is approximating. SOM is also known as Kohonen maps, Kohonen network or Self-Organizing Feature Map(SOFM). SOM is used in various data mining task due to its beneficial properties such as vector quantization, projection and very low computational needs to calculate relative distances between multidimensional vectors. In addition, it does not use a training set which is a time consuming task to create (Brownlee, 2011). SOM was the algorithm chosen to calculate the relative distance between Wikipedia articles' language dispersal. To illustrate the SOM algorithm, I have created a basic SOM application in Java.

The initial SOM size is given by the desired height and width. Height * width gives the amount of neurons used in the SOM. There exist no single optimal size for a self-organized map, but domain knowledge should be combined with empirical tests be used.

> "For maps that are too large for the dataset, unnecessary folds occur and are penalized with a higher error value. The high values for the small maps are partly due to the fact that this measure is almost overly simplistic and suffers from the discrete nature of the output space."(Pölzlbauer, 2004)

Figure 5 shows the data input for the example SOM. This is an image of a forest, and the image consists of several hundred pixels. Each pixel is a three dimensional vector with red, green and blue values.

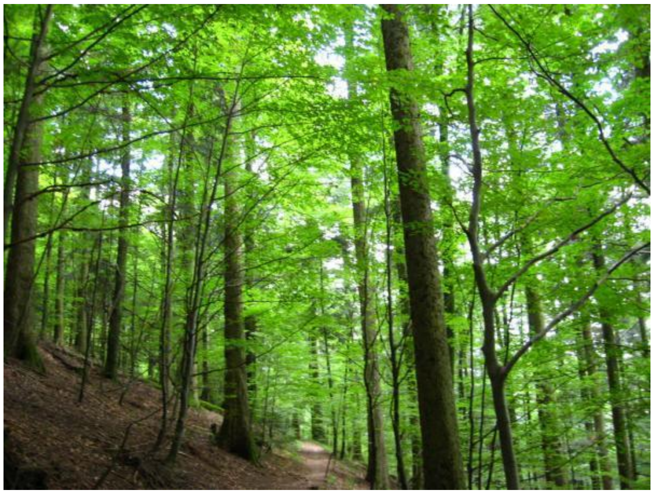


Figure 5 Data input for SOM example

Random data from the input from the Figure 5 will be used for training, this example use 10000 of the in total 270000 pixels in Figure 5. In this implementation one iteration uses one data vector or in this case a pixel, thus in this example there were 10000 training iterations.

The initial value of the neurons in the SOM are random values, in this example the values are from 0 to 255 to represent the RGB value. Figure 6 represents initial state of the example SOM.

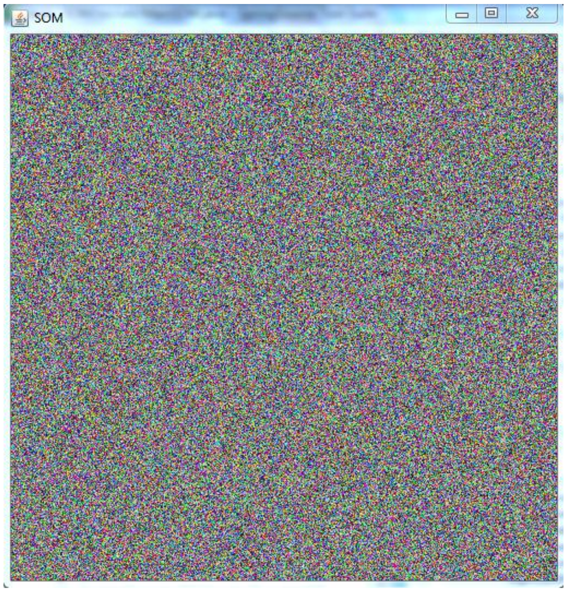


Figure 6 Initial state of the example SOM

The first step in a training iteration is to locate the best match unit(BMU). The BMU is the vector in the SOM that has the shortest distance from the training vector in this example the training pixel. Euclidean distance is a common algorithm used for distance measuring between two vectors for BMU calculation. If several data points have equally short distance, only one is selected at random to be the BMU.

$x^T \cdot y_i$ (Rojas, 1996, page 57) - the irritation of neuron ***i*** by input vector **x** , where y_i is the neuron's position

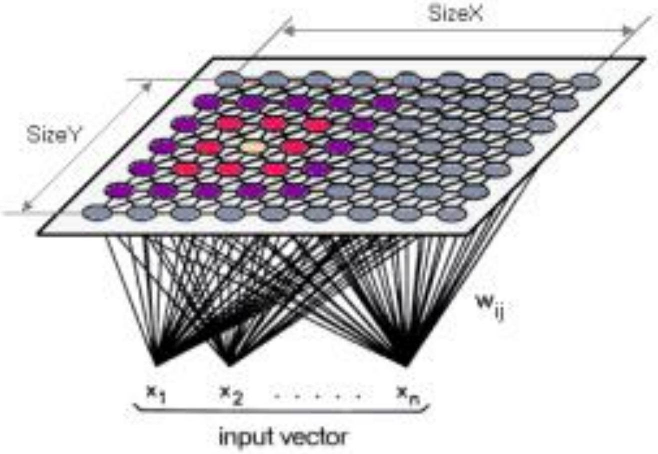


Figure 7 illustration of the dispersal of impact in a training iteration of a SOM (MultiID, 2012)

Formula for Euclidean distance: $dist(x, c) = \sum_{i=1}^{n} (x_i - c_i)^2$

> Where **n** is the number of dimensions, **x** is the training vector and **c** is a given vector in the SOM

After a single BMU has been selected, a neighborhood size is calculated and a adjustment strength is calculated as explained in the book "*Clever Algorithms: Nature-Inspired Programming Recipes".*

> "The neighbors of the BMU in the topological structure of the network are selected using a neighborhood size that is linearly decreased during the training of the network. The BMU and all selected neighbors are then adjusted toward the input vector using a learning rate that too is decreased linearly with the training cycles" (Brownlee, 2011)

The BMU is the vector marked yellow in Figure 7. The vectors in the closest vicinity of the BMU in the SOM are marked red. The red vectors will be strongly adjusted while the purple vectors in the SOM will to a lesser degree be adjusted to match the training vector. The remaining vectors in the SOM will remain unchanged. The adjustment of a vector in the training is disproportionate to the distance from the BMU. Like a rock hitting the water, the ripple effect is strongest in the center and becomes weaker the further away from the center you get.

Figure 8 shows the example SOM after 1000 training iterations of the total 10000 training iterations. Figure 8 have to a smaller degree started to resemble the input data Figure 5.

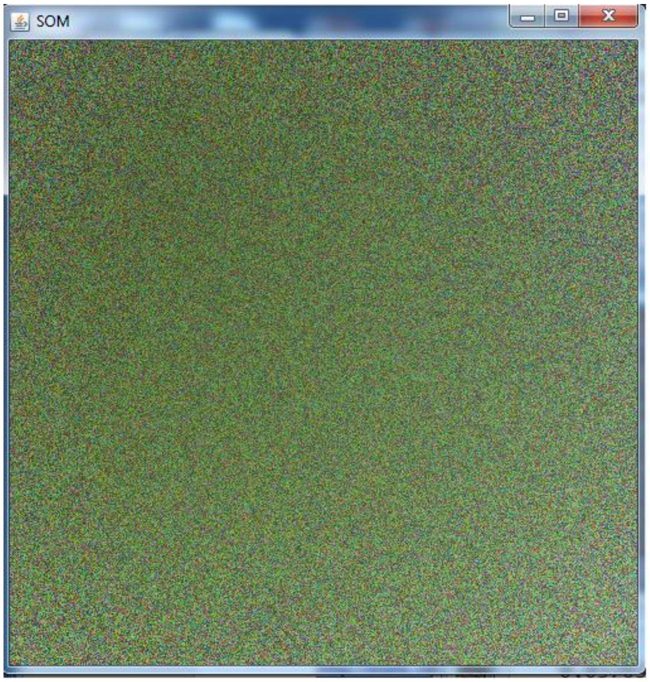


Figure 8 Example SOM after 1000 training iterations

Figure 9 and 10 shows the finished train SOM that can be used to calculate multidimensional relative distance in a two dimensional space from the input data.

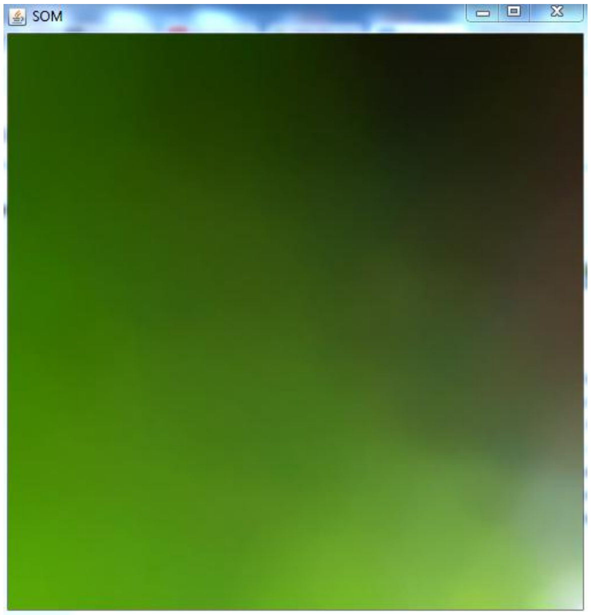


Figure 9 Finished trained SOM 1

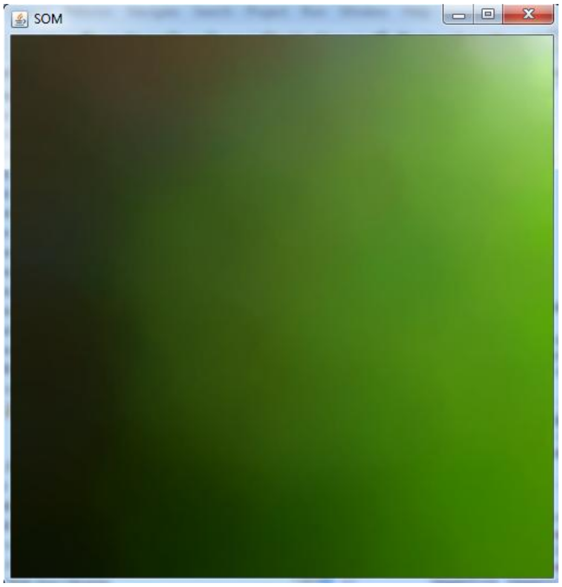


Figure 10 Finished trained SOM 2

To calculate the multidimensional relative distance between two vectors in the input data, in this example two pixels from Figure 5. This is done by finding the BMU of the two data vectors and calculating the distance between the two BMU in the trained map.

SOM is a black box process where two parallel trained maps would seldom be exactly the same. Figure 9 and 10 are two different finished trained SOMs with the same input and arguments. In this project 10 parallel maps were created. The two highest distances and the two lowest distances between two Wikipedia articles regarding language dispersal were removed from the equation, and the remaining 6 scores were added and divided by 6 to get a more correct distance between the Wikipedia articles.

Chapter 4 – Wikipedia

Data processed from Wikipedia was used in this project as the knowledge base. The three main data types from Wikipedia used in the TCSearch2 project, content, structure and metadata will be presented in detail in this chapter. In addition, criticism, systemic bias and reliability of Wikipedia will be presented. Also worth noting is that Wikipedia is divided into two parts, the most known part is the online encyclopedia, it also hosts information provided by the Wikimedia foundation that is not user contributed material regarding Wikipedia. All the references made in this thesis are from the Wikimedia foundation part of Wikipedia.

4.1 Wikipedia history

Wikipedia was launched in January 2001 by Jimmy Wales and Larry Sanger. Wikipedia was launched as a complement to the expert written peer reviewed Internet encyclopedia Nubia. Nubia in its first year only accepted 21 articles. The slow growth of Nubia made Jimmy Wales and Larry Sanger look for other models for Internet based encyclopedia. As Figure 11 shows the growth per month of Wikipedia articles had a fast growth and was at it peaked with 60 000 new articles a month in 2006. The green line shows the expected further growth of number of Wikipedia articles.

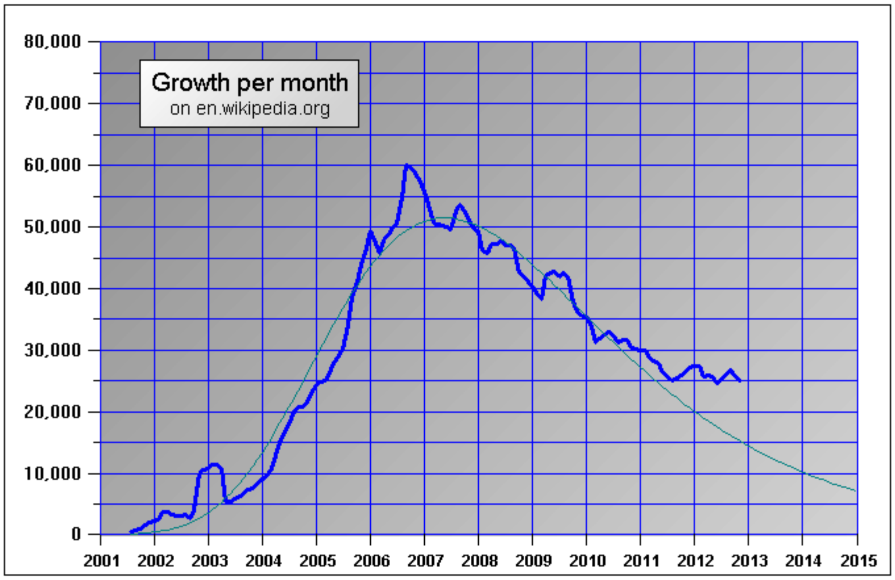


Figure 11 Wikipedia growth per month (Wikipedia G, 2012)

Figure 12 shows the total amount of Wikipedia articles in the English Wikipedia. Currently there are over four million articles in the English Wikipedia, making it the largest English encyclopedia in existence.

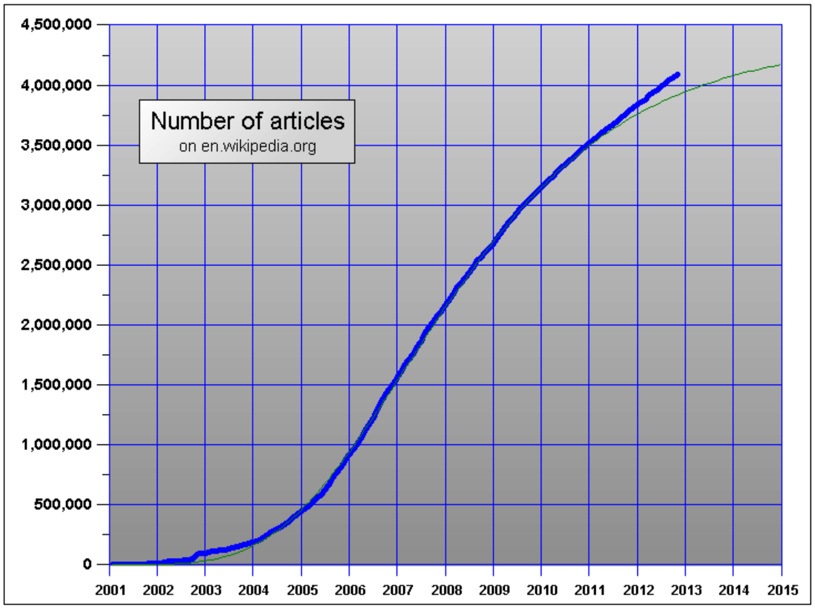


Figure 12 Wikipedia number of articles (Wikipedia G, 2012)

Figure 11 and 12 shows that growth of the numbers of articles has slowed down significantly since 2006. Wikipedia is updated thousands of times an hour from edits, which keeps Wikipedia constantly updated. The amount of edits per article has since the start of Wikipedia in 2001 just grown (Wikimedia Statistics, 2012). The slowdown in new articles but the increase of edits can imply an increased of quality and more features provided by the Wikimedia foundation. Wikipedia has added features such as categories, locking articles, templates, quality flagging articles and portals since its beginning in 2001 (Wikipedia A, 2012).

The number of editors have fallen and it is a sharp decline since 2005 (Meyer, 2012). The number may indicate that Wikipedia is starting to get more centralized with more rules, and this is a reaction to the increase of rules and regulations.

4.2 Wikipedia content

This section explains the different types of Wikipedia content found on Wikipedia. There are 3 large types of text content in Wikipedia: article content, portal content and template content. Article content is the content that the Wikipedia articles mainly consisted of. The article content has a special syntax that reveals structure and how the data is to be presented to the user. An example snippet from the Wikipedia syntax of the Wikipedia article “Semantic Web” history section:

> *== History ==*
>
> *The concept of the ''Semantic Network Model '' was coined in the early sixties by the cognitive scientist [[Allan M. Collins]], linguist [[M. Ross Quillian]] and psychologist [[Elizabeth F. Loftus]] in various publications,<ref name='Collins1969'/><ref*

> *name='Collins1970'/><ref name='Collins1975'/><ref name='M. Ross Quillian'/><ref name='M. Ross Quillian2'/> as a form to represent semantically structured knowledge. It extends the network of [[hyperlink]]ed human-readable [[web pages]] by inserting machine-readable [[metadata]] about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users. The term was coined by [[Tim Berners-Lee]],<ref name="Berners-Lee"/> the inventor of the World Wide Web and director of the [[World Wide Web Consortium]] ("[[W3C]]"), which oversees the development of proposed Semantic Web standards. He defines the Semantic Web as "a web of data that can be processed directly and indirectly by machines."*

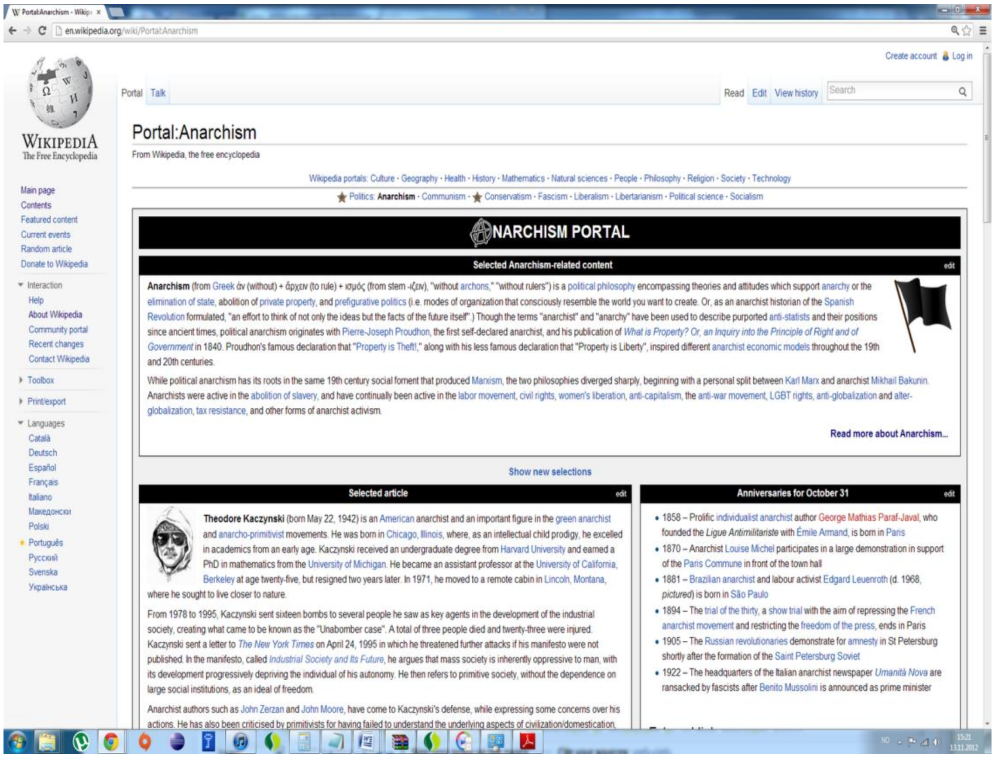
This Wikipedia syntax would be presented as such after being parsed to HTML.

> ## *History*
>
> *The concept of the Semantic Network Model was coined in the early sixties by the cognitive scientist Allan M. Collins, linguist M. Ross Quillian and psychologist Elizabeth F. Loftus in various publications,[1][2][3][4][5] as a form to represent semantically structured knowledge. It extends the network of hyperlinked human-readable web pages by inserting machine-readablemetadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users. The term was coined byTim Berners-Lee,[6] the inventor of the World Wide Web and director of the World Wide Web Consortium ("W3C"), which oversees the development of proposed Semantic Web standards. He defines the Semantic Web as "a web of data that can be processed directly and indirectly by machines."*

In some articles there are additional text boxes with information that links to a portal or a template. Definition of a portal is:

> “Portals are pages intended to serve as Main Pages for specific topics or areas” (Wikipedia C, 2012)

Figure 13 shows the portal for Anarchism. There are only around 500 portals created in the English Wikipedia, with the limited amount of portals this information was not used in this project.



Portal:Anarchism

From Wikipedia, the free encyclopedia

Wikipedia portals: Culture · Geography · Health · History · Mathematics · Natural sciences · People · Philosophy · Religion · Society · Technology

Politics: **Anarchism** · Communism · Conservatism · Fascism · Liberalism · Libertarianism · Political science · Socialism

NARCHISM PORTAL

Selected Anarchism-related content

Anarchism (from Greek ἀν (without) + ἄρχειν (to rule) + ισμός (from stem -ιζειν), "without archons," "without rulers") is a political philosophy encompassing theories and attitudes which support anarchy or the elimination of state, abolition of private property, and prefigurative politics (i.e. modes of organization that consciously resemble the world you want to create. Or, as an anarchist historian of the Spanish Revolution formulated, "an effort to think of not only the ideas but the facts of the future itself".) Though the terms "anarchist" and "anarchy" have been used to describe purported anti-statists and their positions since ancient times, political anarchism originates with Pierre-Joseph Proudhon, the first self-declared anarchist, and his publication of *What is Property? Or, an Inquiry into the Principle of Right and of Government* in 1840. Proudhon's famous declaration that "Property is Theft!," along with his less famous declaration that "Property is Liberty", inspired different anarchist economic models throughout the 19th and 20th centuries.

While political anarchism has its roots in the same 19th century social foment that produced Marxism, the two philosophies diverged sharply, beginning with a personal split between Karl Marx and anarchist Mikhail Bakunin. Anarchists were active in the abolition of slavery, and have continually been active in the labor movement, civil rights, women's liberation, anti-capitalism, the anti-war movement, LGBT rights, anti-globalization and alter-globalization, tax resistance, and other forms of anarchist activism.

Read more about Anarchism...

Show new selections

Selected article

Theodore Kaczynski (born May 22, 1942) is an American anarchist and an important figure in the green anarchist and anarcho-primitivist movements. He was born in Chicago, Illinois, where, as an intellectual child prodigy, he excelled in academics from an early age. Kaczynski received an undergraduate degree from Harvard University and earned a PhD in mathematics from the University of Michigan. He became an assistant professor at the University of California, Berkeley at age twenty-five, but resigned two years later. In 1971, he moved to a remote cabin in Lincoln, Montana, where he sought to live closer to nature.

From 1978 to 1995, Kaczynski sent sixteen bombs to several people he saw as key agents in the development of the industrial society, creating what came to be known as the "Unabomber case". A total of three people died and twenty-three were injured. Kaczynski sent a letter to *The New York Times* on April 24, 1995 in which he threatened further attacks if his manifesto were not published. In the manifesto, called *Industrial Society and Its Future*, he argues that mass society is inherently oppressive to man, with its development progressively depriving the individual of his autonomy. He then refers to primitive society, without the dependence on large social institutions, as an ideal of freedom.

Anarchist authors such as John Zerzan and John Moore, have come to Kaczynski's defense, while expressing some concerns over his actions. He has also been criticised by primitivists for having failed to understand the underlying aspects of civilization/domestication,

Anniversaries for October 31

- 1858 – Prolific individualist anarchist author George Mathias Paraf-Javal, who founded the *Ligue Antimilitariste* with Émile Armand, is born in Paris
- 1870 – Anarchist Louise Michel participates in a large demonstration in support of the Paris Commune in front of the town hall
- 1881 – Brazilian anarchist and labour activist Edgard Leuenroth (d. 1968, *pictured*) is born in São Paulo
- 1894 – The trial of the thirty, a show trial with the aim of repressing the French anarchist movement and restricting the freedom of the press, ends in Paris
- 1905 – The Russian revolutionaries demonstrate for amnesty in St Petersburg shortly after the formation of the Saint Petersburg Soviet
- 1922 – The headquarters of the Italian anarchist newspaper *Umanità Nova* are ransacked by fascists after Benito Mussolini is announced as prime minister

Figure 13 Wikipedia portal for Anarchism

Templates are one of the most common additions to a Wikipedia article. Almost 1 in 5 Wikipedia articles contains a form of template. The definition of a template is:

> "A template is a Wikipedia page created to be included in other pages. Templates usually contain repetitive material that might need to show up on any number of articles or pages. They are commonly used for boilerplate messages, standard warnings or notices, infoboxes, navigational boxes and similar purposes." (Wikipedia D, 2012)

Several templates are scripts, such as a currency converter template. The INRConvert template for example is a script that converts currency between Indian Rupees and US Dollars. (Wikipedia B. 2012). Table 4 shows a few of the different parameters and results of the INRConvert template.

Table 4 Currency converter from Indian Rupees to US Dollar, the INRConvert template (Wikipedia B. 2012).

INRConvert syntax	Results
{{INRConvert|1}}	₹1 (US$0.02)
{{INRConvert|36|b|-2}}	₹36 billion (US$700 million)
{{INRConvert|53|m|0|nolink=yes}}	Rs. 53 million (US$ 1 million)

No template information was used in this project since it is not possible to predict the result of the template by looking at the syntax without knowing the backend logic of the script.

4.3 Wikipedia structure

One of the main reason researchers has seen Wikipedia as a potential general knowledge base is the large amount of links between Wikipedia articles. There is only 50 000 Wikipedia articles of the almost 4 million Wikipedia articles that do not contain any links to other Wikipedia articles.

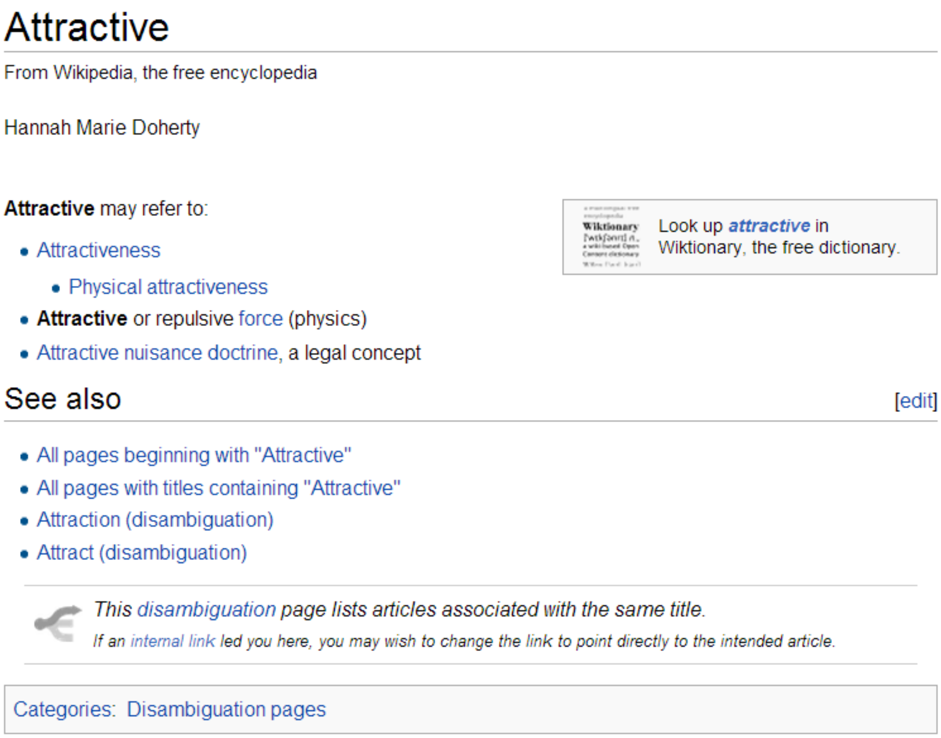
In addition to inter linkage of structure, Wikipedia introduced categories in 2004 (Suchecki, Salah, Gao & Scharnhors, 2012). Categories have a tree structure that contains categories and articles in a hierarchy.

4.4 Wikipedia metadata

Wikipedia have a record of the usages of Wikipedia articles with information such as page view counts. Dating back to December of 2007 there is a record of daily usage of the Wikipedia articles. As the usage of Wikipedia grew the records of Wikipedia page count is now divided into hourly records for page view counts.

4.5 Wikipedia Criticism

Wikipedia being an open encyclopedia where everybody can edit the content it became vulnerable to vandalism, misinformation and disagreement between editors. Often humor seems to be the motivation behind the vandalism such as seen in Figure 14.



Attractive

From Wikipedia, the free encyclopedia

Hannah Marie Doherty

Attractive may refer to:

- Attractiveness
  - Physical attractiveness
- **Attractive** or repulsive force (physics)
- Attractive nuisance doctrine, a legal concept

Look up ***attractive*** in Wiktionary, the free dictionary.

See also [edit]

- All pages beginning with "Attractive"
- All pages with titles containing "Attractive"
- Attraction (disambiguation)
- Attract (disambiguation)

This disambiguation page lists articles associated with the same title.

If an internal link led you here, you may wish to change the link to point directly to the intended article.

Categories: Disambiguation pages

Figure 14 Print screen performed the 21 of November 13:40 GMT Figure 14 of the Wikipedia article Attractive

There also exist very serious cases of vandalism with the spread of racial and sexual kind (Newby, 2012). To protect several particularly sensitive topics such as the holocaust, Hitler, Al-Qaeda, and religious pages several hundreds of different Wikipedia articles are to a degree protected or locked. When a Wikipedia article protected a Wikipedia administrator have to peer review any change of the Wikipedia article before it is published (Wikipedia E, 2012).

4.6 Reliability of Wikipedia

With the growth of Wikipedia it became used for educational purposes even in higher education. Students and researchers started to use Wikipedia as a source in academic papers. A strong opposition was formed against the use of Wikipedia as a citation source. In 2007 several Universities worldwide banned Wikipedia to be used as the single source of information in academic work (McHenry, 2004; Jaschik, 2011). The lack of personal responsibility and proof of academic credentials of the writer(s) of a Wikipedia article was one of the reasons for the ban of Wikipedia articles in academic work (Cohen, 2007). Even one of the founders of Wikipedia Jimmy 'Jimbo' Wales, said in 2006 that students should not use Wikipedia as a source of information in academic work (Orlowski, 2006).

In November 2012 a study of significant size was conducted named “Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources”. The conclusion of the study was as following.

> “The quality of information on depression and schizophrenia on Wikipedia is generally as good as, or better than, that provided by centrally controlled websites, Encyclopaedia Britannica and a psychiatry textbook.” (Reavley, Mackinnon, Morgan, Alvarez-Jimenez, Hetrick, Killackey, Nelson, Purcell, Yap & Jorm, 2012)

Several studies have reached the similar conclusions. Wikipedia with its hordes of users and contributors has an advantage in fields under constant change and need constantly updating the information (Brown, 2011). While several studies points to that Wikipedia in general is more reliable than several more established resources, the studies also conclude that in a few cases the information found on Wikipedia is of considerable lower quality. In the later years, studies have shown a positive trend regarding the reliability of Wikipedia, where in general Wikipedia even outperforms widely used academic textbooks (Reavley et al, 2012; Brown, 2011).

4.7 Coverage of topics and systemic bias

Not all categories have the same amount of articles, and the amount of articles in the different categories seems to reflect the interest of the Wikipedia contributors (Suchecki et al, 2012). The study “Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage” found the following pattern in the coverage of topics in Wikipedia.

> “Wikipedia’s omissions follow a predictable pattern: coverage is best on topics that are more recent or prominent. Using state legislators as an example, I find that the depth of Wikipedia’s coverage improves for legislative leaders, longtime politicians, and legislators with larger constituencies” (Brown, 2011)

The group of Wikipedia contributors is a homogeneous group and is very dominated by English speaking men (Wikimedia, 2012). This may impact what type of information is provided by Wikipedia. As Figure 15 shows, there is a systemic bias to some categories. In addition some categories or subcategories have seen a large increase of Wikipedia articles, while other categories have had little or no increase in amounts of Wikipedia articles. Difference in growth of the categories may be the result of the Wikipedia guidelines on what types of information is notable enough to become a Wikipedia article (Wikipedia F, 2012). Some yearly events will have a new Wikipedia article each year, while a person will in most cases only have one Wikipedia article. One example is the 100 yearly added American beauty pageants competition Wikipedia articles. If this trend continues Wikipedia’s quality as a general knowledge base may decline.

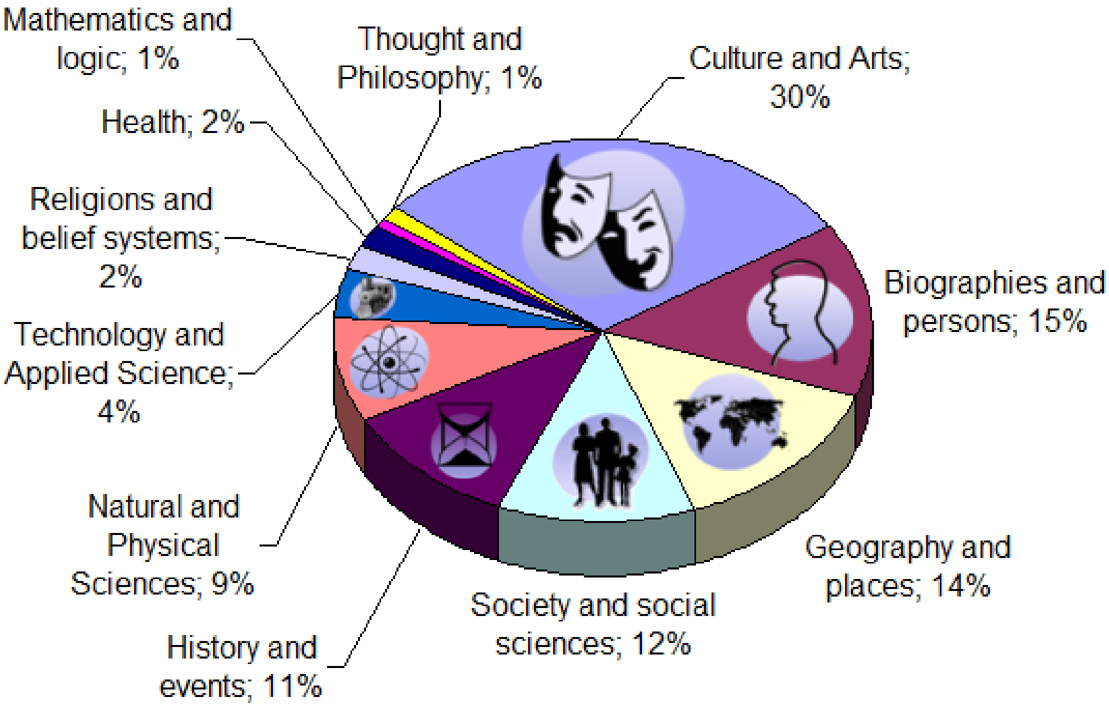


Figure 15 articles divided into categories from 2008 (Kittur, Chi,& Suh, 2009)

Chapter 5 - System development

This project development was divided into two specific tasks. The first task was to develop a program that processed the information provided by Wikipedia into a general knowledge base that could be used for the query augmentation process. The second task was to develop the search engine core TCShearh2. A total of 4 different search engines were developed based on the TCSearch2 core. Main steps for development included the following:

1 Create a knowledge base.
2 Create a term to term search engine.
3 Create a query augmentation search engine.
4 Create an evaluation frontend.
5 Create a prototype frontend.

5.1 Data gathering

There were two methods used to acquire the data used in the development of this project. The Wikipedia article dump was manually downloaded from the Wikipedia backup dump site http://dumps.wikimedia.org/backup-index.html. Data regarding usage of the Wikipedia articles were divided into thousands of files, this would have been a time consuming task to be performed manually. To automate the task of downloading the Wikipedia article usage files, a simple web crawler was created with the aim of downloading automatically all the usage statistics files from http://dumps.wikimedia.org/other/pagecounts-raw/

5.2 Development method

System development is inherent with risk. Learning several new technologies combined with the time constraints of this project was risky. Learning new technologies is a challenging, and time consuming task that is hard to correctly estimate correctly the time it takes to accomplish. Thus, planning with any degree of accuracy is very difficult, of often pointless. In this project, an agile development method Rapid application development(RAD) was used. RAD is centered on minimal planning, but rapid prototyping. After each iteration or sprint, there is a prototype or a mockup created. The prototype or mockup is evaluated, and if it does not perform the task as intended or wanted it is thrown away. If the prototype performed the task as intended it will be further developed.

The system development methodology chosen had an impact on the evaluation process of the system. There has been raised criticism against RAD as a development methodology, claiming that its minimalistic planning makes it hard to control that the development of large (Gerber, van der Merwe & Alberts, 2007). While there was very little low level planning, high level planning and constant evaluation was needed to create a cohesive working product (Pfleeger & Atlee, 2005, page 190-194).

5.3 Iterations

As discussed in the introduction, there are two distinct tasks in the development process of the TCSearch2 system. While most of the resources in this project were used in the first task creating the knowledge base, both systems were developed in parallel. A detailed account of the development is given in Table 5.

Table 5 Iterations while development of the TCSearch2 project

Iteration number	Duration	Activities
1	January 02.- January 30.	Setup of computer environment.
2	January 31. - February 15.	Manually downloading the Wikipedia article content dump. Developed a web crawler for downloading the hundreds of Wikipedia user statistic data logs.
3	February 16. - Mars 15.	Continuing downloading user statistics data and uncompressing the data to disk. Setup of MySQL database and import of Wikipedia database dump files. Started search engine development.
4	Mars 16. - April 15.	Creating the Self-organizing map application. Error checking computer and re-import the database due to corrupted data due to faulty memory. Optimizing MySQL database configuration based on the hardware. Continuing the search engine development.
5	April 16. - May 15.	Processing the Wikipedia database dump and creating the logic for the knowledge base and search engine logic. The first prototype of the search index and prototype for the first knowledge base prototype. Creating a Wikipedia article structure with the different sections corresponding zone weighting. Continuing the search engine development.
6	May 16. - June 9.	Speeding up the term normalization process of Wikipedia content dump and developed the first working search engine prototype. Adding redirect tiles processing capability to improve the link connectivity calculation, and adding disambiguates processing capability. Adding TF-

		IDF ranking in the search engine.
Summer holiday	June 10. - August 10.	Summer internship at Yahoo! Technology Norway
7	August 11.- September 01.	Shifting the search engine storage from memory to a database. Improving the term index by adding the terms used to describe the link in other pages. Training the SOM with the user statistic data and link connection values. Created the first augmented query algorithm.
8	September 01. - September 15.	Improving the search time by improving the search index hashing and upgrading the hardware for the database. Adding term zone weighting with italic and bold structure. Developed the augmented query algorithm with user statistic metadata.
9	September 16. - November 15.	Creating the evaluation and a prototype web interface. Improve the title zone weight value. Adding an implementation of PageRank to one of the search engines.

5.4 Query Augmentation implementation

While the practice of query augmentation is well known and tested, there are several different possible approaches of the implementation. This section describes in detail this projects implementation of the query augmentation process.

Figure 16 shows a representation of the documents in the knowledge base randomly spread out in a two dimensional space with each dot represents a Wikipedia article. Each document in the knowledge base is stored as an inverted index with the containing terms and value of the terms in this document.

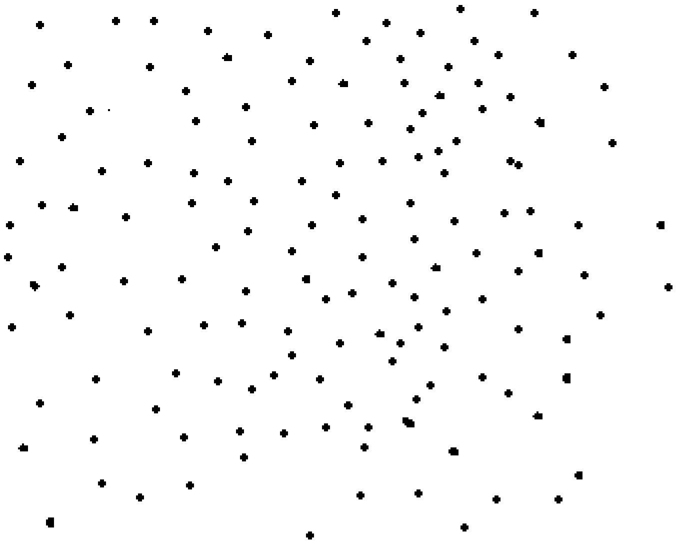


Figure 16 A representations of the documents in the knowledge base

The first step in the query augmentation process is the stimulation of the documents in the knowledge base based on the input of the user. The method for calculating the documents stimulation is the sum of the TF-IDF with a combination of term zone weighting.

In Figure 17 the documents simulation value is marked in red. The strength of the stimulation is shown by the size of the red circles.

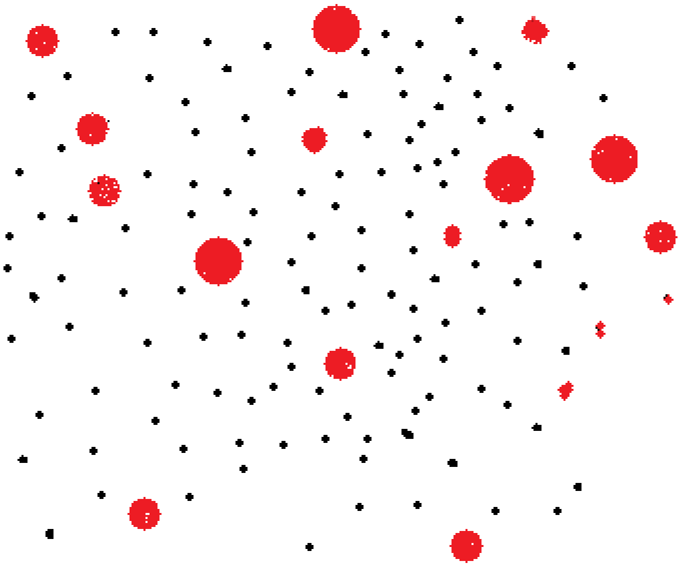


Figure 17 Knowledge base documents stimulated based on user input

With the interconnected nature of Wikipedia, there are tens of millions of internal links between documents in the knowledge base. Links between the documents also have different values associated to them. In the same manner as terms, links were processed using TF-IDF and zone score weighting. This resulted in an additional inverse index for each document in the knowledge base containing the links found in the document and the values associated with the links.

It was in this step the language dispersal distance between the BMU's of the documents was used to adjust the value of the links in the inverse index. In addition, documents with a high amount of readers would receive a minor bonus, to boost popular Wikipedia articles within a topic.

Using the link to further stimulate the documents in the knowledge base, documents not previously simulated, could be stimulated by connectivity of documents stimulated in the first step of the stimulation process.

In Figure 18 the links between the documents are drawn, and the width of the lines between the dots is proportional to the connection value of the links between the documents.

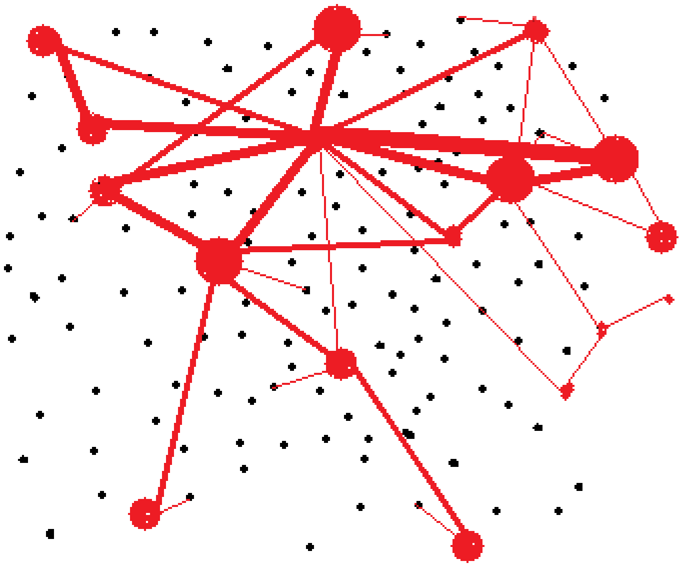


Figure 18 Links between documents in the knowledge base

The final stimulation value of a document is the sum of the connection values based on connection to stimulated documents, added to the initial stimulation value. Thus, even documents with low initial term stimulation, or even no initial term stimulation, can be highly stimulated through connectivity.

Figure 19 shows the stimulation after the final step in the stimulation of the documents in the knowledge base.

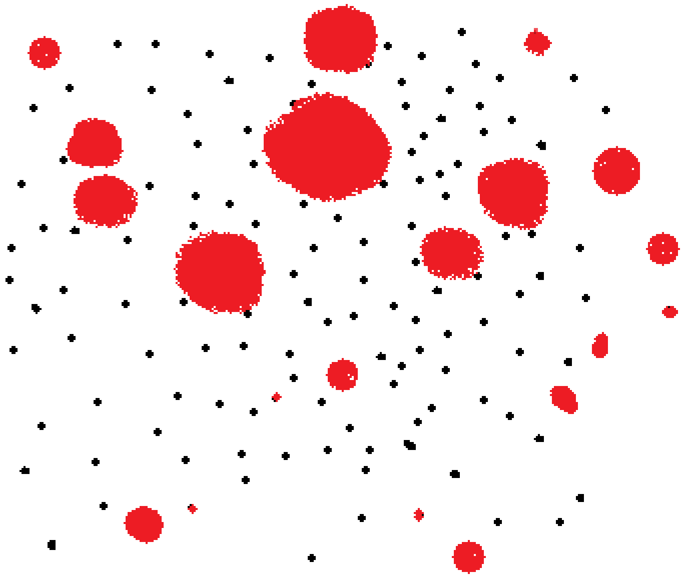


Figure 19 Final stimulation value of documents

The aim of the query augmentation process in this project is to extract terms with a semantic connection to the original input terms. All terms from all stimulated documents were added in the extended query table based on the terms value in the document and the document's final stimulation value. The selected amount of top valued terms in the extended query table was then used as the augmented query. In order to keep the original input the most significant part of the augmented query, the original terms value in the extended query table would increased by the highest value in the extended query table. Thus, the original input will always have a higher impact in the augmented query than any other term in the query.

Domain knowledge obtained during development and empirical trials were used to adjust the amount of terms to include in the augmented query. A limited trial performed suggested that large topics that are not ambiguities have a high amount of related terms that are chosen correctly by the query augmentation process. An example is the query "*Bachelor degree".* With an augmented query of 30 extended terms, all of the terms have a semantic relation to the input in this example as seen below.

```
{bachelor=0.2469702627489005,
degre=0.24234324192452933,
doctor=0.027932585968155742,
undergradu=0.026399468118924962,
academ=0.025838137805089104,
master=0.025824600900951417,
scienc=0.025766983000533695,
educ=0.021274040275343718,
univers=0.0209639456794497,
postgradu=0.019370290686213498,
studi=0.017953082804601988,
bologna=0.017901405957548258,
diplom=0.017468673139735852,
student=0.017425430624957727,
medicin=0.01742233228152799,
diploma=0.017020522537424698,
higher=0.016848743374262457,
magist=0.01671437257321428,
graduat=0.016391232071660416,
cours=0.016275720467595827,
laurea=0.016046882443104468,
baccalaureus=0.01579262809966535,
art=0.015264812357541247,
law=0.01525856463787143,
engin=0.01453528164839405,
requir=0.014309127208834423,
qualif=0.014006761346815187,
research=0.013786662097977718,
institut=0.013650848979351595,
profession=0.013242947631562437}
```

A second example is with the query “finger”, the augment query was of a far lower quality than the “*Bachelor degree”* example. Several terms in the extended query were not relevant terms, such as “or”, “are” and “can”, as seen in the augmented query below. One reason for the low quality of the augmented query is that the input “finger” is ambiguities. Being an ambiguities term means that the term have several different meanings, as seen on http://en.wikipedia.org/wiki/Finger_(disambiguation).

```
{finger=0.724035785995667,
thumb=0.03190755829444981,
hand=0.022483854750906217,
zinc=0.01829444773374983,
digitus=0.01623315106149635,
instrument=0.01584456719557701,
digit=0.01548277278180607,
human=0.013483005151549788,
music=0.013336221526475506,
muscl=0.01299177856684266,
or=0.012820568049133936,
index=0.012286521348189177,
anatomi=0.012189178421362782,
extensor=0.012028235447570697,
are=0.01137169782286619,
ring=0.011362709946877294,
can=0.011351435933156013,
use=0.01125070988405395,
string=0.0106973872374442,
guitar=0.010547675059336833}
```

Lowering the amount of terms added in the augmented query to 10 significantly reduced non-relevant terms.

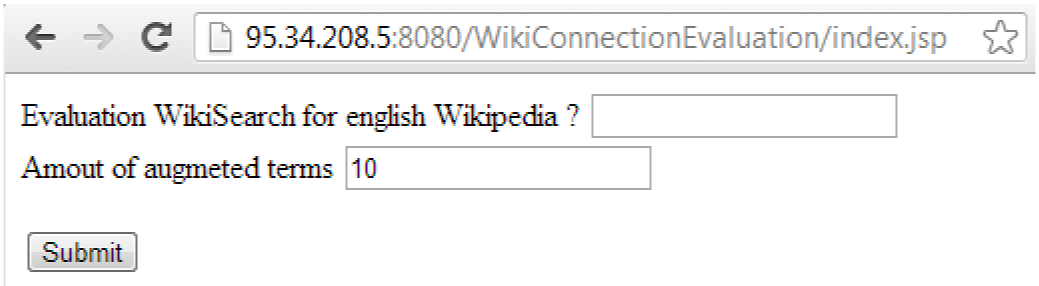
```
{finger=0.8189580333929527,
thumb=0.036090690124189514,
hand=0.02543152400205416,
zinc=0.020692879036964595,
digitus=0.01836134308031749,
instrument=0.017921815249239484,
digit=0.01751258901025331,
human=0.01525064865123441,
music=0.015084621458587,
muscl=0.014695021477076346}
```

The possible negative impact of including non-descriptive terms can be both reduced result quality and increased computational needs. The standard amount of added terms in the augmented query was set to 10. Lowering the amount of extended terms in the augmented queries from 30 prevented a possible decline in result quality due the inclusion of non-descriptive terms in the augmented query. Users would also be able to manually set the amount of terms added to the augmented query for each query.

5.5 Graphical user interface development

In this project there was only minimal focus on the graphical user interface(GUI). This was due to the project's intended purpose. As stated in the introduction, this project is intended to be a plug-in module to existing search engines. Only GUI development for evaluation and prototype purposes was developed, with the prototype being developed for researchers to compare results with the TCSearch2 search engines. I will first present the evaluation interface and afterwards present the prototype interface.

The main goal of the GUI development was to create an evaluation interface for the test subjects. The intro page for the evaluation search engine was created to be as simple and uncomplicated as possible, as seen in Figure 20.



95.34.208.5:8080/WikiConnectionEvaluation/index.jsp

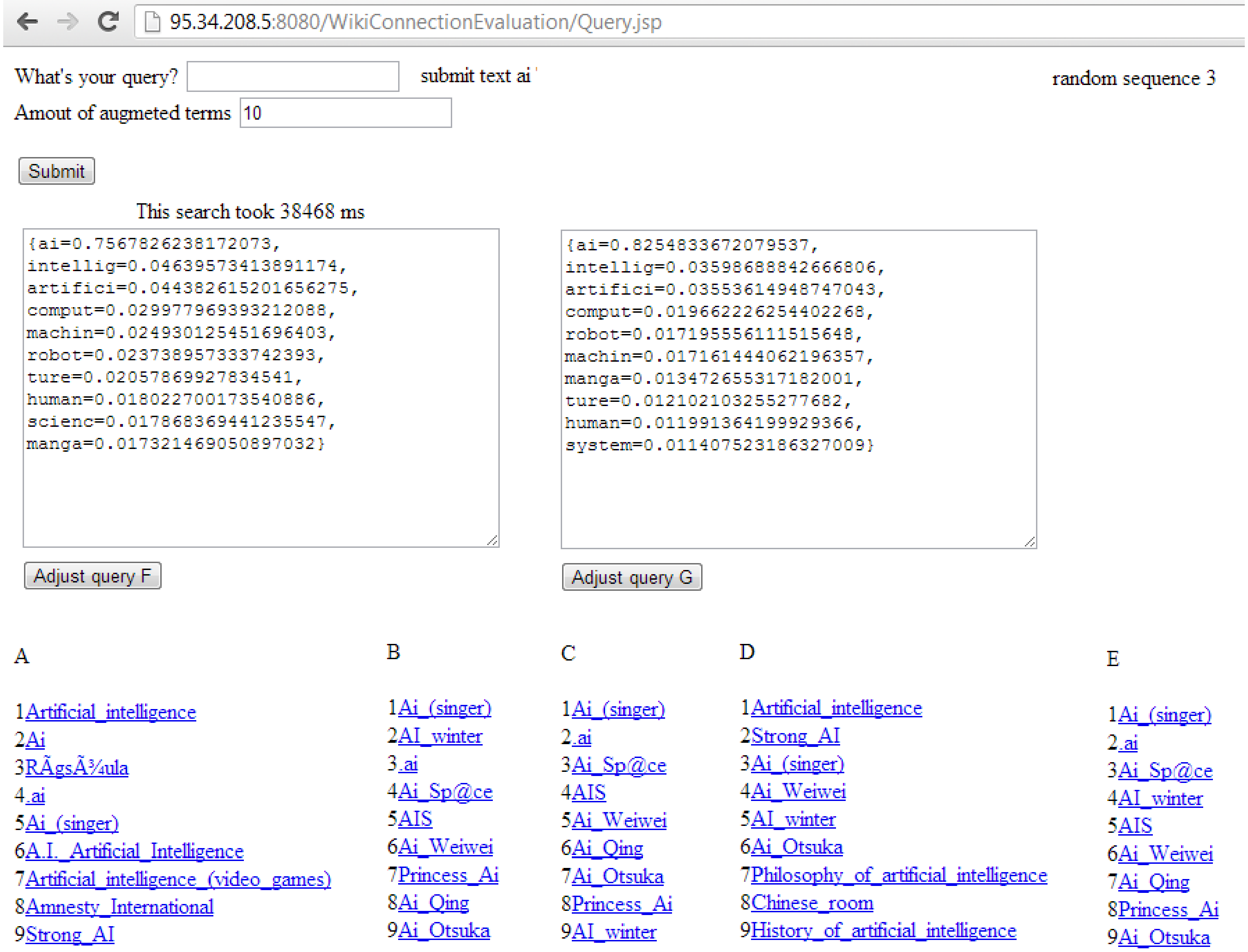
Evaluation WikiSearch for english Wikipedia ?

Amout of augmeted terms 10

Figure 20 Intro to the evaluation interface

After the test subject entered a query, a new layout with the different search results and two different augmented queries text boxes is shown. This was created to resemble popular existing search engine. As seen in Figure 21, the different search results are given code names.

Additionally, the different search engines appear in a random sequence for the test subjects. But for the evaluator, the number in the top right corner is the key to decipher the different search engine results.



What's your query? submit text ai

random sequence 3

Amout of augmeted terms 10

This search took 38468 ms

```
{ai=0.7567826238172073,
intellig=0.04639573413891174,
artifici=0.044382615201656275,
comput=0.029977969393212088,
machin=0.024930125451696403,
robot=0.023738957333742393,
ture=0.02057869927834541,
human=0.018022700173540886,
scienc=0.017868369441235547,
manga=0.017321469050897032}
```

```
{ai=0.8254833672079537,
intellig=0.03598688842666806,
artifici=0.03553614948747043,
comput=0.019662226254402268,
robot=0.017195556111515648,
machin=0.017161444062196357,
manga=0.013472655317182001,
ture=0.012102103255277682,
human=0.011991364199929366,
system=0.011407523186327009}
```

A

1Artificial_intelligence
2Ai
3RÃgsÃ¾ula
4.ai
5Ai_(singer)
6A.I._Artificial_Intelligence
7Artificial_intelligence_(video_games)
8Amnesty_International
9Strong_AI

B

1Ai_(singer)
2AI_winter
3.ai
4Ai_Sp@ce
5AIS
6Ai_Weiwei
7Princess_Ai
8Ai_Qing
9Ai_Otsuka

C

1Ai_(singer)
2.ai
3Ai_Sp@ce
4AIS
5Ai_Weiwei
6Ai_Qing
7Ai_Otsuka
8Princess_Ai
9AI_winter

D

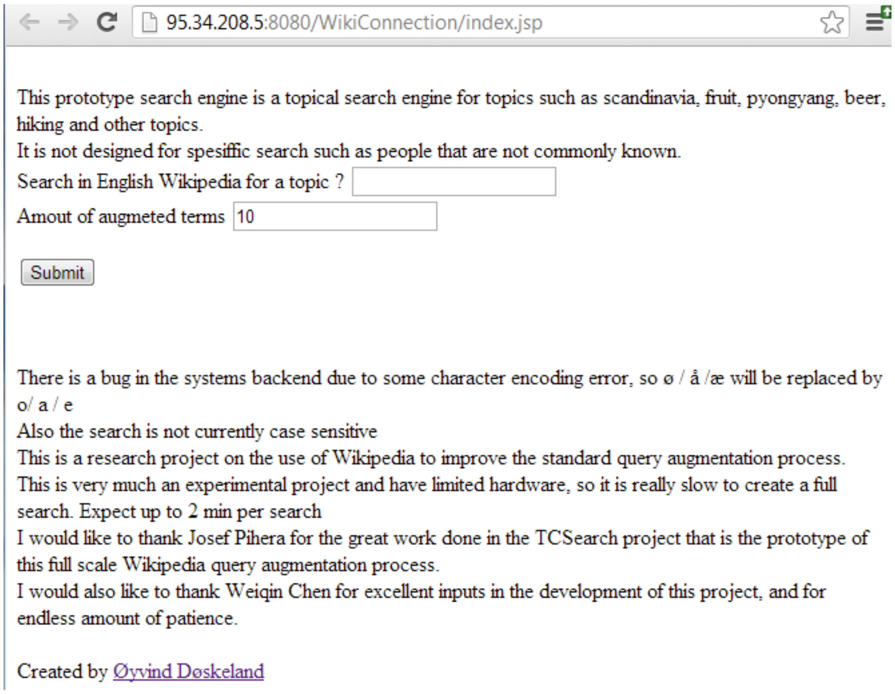
1Artificial_intelligence
2Strong_AI
3Ai_(singer)
4Ai_Weiwei
5AI_winter
6Ai_Otsuka
7Philosophy_of_artificial_intelligence
8Chinese_room
9History_of_artificial_intelligence

E

1Ai_(singer)
2.ai
3Ai_Sp@ce
4AI_winter
5AIS
6Ai_Weiwei
7Ai_Qing
8Princess_Ai
9Ai_Otsuka

Figur 21 Main evaluation page

The prototype has a similar outline to the evaluation pages. The front page of the prototype includes a minimal introduction, list of known bugs, acknowledgments and a link to the social website LinkedIn for potential communication to other users or interested parties. See Figure 22.



95.34.208.5:8080/WikiConnection/index.jsp

This prototype search engine is a topical search engine for topics such as scandinavia, fruit, pyongyang, beer, hiking and other topics.
It is not designed for spesiffic search such as people that are not commonly known.
Search in English Wikipedia for a topic ?
Amout of augmeted terms 10

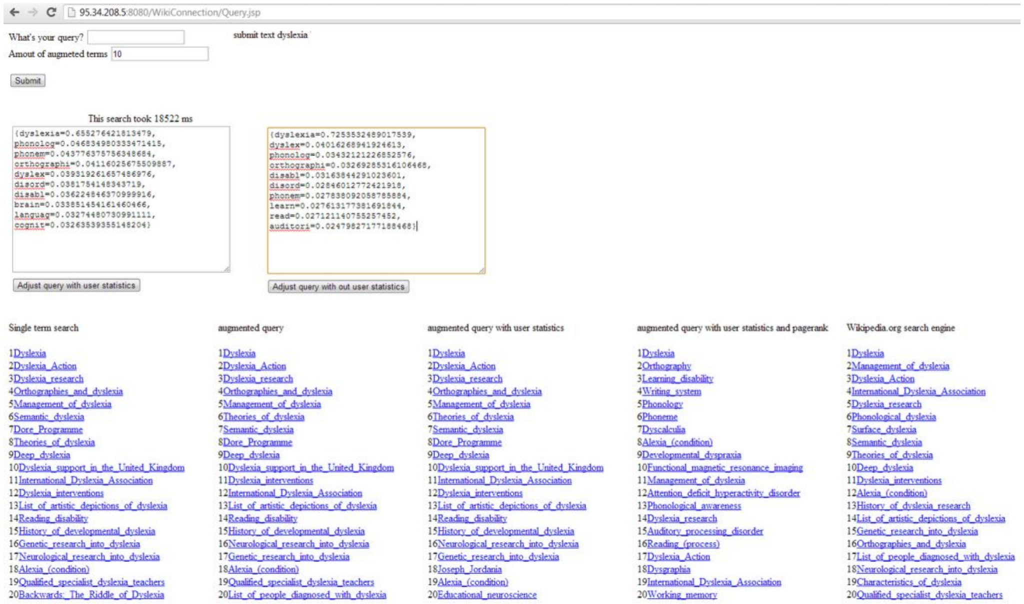
Submit

There is a bug in the systems backend due to some character encoding error, so ø / å /æ will be replaced by o/ a / e
Also the search is not currently case sensitive
This is a research project on the use of Wikipedia to improve the standard query augmentation process.
This is very much an experimental project and have limited hardware, so it is really slow to create a full search. Expect up to 2 min per search
I would like to thank Josef Pihera for the great work done in the TCSearch project that is the prototype of this full scale Wikipedia query augmentation process.
I would also like to thank Weiqin Chen for excellent inputs in the development of this project, and for endless amount of patience.

Created by Øyvind Døskeland

Figure 22 Intro to prototype

After a user inserts an input and press submit, the user is presented with the results, Figure 23.



95.34.208.5:8080/WikiConnection/Query.jsp

What's your query?
Amout of augmeted terms 10

submit text dyslexia

Submit

This search took 18522 ms

```
{dyslexia=0.655276421813479,
phonolog=0.04683498033471415,
phonem=0.043776375756348684,
orthographi=0.04116025675509887,
dyslex=0.039319261657486976,
disord=0.0381754148343719,
disabl=0.036224846370999916,
brain=0.033851454161460466,
languag=0.03274480730991111,
cognit=0.03263539355148204}
```

Adjust query with user statistics

```
{dyslexia=0.7253532489017539,
dyslex=0.04016268941924613,
phonolog=0.03432121226852576,
orthographi=0.03269285316106468,
disabl=0.03163844291023601,
disord=0.02846012772421918,
phonem=0.027838092058785884,
learn=0.027613177381691844,
read=0.027121140755257452,
auditori=0.02479827177188468}
```

Adjust query with out user statistics

Single term search	augmented query	augmented query with user statistics	augmented query with user statistics and pagerank	Wikipedia.org search engine
1Dyslexia	1Dyslexia	1Dyslexia	1Dyslexia	1Dyslexia
2Dyslexia_Action	2Dyslexia_Action	2Dyslexia_Action	2Orthography	2Management_of_dyslexia
3Dyslexia_research	3Dyslexia_research	3Dyslexia_research	3Learning_disability	3Dyslexia_Action
4Orthographies_and_dyslexia	4Orthographies_and_dyslexia	4Orthographies_and_dyslexia	4Writing_system	4International_Dyslexia_Association
5Management_of_dyslexia	5Management_of_dyslexia	5Management_of_dyslexia	5Phonology	5Dyslexia_research
6Semantic_dyslexia	6Theories_of_dyslexia	6Theories_of_dyslexia	6Phoneme	6Phonological_dyslexia
7Dore_Programme	7Semantic_dyslexia	7Semantic_dyslexia	7Dyscalculia	7Surface_dyslexia
8Theories_of_dyslexia	8Dore_Programme	8Dore_Programme	8Alexia_(condition)	8Semantic_dyslexia
9Deep_dyslexia	9Deep_dyslexia	9Deep_dyslexia	9Developmental_dyspraxia	9Theories_of_dyslexia
10Dyslexia_support_in_the_United_Kingdom	10Dyslexia_support_in_the_United_Kingdom	10Dyslexia_support_in_the_United_Kingdom	10Functional_magnetic_resonance_imaging	10Deep_dyslexia
11International_Dyslexia_Association	11Dyslexia_interventions	11International_Dyslexia_Association	11Management_of_dyslexia	11Dyslexia_interventions
12Dyslexia_interventions	12International_Dyslexia_Association	12Dyslexia_interventions	12Attention_deficit_hyperactivity_disorder	12Alexia_(condition)
13List_of_artistic_depictions_of_dyslexia	13List_of_artistic_depictions_of_dyslexia	13List_of_artistic_depictions_of_dyslexia	13Phonological_awareness	13History_of_dyslexia_research
14Reading_disability	14Reading_disability	14Reading_disability	14Dyslexia_research	14List_of_artistic_depictions_of_dyslexia
15History_of_developmental_dyslexia	15History_of_developmental_dyslexia	15History_of_developmental_dyslexia	15Auditory_processing_disorder	15Genetic_research_into_dyslexia
16Genetic_research_into_dyslexia	16Neurological_research_into_dyslexia	16Neurological_research_into_dyslexia	16Reading_(process)	16Orthographies_and_dyslexia
17Neurological_research_into_dyslexia	17Genetic_research_into_dyslexia	17Genetic_research_into_dyslexia	17Dyslexia_Action	17List_of_people_diagnosed_with_dyslexia
18Alexia_(condition)	18Alexia_(condition)	18Joseph_Jordania	18Dysgraphia	18Neurological_research_into_dyslexia
19Qualified_specialist_dyslexia_teachers	19Qualified_specialist_dyslexia_teachers	19Alexia_(condition)	19International_Dyslexia_Association	19Characteristics_of_dyslexia
20Backwards:_The_Riddle_of_Dyslexia	20List_of_people_diagnosed_with_dyslexia	20Educational_neuroscience	20Working_memory	20Qualified_specialist_dyslexia_teachers

Figur 23 Main page prototype

5.6 Tools

The tools in this project are divided into hardware tools and software tools. Several decisions in the developments were made on characteristics of the hardware available in this project thus the hardware used in this project is listed. In the software section the different software used in the development and deployment are presented.

5.6.1 Hardware

Large amount of data and high computational demands created the need to acquire new hardware for development and testing of the TCSearch2. Due to a natural disaster in Thailand, causing a lack of components, the development of this project was delayed in the start of this project. In addition some of the hardware parts were faulty and halted the development process for several weeks.

The hardware used in this system was as following:
CPU**:** Intel i7-3770
Memory: 36 GB
Disk: 128 GB SSD

5.6.2 Software

This section presents the software used in the development and deployment.

Java
Java is the programming language used in the development in this project. Java was chosen as the developing language due to large amount of existing Wikipedia related libraries. It is also the programming language that I am most comfortable with, and there is extensive literature and examples available. In addition to standard Java, Java Servlet Page(JSP) was used for connecting the frontend to the backend. JSP is a technology for developing dynamically generated web pages based on HTML and XML. JSP is similar to the more well known PHP but uses the Java programming language.

MySQL
The database used in this project was MySQL. MySQL is the world's most used open source relational database. It is also used by large companies such as Google and Yahoo!’s Flicker, but Google and Flickr do not use MySQL for the web search(MySQL A. 2012; MySQL B. 2012; MySQL C. 2012). MySQL was used since it is free and easy to set up. I used EasyPHP that included an installation of MySQL for this project.

Hyper Text Markup Language
For the frontend, Hyper Text Markup Language**(**HTML) was used for the interface development.

Integrated development environment (IDE)

SpringSource was used as the IDE of this project. SpringSource was chosen based on its similarity with Eclipse IDE, but SpringSource included better support for web development with integration of Apache Tomcat.

Apache Tomcat 7

For deployment of the web application, Tomcat was chosen since I had used it several times before and it was integrated into SpringSource, making testing during development considerable less time consuming.

Windows 7 Professional

Windows 7 is the most used OS in 2012 (W3School. 2012). While I am more used to Mac OS X, the cost of an Apple computer was magnitudes higher and had other minor drawbacks for the development of this project. The Professional edition was chosen since it was the version that supported the large amount of memory used in this project and it was provided free of charge through the University.

Google documents

This paper was written in Google documents. This allowed for access wherever there was an Internet connection. Google documents also allow several people to comment simulations in the same document, and allowed the supervisor of this thesis to constantly monitor and comment the progress of this thesis.

Chapter 6 - Evaluation

In this chapter the evaluation process and its results are presented. The aim of the evaluation in this project was to gather data in order to address the research questions set out in the introduction chapter. In addition, the evaluation has been an integrated part of the development process. During the development process, an evaluation was conducted after each iteration, under the precondition of a testable TCSearch2 search engine.

Information Retrieval evaluation has evolved to be a highly empirical discipline. To demonstrate differences in search engine results require a thorough and a careful evaluation process. The single most important key measure to differentiate search engines is user happiness. While there exists several quantitative methods for measuring the quality or accuracy of a result from a search engine such as recall and precision, quantitative evaluation alone cannot clearly evaluate user happiness. In addition, evaluating the distance between the expectation of the user and the result of search engines is best evaluated through qualitative methods (Manning et al, 2008, page 139).

While user happiness is the key utility measure of the quality, other forms of evaluation methodology were also used in the evaluation process. One of the reasons for choosing several different evaluation methods was to get a broader range of data that could be used for comparison between the different search engines. Also, due to the high costs connected to qualitative studies, a quantitative experiment was chosen in order to increase the amount of unique queries used in the evaluation process, making the evaluation more representative.

Query inputs used for the evaluation were extracted from several different sources. The sources used were the most searched terms on AOL and Google in 2011. In addition, a minimalistic crawler extracted terms from the Wikipedia main page based on the amount of readers of the Wikipedia article in question. Only Wikipedia articles that were in the top 10% of the most read Wikipedia articles were used in this process.

One of the aims of the evaluation was to create an experiment to test the differences between the users' expectations and the actual results of several different search engines. The evaluation took into consideration that test subjects had clearer expectations about the type of results they expected in relation to popular topics compared to less known topics.

In total 47 terms were selected for the evaluation. The 47 terms were divided into three lists. The first list consists of 10 terms that were used for the quantitative evaluation during the development of this project. The second list consists of 30 terms and was used in the final quantitative evaluation. The remaining 7 terms were used in the qualitative study. 50 random terms is the recommended minimum amounts of queries to perform a representable quantitative study (Manning et al, 2008. page 140). The method used to extract terms for the quantitative evaluation aimed at collection a list of 30 germane queries. All the 30 selected terms to be used in the evaluation were terms that were popularly used and believed to be of greater importance than 50 random terms. In addition, the total amount of different queries will be over 50 when considering both the qualitative and quantitative study.

Result quality from a query is dependent on the information need from the query. The lists provided by Google and AOL have specified the topic of each query, and the topic will then serve as the information need in the evaluation (AOL, 2012; Searches Organic SEO, 2012). With regard to Wikipedia the terms selected to be used in the evaluation was the link text of a Wikipedia article. Below is a typical HTML link.

```
<a href="/wiki/Judaism" title="Judaism">Jewish</a>
```

The Wikipedia article has the title “Judaism” but the link text is “Jewish”. The information need based on the input query “Jewish”, is centered on the content in the Wikipedia article “Judaism”.

One of the research questions is regarding the TCSearch2 search engines compare against existing search engines. Comparing the TCSearch2 with live search engines, such as Google and Wikipedia’s search engine, is challenging since both Google and Wikipedia’s search engine is under constant development, and is constantly gathering new data for their search index. Thus the results of the evaluation may not be reproducible, since the data from the external sources may have been altered. To minimize the impact of change due to a newer search index of Google and Wikipedia, queries with several Wikipedia article created after the Wikipedia dump this project used was not used in the evaluation

Information need is subjective and thus can be interpreted differently from what has been done in this study. Therefore, it was of utmost importance to publish all the evaluation data. The data collected in this project is in Appendix A and B.

6.1 Evaluation during development

The use of RAD as development methodology prescribes constant evaluation to track progress. After the 4th iteration, the evaluation started on the first functional prototype. The evaluation only evaluates the existing features in an iteration. This evaluation was only for the TCSearh2 based search engines, so no external search engines were included in this evaluation.

10 queries were randomly selected from the list extracted from AOL, Wikipedia and Google as described in the previous section. The following list contains the queries used in the evaluation during the development:

1 Youtube
2 xnxx
3 bbc
4 cnn
5 ikea
6 Japan Earthquake
7 Bridesmaids
8 Dollar

9 Italy

Only the top 20 results from each search engine were used. As seen in Figure 24, less than 3% of the traffic from a search is from results lower than the 20th result (Chitika Insights. 2010). Selecting only a fixed amount of top hits is called result pooling, and is often used for evaluation purposes (Manning et al, 2008, 159-160)

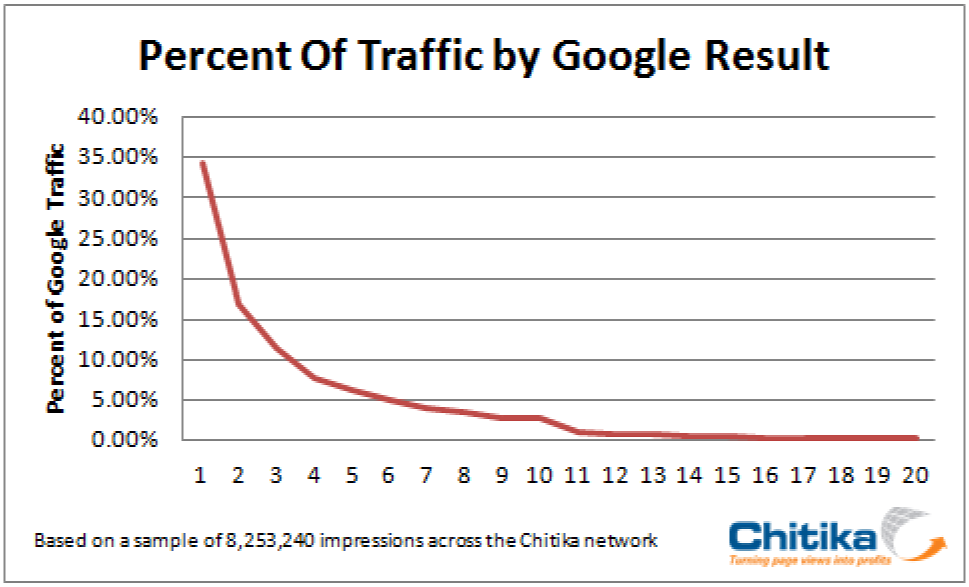


Figure 24 Percent of traffic by Google result (Chitika Insight, 2012)

Table 6 presents the result of the search simulation performed during the development. The result of this evaluating is the precision of each search engine during development. Precision is the fraction of retrieved documents that are relevant for the query; in this case the precision was the fraction of relevant documents in the top 20 results. Assessment of the relevancy of the documents in the result list was based on the information need extracted along with the queries used in the evaluation. A binary unranked classification for relevancy was used. Ideally recall would also be calculated. Recall is the amount of relevant documents found in the result list divided by the amount of relevant documents in the corpus. With the limited amount of time and resources available, this was not feasible with around 4 million documents in the corpus.

Table 6 Results of the search simulation results preformed during development

Iteration	Standard query	Standard augmented query	Augmented query with user statistic data	Augmented query with user statistics and PageRank
6	0,450	-	-	-
7	0,420	0,120	-	-
8	0,600	0,510	0,28	-

9	0,600	0,705	0,725	0,755

6.2 Final evaluation

In order to address the research questions, both qualitative and quantitative evaluation methods were used for the final evaluation. The final evaluation was performed after all development tasks were finalized.

One research question inquires whether Wikipedia, as a knowledge base, can be used for a query augmentation process in a search engine and give improved results compared to a term to term based search engine. The evaluation aim for this research question was to gather data on the result quality difference between a term to term search engine and a query augmented search engine that used Wikipedia as a knowledge base built on the same search engine core.

The second research question inquires whether including the usage statistics data of the different Wikipedia articles can have a positive impact in the query augmentation process when using Wikipedia as a knowledge base. Evaluation aim for this research question was to gather data on the result quality difference in results from an augmented queries search engine, with and without the use of Wikipedia article usage data.

The last research question inquires whether the results from search engines developed in this project are comparable to the leading web and domain search engines. Evaluation aim for this research question was to gather comparable results from all the different search engines both TCSearch2 based and external search engines.

6.2.1 Qualitative evaluation

The qualitative evaluation experiment in this project was a laboratory study. Laboratory studies are conducted in a fixed space and time. Users have assignments and are observed during the experiment. Comments and opinions outside the parameter of the assignments are recorded for analyses. Guidelines on established design guidelines for qualitative questionnaires are proposed by Jeffrey Rubin (Rubin, 1994):

1. Use the research questions as basis for the goals of the questionnaires questions.
2. The questionnaires should be developed for distribution before, after or during a test session.
3. Questions in a qualitative study should be directed at collecting data that is not easy to acquire during quantitative studies. Data that cannot be observed in a quantitative experiment are feelings, opinions and reasoning for their answers.
4. Formulation of the questions should be designed for simplicity and brevity. Complicated instructions should be minimized.

Developing an evaluation environment and questions that would not influence the users was a challenging task. Before conducting the qualitative evaluation a pilot study was performed. Reason for conducting the test evaluation was to estimate the time needed to conduct an evaluation. After the test evaluation, some minor changes were done to prevent any bias during the evaluation.

To be able to evaluate the result quality of a query, some knowledge is needed to make an educated reply. To prevent the test subjects giving an uneducated reply in the evaluation, all the results were links that the test subjects could follow during the evaluation. Furthermore, if some terms in one of the augmented queries were unfamiliar to the test subjects, they could use the method of their choice to acquire the information needed.

In total, 7 people performed the evaluation, and each test subject used an average of 250 minutes. The most representable group of test subjects, with the resources available, was assembled. Giving particular weight to diverse educational background and even mix of students and working test subjects was in focus in order to make the test subject group representable. None of the test subjects had the same education, and half of them were students and half of them are working. One of the test subjects was an expert, and has held Information Retrieval courses at University level.

The following 7 steps were included. The assignments were given during the evaluation. The test subjects were given the assignments in step 2 only after finishing step 1 and so on.

Step 1)
For each of the terms on the list, write down up to 10 topics that are considered to be the information need from the term.

Hitler
Ireland
Scandinavia
Facebook
Wine
Soccer
Citrus

Step 2)
Use the terms from step 1 and perform search, steps 2-5 use the same search results.
Which of the following result lists is the closest to your expectations, rank from closest to furthest away from your expectations.

Step 3)
There are two text boxes, F and G both representing two queries.
Which of the queries do you prefer?

Step 4)
Rank the result lists from best to worst based on quality.

Step 5)
Which of the result lists would you categorize as good results?

Step 6)
Search two times with inputs of your choice that you have knowledge of.
Rank the result lists from best to worst based on quality.

Step 7)
Have you experienced one or several scenarios with a widely used search engine where the results from the search could have be improved by adding terms that share a semantic connection to the search query, to prevent getting search results that only matches the query without matching the context of the query ?

In the evaluation environment the 4 different TCSearch2 search engines and the Wikipedia search engine was presented to the test subjects. Ranking of the 4 different TCSearch2 search engines allow for isolation of the impact of the different features implemented. One normal problem with evaluation with test subjects is the impact of user interface design issues when testing different search engines. In this evaluation environment this was prevented by using an identical user interface for all the search engines. In addition giving the different search engines random code names and positions could prevent earlier results quality from affecting the evaluation of later evaluation of result quality (Manning et al, 2008, 139).

Not being able to add the result list from Google's search engine to the evaluation environment was the reason why Google was not included in the qualitative evaluation.

6.2.2 Quantitative evaluation

In order to improve the stringency of the evaluation two independent types of evaluation were conducted. Thus, a quantitative evaluation was performed. In addition, the last research question was not fully evaluated by the qualitative evaluation process with the lack of studies that included Google's search engine. A search simulation was performed to gather evaluation data on all the research questions given in Chapter 1. The list of the queries and the information need associated with each query is given in Appendix B.

To collect search results from Google, the site search flag was used. So a Google search for USA for the English Wikipedia on Google would result in the following query:

> "site:en.wikipedia.org/wiki/ USA"

In addition to the Wikipedia articles, several other document types can appear in a result list from Google. All documents that were not Wikipedia articles that were hosted on "en.wikipedia.org/wiki/" were ignored.

The search simulation was performed manually, and this experiment collected the precision of the different search engines. Pooling was used on the result from the search engines, only the top 20 results from each search engine were included in this experiment. The evaluation interface of the TCSearch2 was used to remove any potential bias of the evaluator when comparing the four TCSearch2 search engines and Wikipedia. Google uses personalization of the result list based on previous searches made; to avoid this personalization all cookies were disabled in the browser to prevent that Google would personalize the search results. Results of this evaluation will be presented in the Quantitative evaluation result section of this chapter.

6.3 Final evaluation result

6.3.1 Qualitative evaluation results

In this section the results from the qualitative evaluation will be presented. This result is based on the user response in Appendix A. Some abbreviations are used in the figures and in this section as seen below.

TCSearch2 term to term / TCSearch2 TT
TCSearch2 query augmented /TCSearch2 QA
TCSearch2 query augmented with usage statistics / TCSearch2 QAUS
TCSearch2 query augmented with usage statistics and PageRank /TCSearch2 QAUSPR

Figure 25 shows the percentage of the results for each search engine that were classified as good results by the test subjects.

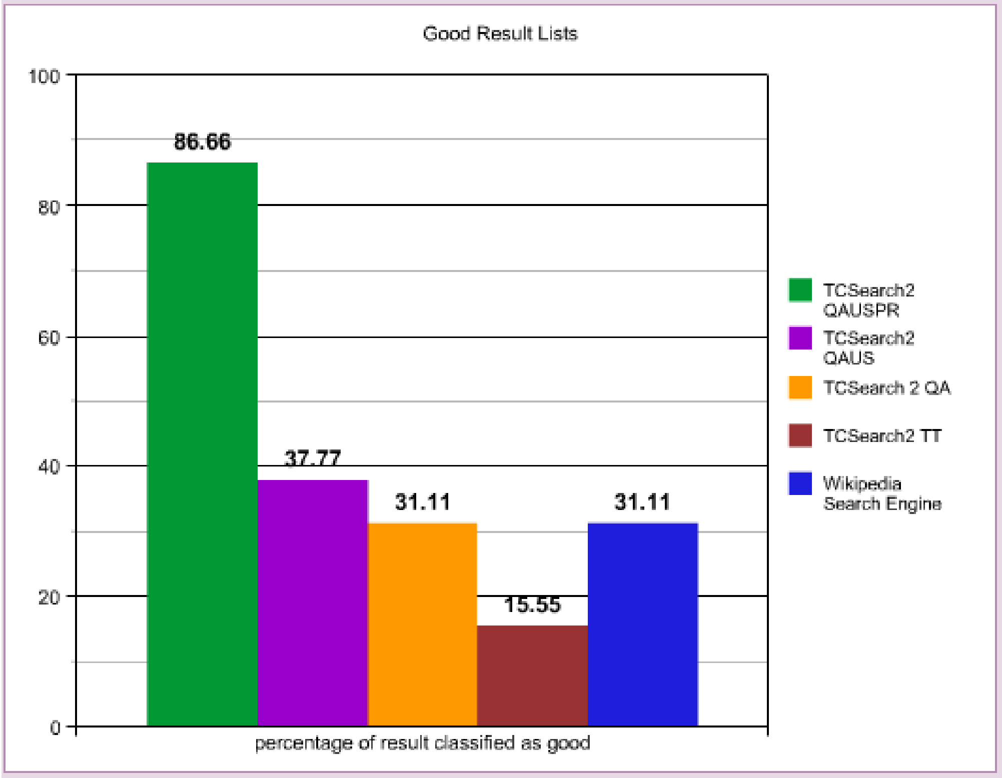


Figure 25 Result lists classified as good by the test subjects

As Figure 25 shows the TCSearch2 QAUSPR search engine had a total of 86,66% of the results list classified as good result lists. Compared to the other search engine this seems to be a high percentage of good result, with over double the amount of results classified as good as the second highest rated search engine. In second place, the TCSearch2 QAUS with 37,77% of the result classified as good. In third place both Wikipedia's search engine and the TCSearch2 QA search engine have the same score of 31,11%. In last place the TSearch2 TT search engine with just 15,55% of the results were classified as good.

With the same search engine core the best TCSearch2 search engine had over five times the amount of results as the lowest performing TCSearch2 search engine. Furthermore, the best performing TCSearch2 search engine almost had three times the amount of result lists classified as good compared to Wikipedia's search engine.

The impact of using the Wikipedia article usage statistics data in the query augmentation process used in the TCSearch2 QAUS search engine, compared to the in the query augmentation process without the use of Wikipedia article usage statistics data as in the TCSearch2 QU search engine was an increase of 6% higher classification of results lists as good.

In Figure 26 the different search engines dispersal of the rank positions based on user expectation of the search engine result are shown from best to worst where 1 is best and 5 is worst. In addition, 0 is when the results did not meet the expectations of the users and could not be included in the ranking. In total 47 queries were used in this evaluation.

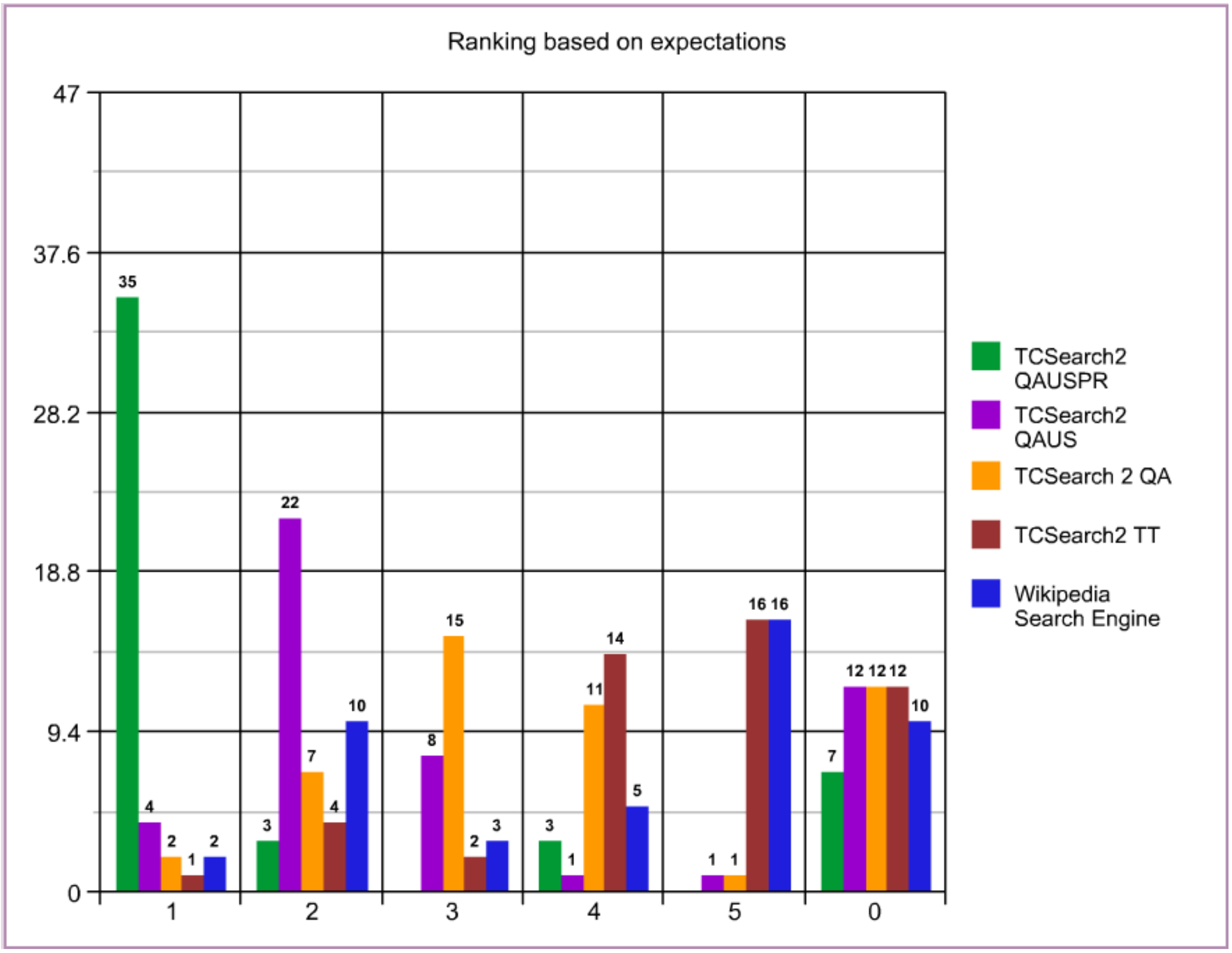


Figure 26 Dispersal of the rank positions based on user expectation

The TCSearch2 QAUSPR is the search engine that best meet the expectations of the users, of the 47 queries in the evaluation it was ranked closest to the users' expectation in 35 cases. The TCSearch2 QAUS was the search engine that came second closest to the users' expectation with over 26 of the 47 cases being on first or second place. Wikipedia's search engine was the one with the second lowest score on user expectation only performing better than the TCSearch2 TT search engine.

In Figure 27 the different search engines are distributed based on the quality of result lists.

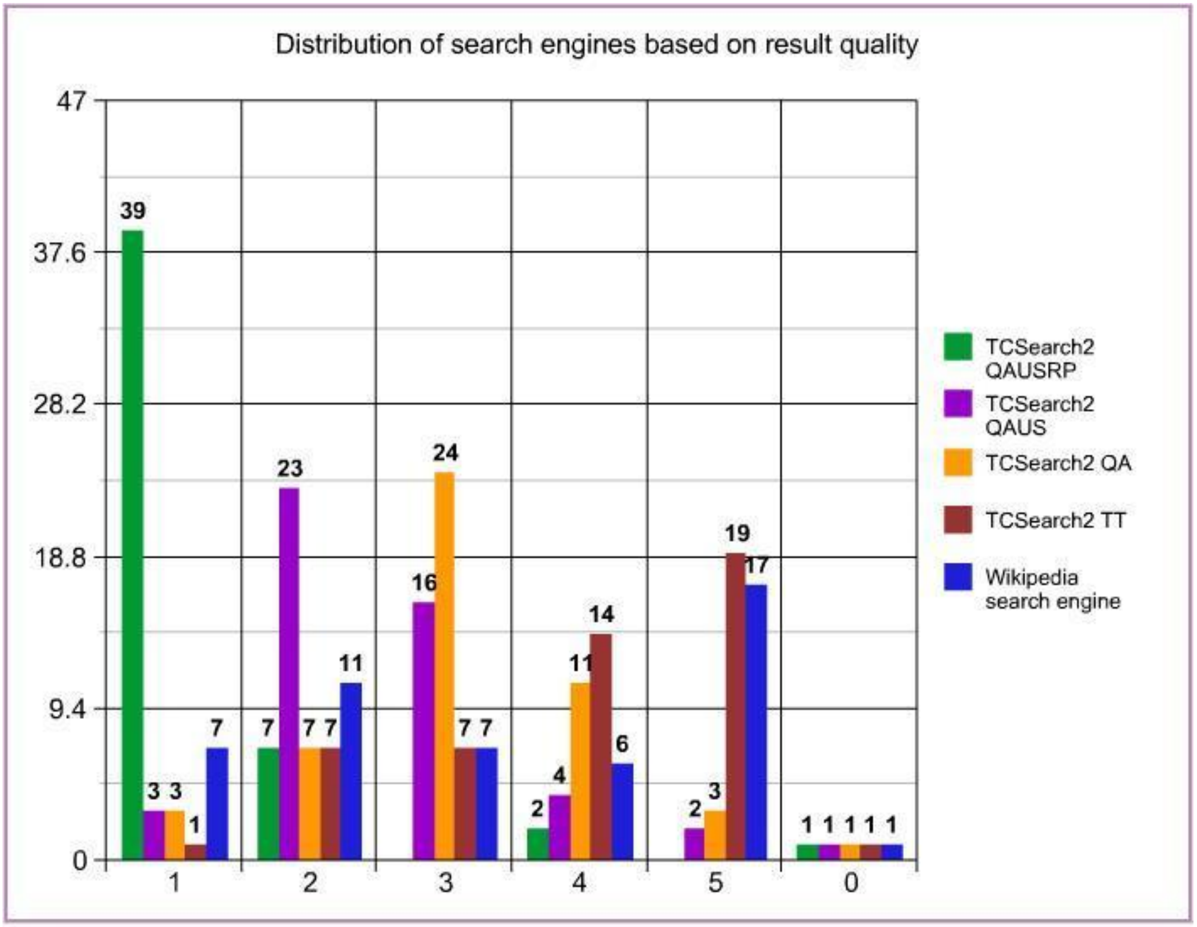


Figure 27 Distributed based on the quality of result lists

Regarding result quality the TCSearch2 QAUSPR outperformed the other search engines with 82% of the cases being ranked as the search engine with the best result quality. While Wikipedia had the second highest amount of first place positions the TCSearch QAUS did in general have a higher quality of results than Wikipedia's search engine.

Compared to the expectation evaluation, Wikipedia's search engine performed significantly better, with a result quality comparable to the TCSearch QA search engine. The TCSearch2 TT is the search engine with the lowest result quality in average.

The last query evaluation was with terms that the test subjects could freely choose. A total of 14 unique queries were searched once. Due to the fact that this evaluation was with such a limited amount of queries the evaluation data was separated from the other query evaluation data, and is presented independently in Figure 28. The very limited size of this study should give this data significantly less importance compared to the larger studies performed.

In Figure 28 the distribution of the queries selected freely from the test subjects is presented based on the result list quality.

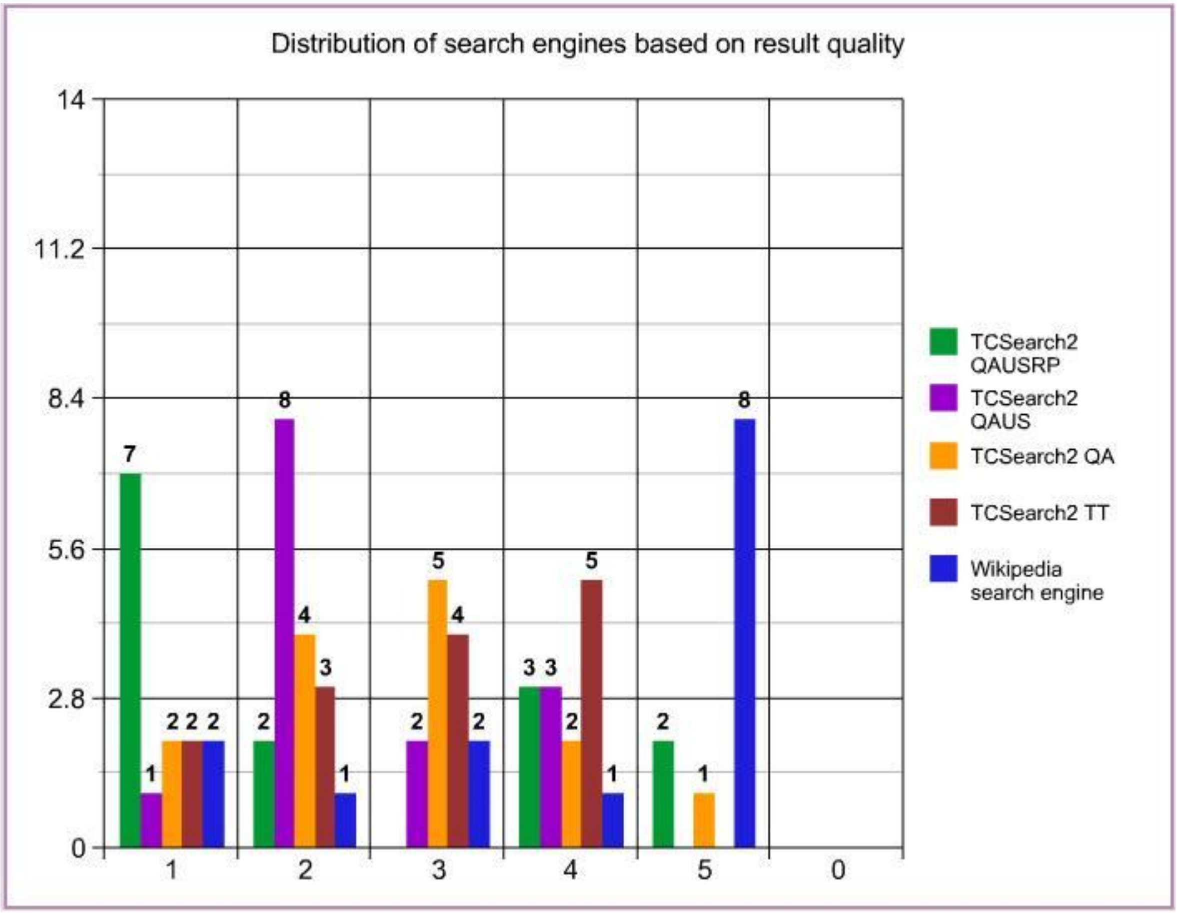


Figure 28 Distributed based on the quality of result lists

The scores of Wikipedia search engine were particularly lower when the test when the subject feely could choose the query, Compared to the score of the other parts of the qualitative evaluation. The weak result of Wikipedia in this part of the qualitative evaluation can be due to limited size of queries preformed in this part of the qualitative evaluation.

Average percentage of the test subject's preference of the augmented queries is presented in Figure 30.

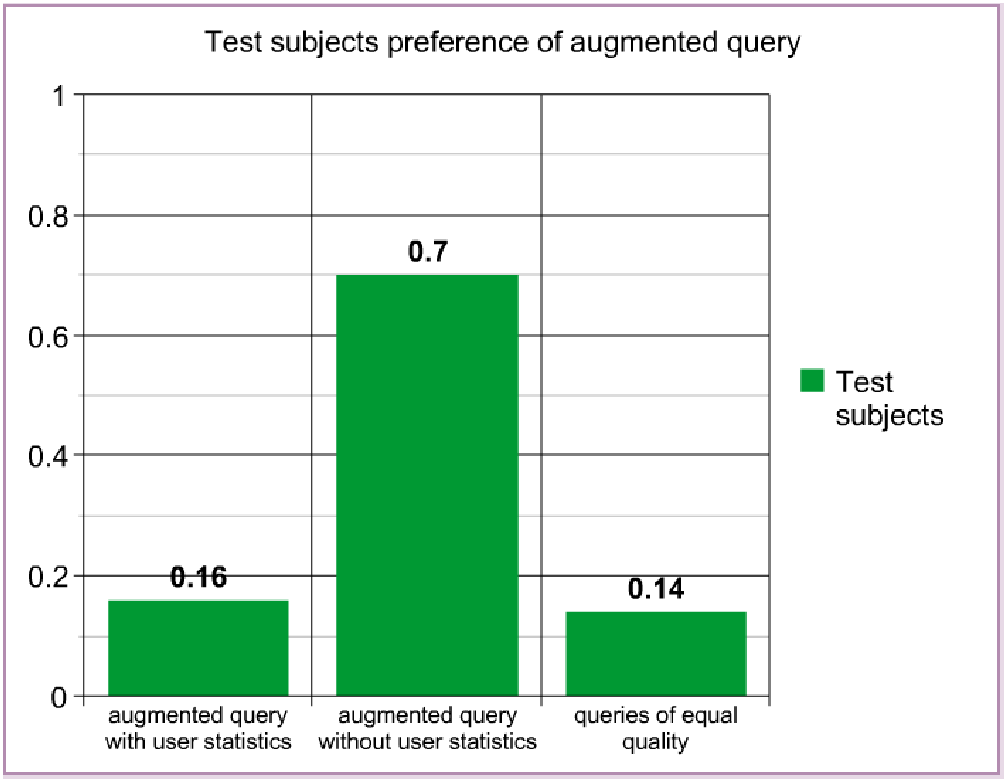


Figure 29 Test subject's preference of the augmented queries

While the search engine result lists based on the augmented query with user statistics was preferred in all the evaluations performed, the augmented query itself without user statistics was preferred in 70% of the cases, and in only 14% of the cases the queries was perceived to be of equal quality.

The last step of the qualitative step was to answer if the user believed that the idea behind the query augmentation process query extensions could improve their search results quality in general. All test subject answered that they thought this approach would be in general positive for the search result quality.

6.3.2 Quantitative evaluation results

Result of the quantitative experiment is presented in Figure 30.

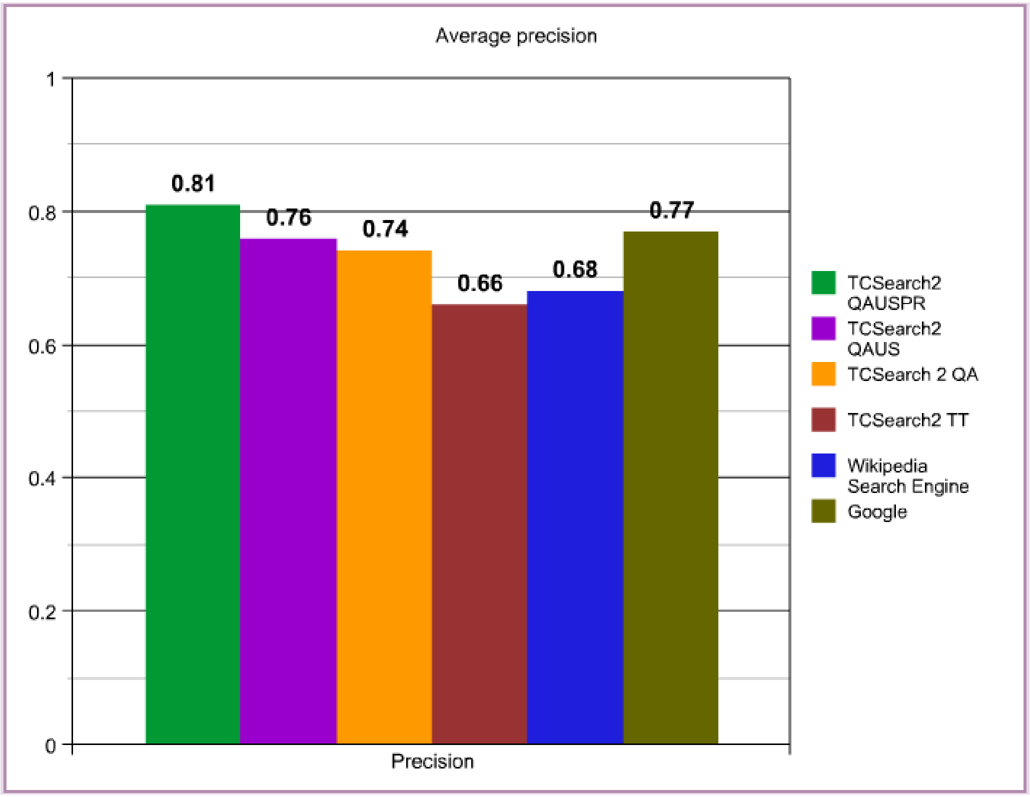


Figure 30 Average precision based on search simulation results

While the Wikipedia search engine is a domain specific search engine for Wikipedia articles, Google's web search engine had a significant higher level of precision. Only the TCSearch2 QAUSPR has a higher precision score than Google in this experiment.

A similar trend was found in the quantitative evaluation data as the qualitative evaluation regarding the ranking of the four TCSearch2 search engines and the Wikipedia search engine. Precision alone is not a perfect indicator of the quality of the search results, but can be an indicator of one aspect of result quality.

6.4 System evaluation

If the aim of this project was to create a search engine several system evaluations should have been performed, such as: How fast does it index? How fast does it search? How expressive is the search language? Size of document collection? (Manning et al, 2008, 155)

The aim of this project was only to create a module that uses the already existing features of the search engine it will be integrated with. Given that the aim of this project is not to create the a new search engine the question from Manning et al regarding system evaluation were not considered to be relevant.

Chapter 7 - Conclusion and future development

7.1 Conclusion

Several different search engines were created for the purpose of answering the research questions raised by this thesis. These engines were constructed on the same core, the TCSearch2. Using the same search engine core and only adding one feature at a time made it possible to measure the differences of search results for each feature added.

Previous studies, such as “A knowledge-based search engine powered by Wikipedia” by Witten and Nichols, have found that the query augmentation process of query expansion using a knowledge base automatically processed from Wikipedia often has a positive impact on the results. Those findings were reconfirmed by evaluation data in this project: Wikipedia can serve as a general knowledge base, which can then be used for query augmentation.

One special case was found during the evaluation. When searching for the term “xnxx”, the term “xnxx” is marked as an adult content by search engine statistics site Searches Organic (Searches Organic SEO. 2012). Due to the policies on adult content on Wikipedia there was a hole in Wikipedia’s ability to be used as a general knowledge base. Potentially there could be several additional holes in Wikipedia that could impact the result quality when using Wikipedia data to create a general knowledge base.

The positive impact of PageRank in the result in the studies was significant. One of the possible reasons for this is that links between Wikipedia articles have a semantic correlation that is stronger than the average links between web pages. Thus, the results are not believed to be reproducible with a different corpus exhibiting weaker semantic interconnections.

The new method of using language dispersal to adjust link connection values based on Wikipedia article usage statistic data were used in this project, such method has not been used before in earlier research on using Wikipedia as a general knowledge base for query augmentation. While there is limited amount of evaluation data, there is a clear positive trend on the impact of including usage statistics in the query augmentation process. Several previous studies on the impact of usage statistics in Information Retrieval field have had similar findings (Hu et al. 2007; Agichtein et al. 2006). Using the new method in combination with other usage statistics methods such as giving popular topics an added bonus, future studied is needed to conclude on the impact of using language dispersal to adjust link connection value between Wikipedia articles.

The last research question was regarding the result quality difference of the TCSearch2 search engines compared with existing search engines. Based on the evaluation data several of the TCSearch2 search engines have a result quality that is in average higher than Wikipedia’s search engine. Given the limited resources used on developing the TCSearch2 this was an unexpected finding. The method used to extracting the query terms in this study may have extracted queries where the Wikipedia search engine underperformed with regards to result quality, compared to a study using a different method of extracting evaluation queries. The very limited amount of searches performed freely by the test subjects suggest that this is not the case. The studies performed were not representable enough to make a conclusion, but a strong indication in the evaluation data found that Wikipedia's search engine is outperformed by several TCSearch2 search engines.

With an even smaller amount of evaluation data for comparing the result quality between the TCSearch2's different search engines with Google it is impossible to draw any definitive conclusion. The findings in the experiment of precision show that for the limited evaluation data that some of the TCSearch2 search engines did perform comparable to Google's search engine for the subsection of the Internet Wikipedia. This finding, though inconclusive, is surprising with regards to the resources at Google's disposal and the tight integration of Google with Wikipedia (Roberts, 2012). An explanation for the results found in this study could be that Google's aim of diversification of the results lowered the quality of the results compared to the information needs stated in this study.

Another surprising finding was the selection of what augmented query the user preferred. The evaluation data shows that users were not able to evaluate what augmented query would achieve the best result. All test subjects preferred the augmented query without usage statistics, but all test subjects preferred the result list of the augmented query with usage statistics. This finding indicates that users are not able to recognize what query will give the best result. This can suggest that manual evaluation of augmented queries is not advisable.

The last question asked in the qualitative questionnaire, was if users believed that their search results could be significantly improved if search engines used some form of query augmentation by query expansion. All of the participants believed that search results would see an improvement from query augmentation. This result indicates that the current dominating search engines do not provide the users with an adequate result quality, at least for some queries.

To sum it up, this thesis has been successful in acquiring evaluation data on all the research questions posed in the introduction. This thesis have both confirmed previous findings in other studies and produced strong indications of new findings. While the amount of evaluation data was limited due to the time constraints of this project, there are some clear indications that were found in the evaluation data.

7.2 Future Work

One of the weaknesses of this thesis is that the article popularity and article language dispersal were not separately evaluated due to time constraints. Several parts of Wikipedia have not been used and could possibly improve the quality of Wikipedia as a general knowledge base for query augmentation. One example of this is analyzing the behavior of Wikipedia contributors in addition to the Wikipedia users. A huge structural part of Wikipedia, the categories, was not used in this project, and could improve the connection value between different articles. Finally, Wikipedia also have a quality ranking of the articles that could have been used to boost the value of well written articles.

Storage of this system is based on MySQL since I was familiar with this storage structure. Performance in MySQL for this particular task was not optimal and should in future work be replaced by a storage engine that does not use relational tables. To make matters worse the table engine used in MySQL for this application has been deprecated during the development of this project due to lack of good multithreaded support.

To improve the augmented query process it should include multi word term. Free text search in addition to the bag of word model could improve the correctness of the stimulation of the articles in the query augmentation process. This project used a very blunt normalization process both in regards to character normalization and stemming. Terms like “C++” were normalized to “c” in this project. Using improved engine indexer, dictionary and extensive special term lists could have create a more correct simulation of documents in the knowledge base. Improved document stimulation would have improved the query augmentation process, and would in turn result in a better result quality.

Some queries greatly benefit from query augmentation, while other queries do not have any benefits from query augmentation. Creating a method that automatically finds the optimal amount of terms to add in a query, could improve both result quality and lower the computational needs in this project.

References

1 S.F Adafre & M. de Rijke. 2005. “Discovering Missing Links in Wikipedia”. [ONLINE] Available at: http://tinyurl.com/bw6fn84 . [Accessed 14 December 2012].

2 E Agichtein, E Brill & S Dumais. 2006. “Improving Web Search Ranking by Incorporating User Behavior Information “. [ONLINE] Available at: http://research.microsoft.com/en-us/um/people/sdumais/sigir2006-fp345-ranking-agichtein.pdf. [Accessed 14 December 2012].

3 AOL. 2012. “Top search topics, trends & stories around the web - Aol Hot Searches”. [ONLINE] Available at: http://hotsearch.aol.co.uk/us-review/ . [Accessed 14 December 2012].

4 A.R Brown .2011. “Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage”. [ONLINE] Available at: http://adambrown.info/docs/research/brown-2011-wikipedia-as-a-data-source.pdf . [Accessed 14 December 2012].

5 J Brownlee. 2011. “Clever Algorithms: Nature-Inspired Programming Recipes. Self-Organizing Map”. [ONLINE] Available at: http://www.cleveralgorithms.com/nature-inspired/neural/som.html . [Accessed 14 December 2012].

6 S Chernov, T Iofciu, W Nejdl & X Zhou. 2006. “Extracting Semantic Relationships between Wikipedia Categories” [ONLINE] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.5507&rep=rep1&type=pdf. [Accessed 14 December 2012].

7 Chitika Insights. 2010. “The Value of Google Result Positioning”. [ONLINE] Available at: http://insights.chitika.com/2010/the-value-of-google-result-positioning/ . [Accessed 14 December 2012].

8 N Cohen. 2007. “Wikipedia on an academic hit list”. [ONLINE] Available at: http://www.taipeitimes.com/News/editorials/archives/2007/02/27/2003350261 . [Accessed 14 December 2012].

9 Experian. 2012. “Experian Marketing Services Reports Google Share of Searches at 65 Percent in May 2012”. [ONLINE] Available at: http://press.experian.com/United-States/Press-Release/experian-marketing-services-reports-google-share-of-searches-at-65-percent-in-may-2012.aspx. [Accessed 14 December 2012].

10 T Finin, Y Peng, R.S Cost, J Sachs, A Joshi, P Reddivari, R Pan, V Doshi & L Ding. 2004. “Swoogle: a search and metadata engine for the semantic web”. [ONLINE]

Available at: http://dl.acm.org/citation.cfm?id=1031289 . [Accessed 14 December 2012].

11 S Gauch & J.B Smith. 1993. “An Expert System for Automatic Query Reformulation”. [ONLINE] Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.5835 . [Accessed 14 December 2012].

12 A Gerber, A van der Merwe & R Alberts. 2007. “Practical Implications of Rapid Development Methodologies”. [ONLINE] Available at: http://csited.org/2007/73GerbCSITEd.pdf . [Accessed 14 December 2012].

13 Google Forum. 2012. “Every millisecond counts” [ONLINE] Available at: http://www.googleforum.co.uk/millisecond-counts . [Accessed 14 December 2012].

14 Google Search Result. 2012. Video search for “SS” [ONLINE] Available at: http://www.google.no/search?q=ss&oq=ss&sugexp=chrome,mod=6&sourceid=chrome&ie=UTF-8#q=ss&hl=en&prmd=imvns&source=lnms&tbm=vid&sa=X&ei=3fOfULTULI_04QTIyYCoCQ&ved=0CAwQ_AUoAw&bav=on.2,or.r_gc.r_pw.r_qf.&fp=6ba02b0ea76834a3&bpcl=38093640&biw=1167&bih=1035 . [Accessed 1 December 2012].

15 M Gruninger & J Lee. 2002. “Ontology applications and design”. [ONLINE] Available at: http://alarcos.inf-cr.uclm.es/doc/masi/doc/lec/parte3/gruninger-cacm02.pdf . [Accessed 14 December 2012].

16 X Han & J Zhao. 2009. “Named entity disambiguation by leveraging Wikipedia semantic knowledge” [ONLINE] Available at: http://dl.acm.org/citation.cfm?id=1645983 [Accessed 14 December 2012].

17 M Hu, E.P Lim, A Sun, H.W. Lauw & B.Q Vuong. 2007. “On Improving Wikipedia Search using Article Quality” [ONLINE] Available at: http://www.cais.ntu.edu.sg/~axsun/paper/sun_widm07.pdf [Accessed 14 December 2012].

18 J Huang, D.J Abadi & K Ren. 2011. “Scalable SPARQL Querying of Large RDF Graphs”. [ONLINE] Available at: http://cs-www.cs.yale.edu/homes/dna/papers/sw-graph-scale.pdf . [Accessed 14 December 2012].

19 Internet World Stats. 2012. “World Internet Users Statistics Usage”. [ONLINE] Available at: http://www.internetworldstats.com/stats.htm . [Accessed 14 December 2012].

20 S Jaschik. 2011. “A Stand Against Wikipedia” Available at: http://www.insidehighered.com/news/2007/01/26/wiki . [Accessed 14 December 2012].

21 A Kittur, E.H Chi & B Suh. 2009. “What’s in Wikipedia? Mapping Topics and Conflict Using Socially Annotated Category Structure”. [ONLINE] Available at: http://www-users.cs.umn.edu/~echi/papers/2009-CHI2009/p1509.pdf . [Accessed 14 December 2012].

22 R Kraft, C.C Chang, F Maghoul & R Kuma. 2006. “Searching with Context”. [ONLINE] Available at: http://www2006.org/programme/item.php?id=3015 . [Accessed 14 December 2012].

23 S Lovic, M Lu & D Zhang. 2006. “Enhancing Search Engine Performance Using Expert Systems”. [ONLINE] Available at: http://tinyurl.com/cm8cmnx . [Accessed 14 December 2012].

24 W.Y Ma, 2008. “Web Information Retrieval – History and Future Trends”. [Online] Available at: http://research.microsoft.com/en-us/collaboration/global/asia-pacific/talent/webirhistoryfuturetrends.pdf [Accessed 14 December 2012]

25 D.C Manning, P Raghavan & H Schütze . 2008. “Introduction to information retrieval”. 1 Edition. Cambridge University Press

26 P McCorduck, 2004, “Machines Who Think” 2 edition Peters

27 R McHenry. 2004. “The Faith-Based Encyclopedia” [ONLINE] Available at: http://www.ideasinactiontv.com/tcs_daily/2004/11/the-faith-based-encyclopedia.html . [Accessed 14 December 2012].

28 R Meyer. 2012. “3 Charts That Show How Wikipedia Is Running Out of Admins”. [ONLINE] Available at: http://www.theatlantic.com/technology/archive/2012/07/3-charts-that-show-how-wikipedia-is-running-out-of-admins/259829/ . [Accessed 14 December 2012].

29 MultiID. 2012. “SOM illustration”. [ONLINE] Available at: http://www.multid.se/genex/SOM_illustration.png . [Accessed 14 December 2012].

30 MySQL A. 2012 “MySQL Customer: Google” [ONLINE] Available at: http://www.mysql.com/customers/view/?id=555 . [Accessed 14 December 2012].

31 MySQL B. 2012 “MySQL Customer: Flicker” [ONLINE] Available at: http://www.mysql.com/customers/view/?id=720 . [Accessed 14 December 2012].

32 MySQL C. 2011. “Dispelling the Myths”. [ONLINE] Available at: http://web.archive.org/web/20110606013619/http://dev.mysql.com/tech-resources/articles/dispelling-the-myths.html . [Accessed 14 December 2012].

33 NeuroSolutions. 2012. “What is a Neural Network?”. [ONLINE] Available at: http://www.nd.com/welcome/whatisnn.htm . [Accessed 14 December 2012].

34 J Newby. 2012. “Mia Love Wikipedia page vandalized with misogynistic, racial slurs; media silent”. [ONLINE] Available at: http://www.examiner.com/article/mia-love-wikipedia-page-vandalized-with-misogynistic-racial-slurs-media-silent . [Accessed 14 December 2012].

35 A Orlowski. 2006. “Avoid Wikipedia, warns Wikipedia chief”. [ONLINE] Available at: http://www.theregister.co.uk/2006/06/15/wikipedia_can_damage_your_grades/ . [Accessed 14 December 2012].

36 S.L Pfleeger & J.M Atlee. 2005. “Software Engineering : Theory and Practice”. 3 Edition. Pearson

37 G Pölzlbauer. 2004. “Survey and Comparison of Quality Measures for Self-Organizing Maps”. [ONLINE] Available at: http://www.ifs.tuwien.ac.at/~poelzlbauer/publications/Poe04WDA.pdf . [Accessed 14 December 2012].

38 M.F Porter. 1997 “An algorithm for suffix stripping Readings in Information Retrieval (The Morgan Kaufmann Series in Multimedia Information and Systems)”. 1st Edition. Morgan Kaufmann Publishers, Inc.

39 N.J Reavley , A.J Mackinnon, A.J Morgan, M Alvarez-Jimenez, S.E Hetrick, E Killackey, B Nelson, R Purcell, M.B Yap & A.F Jorm. 2012. “Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources.” [ONLINE] Available at:

40 J.J Roberts. 2012. “Google shakes up search with new Wikipedia-like feature — Tech News and Analysis”. [ONLINE] Available at: http://gigaom.com/2012/05/16/google-shakes-up-search-with-new-wikipedia-like-feature/ . [Accessed 14 December 2012].

41 R Rojas. 1996. “*Neural Networks: A Systematic Introduction”*. 1 Edition. Springer.

42 J Rubin. 1994. “Handbook of usability testing : how to plan, design, and conduct effective tests”. Wiley.

43 S Russell & P Norvig , 2003, “Artificial Intelligence: A Modern Approach”, 2 edition Prentice Hall

44 Searches Organic SEO. 2012. “Top Google Searches in 2012”. [ONLINE] Available at: http://seattleorganicseo.com/sosblog/top-google-searches-in-2012-the-most-popular-keywords-study-version-3/#top25keywords2012 . [Accessed 14 December 2012].

45 Statistic Brain. 2012. “Google Annual Search Statistics” [ONLINE] Available at: http://www.statisticbrain.com/google-searches/ . [Accessed 14 December 2012].

46 K Strauss. 2012. “Artificial Intelligence is the Next Step in Search (and everything else)”. [ONLINE] Available at: http://www.forbes.com/sites/karstenstrauss/2012/06/20/artificial-intelligence-is-the-next-step-in-search-and-everything-else/ . [Accessed 14 December 2012].

47 M Strube & S. P Ponzetto. 2006. ”WikiRelate! Computing Semantic Relatedness Using Wikipedia” . [ONLINE] Available at: http://www.aaai.org/Papers/AAAI/2006/AAAI06-223.pdf . [Accessed 14 December 2012].

48 K Suchecki, A.A.A Sala, C Gao & A Scharnhors .2012. “Evolution of Wikipedia’s category structure” . [ONLINE] Available at: http://arxiv.org/pdf/1203.0788.pdf . [Accessed 14 December 2012].

49 A Efrati. 2011. “Google Notches One Billion Unique Visitors Per Month”. [ONLINE] Available at: http://blogs.wsj.com/digits/2011/06/21/google-notches-one-billion-unique-visitors-per-month/ . [Accessed 14 December 2012].

50 W3School. 2012. “OS Statistics” [ONLINE] Available at: http://www.w3schools.com/browsers/browsers_os.asp . [Accessed 14 December 2012].

51 Wikimedia. 2012. “Nine out of ten Wikipedians continue to be men: Editor Survey”. [ONLINE] Available at: http://blog.wikimedia.org/2012/04/27/nine-out-of-ten-wikipedians-continue-to-be-men/ . [Accessed 14 December 2012].

52 Wikimedia Statistics. 2012. “Edits per article” [ONLINE] Available at: http://stats.wikimedia.org/EN/TablesArticlesEditsPerArticle.htm . [Accessed 14 December 2012].

53 Wikipedia A. 2012. “Wikipedia:About” [ONLINE] Available at: http://en.wikipedia.org/wiki/Wikipedia:About . [Accessed 14 December 2012].

54 Wikipedia B. 2012. “Template:INRConvert “ [ONLINE] Available at: http://en.wikipedia.org/wiki/Template:INRConvert . [Accessed 14 December 2012].

55 Wikipedia C. 2012 “Wikipedia:Portal”. [ONLINE] Available at: http://en.wikipedia.org/wiki/Wikipedia:Portal . [Accessed 14 December 2012].

56 Wikipedia D. 2012. “Wikipedia:Template”. [ONLINE] Available at: http://en.wikipedia.org/wiki/Help:Template . [Accessed 14 December 2012].

57 Wikipedia E. 2012. “Wikipedia:Protection policy”. [ONLINE] Available at: http://en.wikipedia.org/wiki/Wikipedia:Protection_policy . [Accessed 14 December 2012].

58 Wikipedia F. 2012. “Wikipedia:Notability”. [ONLINE] Available at: http://en.wikipedia.org/wiki/Wikipedia:Notability . [Accessed 14 December 2012].

59 Wikipedia G. 2012. “Wikipedia:Modelling Wikipedia's growth”. [ONLINE] Available at: http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth . [Accessed 14 December 2012].

60 Wikipedia H. 2012. “Wikipedia:Database download”. [ONLINE] Available at: http://en.wikipedia.org/wiki/Wikipedia:Database_download . [Accessed 14 December 2012].

61 D.M.I Witten & D.M Nichols. 2007. “A knowledge-based search engine powered by Wikipedia” [ONLINE] Available at: http://researchcommons.waikato.ac.nz/handle/10289/5379 . [Accessed 14 December 2012].

Appendix A Qualitative user response

All of the qualitative data from the test subjects were of considerable length only a subset of the total amount of data gathered from the study is included in the thesis. Only data from 2 of the total 7 test subjects is included. The remaining amount of user response data is located online at https://docs.google.com/document/d/1lMGfGPV5477nK6ztifqcJlihfKSX9XYFAITD9EmzCeM/edit

The same abbreviation of the different search engines is as in Chapter 6.

User 01

Step 1-5

Query term: Hitler

Expected information need:
Adolf Hitler
Nazi Germany
World war 2
Events in 1945
Events in 1939

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
TCSearch2 QAUS	TCSearch2 QAUS	TCSearch2 QAUS
TCSearch2 QA	TCSearch2 QA	TCSearch2 QA
TCSearch2 TT	Wikipedia	
Wikipedia	TCSearch2 TT	

Query augmentation preferred:
Query augmentation without user statistics

Comments or opinions given during this query:
None

Query term: Ireland

Expected information need:
Northern-Ireland
Irish
Commonwealth
Great Britain
Conflicts regarding Ireland

Search result ranked on users expectation	Ranked result	Good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
TCSearch2 TT	TCSearch2 TT	
Wikipedia	Wikipedia	
TCSearch2 QAUS	TCSearch2 QAUS	
TCSearch2 QA	TCSearch2 QA	

Query augmentation preferred
Query augmentation without user statistics

Comments or opinions given during this query:
“When I searched for Ireland I expect it to be at the top, everything else is just stupid”

Query term: Scandinavia

Expected information need:
Norway
Sweden
Denmark
Finland
Iceland

Search result ranked on users expectation	Ranked result	Good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
TCSearch2 QAUS	Wikipedia	
Wikipedia	TCSearch2 QAUS	
TCSearch2 QA	TCSearch2 QA	
TCSearch2 TT	TCSearch2 TT	

Query augmentation preferred
Query augmentation without user statistics

Comments or opinions given during this query:
None

Query term: Facebook

Expected information need:
Privacy on social media
Social media
Mark Zuckerberg
The Social Network

Search result ranked on users expectation	Ranked result	Good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
TCSearch2 QAUS	TCSearch2 QAUS	TCSearch2 QAUS
Wikipedia	Wikipedia	
TCSearch2 QA	TCSearch2 QA	
TCSearch2 TT	TCSearch2 TT	

Query augmentation preferred
Query augmentation without user statistics

Comments or opinions given during this query:
None

Query term: Wine

Expected information need:
Linux software for emulation Windows
Grapes
France
Red wine
White wine

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
Wikipedia	TCSearch2 QAUSPR	TCSearch2 QAUSPR
TCSearch2 QAUSPR	Wikipedia	Wikipedia
-	TCSearch2 QAUS	
-	TCSearch2 QA	
-	TCSearch2 TT	

Query augmentation preferred:
Query augmentation without user statistics

Comments or opinions given during this query:
None

Query term: Soccer

Expected information need:
Football
World Cup
Teams such as Manchester United

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
-	-	
-	-	
-	-	
-	-	
-	-	

Query augmentation preferred:
The query augmentation was of equal quality

Comments or opinions given during this query:
“None of the result list were acceptable result lists”

Query term: Citrus

Expected information need:
Lemon
Fruite
Orange
Trees

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
TCSearch 2 QAUSPR	TCSearch 2 QAUSPR	TCSearch 2 QAUSPR
TCSearch 2 QAUS	Wikipedia	
TCSearch 2 QA	TCSearch 2 QAUSPR	
Wikipedia	TCSearch 2 QA	
TCSearch 2 TT	TCSearch 2 TT	

Query augmentation preferred:
Query augmentation without user statistics

Comments or opinions given during this query:
None

Step 6)

Query provided by the test subject: Programming

Search result ranked on the result
Wikipedia
TCSearch2 QAUS
TCSearch2 QA
TCSearch2 TT
TCSearch2 QAUSPR

Query provided by the test subject: Coq

Search result ranked on the result
TCSearch2 TT
TCSearch2 QAUS
TCSearch2 QA
Wikipedia
TCSearch2 QAUSPR

Step 7)

“I have experienced that I have performed a search, and the articles I have got as a result contains what I have searched for but not in context. So in short yes, but getting documents that don't contain the term search for is not good.”

User 02

Step 1-5

Query term: Hitler

Expected information need:
World war 2
Holocaust
Germany

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
-	TCSearch2 QAUS	TCSearch2 QAUS
-	TCSearch2 QA	TCSearch2 QA
-	TCSearch2 TT	
-	Wikipedia	

Query augmentation preferred:
Query augmentation without user statistics

Comments or opinions given during this query:
None

Query term: Ireland

Expected information need:
Kelter
Dublin
Catholicism in Ireland
Iconic Irish beverages such as Guinness

Search result ranked on users expectation	Ranked result	Good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
Wikipedia	TCSearch2 QAUS	
TCSearch2 QAUS	TCSearch2 QA	
TCSearch2 QA	Wikipedia	
TCSearch2 TT	TCSearch2 TT	

Query augmentation preferred
Query augmentation without user statistics

Comments or opinions given during this query:
“Ranking is very important in the result set”

Query term: Scandinavia

Expected information need:
The Scandinavian countries (Norway, Sweden and Denmark)
Scandinavian people

Search result ranked on users expectation	Ranked result	Good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
Wikipedia	Wikipedia	
TCSearch2 QAUS	TCSearch2 QAUS	
TCSearch2 QA	TCSearch2 QA	
TCSearch2 TT	TCSearch2 TT	

Query augmentation preferred
Query augmentation without user statistics

Comments or opinions given during this query:
None

Query term: Facebook

Expected information need:
facebook.com
Social media
Mark Zuckerberg
The Social Network

Search result ranked on users expectation	Ranked result	Good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
TCSearch2 QAUS	TCSearch2 QAUS	TCSearch2 QAUS
TCSearch2 QA	TCSearch2 QA	TCSearch2 QA
TCSearch2 TT	TCSearch2 TT	TCSearch2 TT
Wikipedia	Wikipedia	Wikipedia

Query augmentation preferred
Query augmentation with statistics

Comments or opinions given during this query:
“The terms in the augmented query gave me information to a larger extend than the result set. And it is easy to see what the search engines thinks you are looking for.”

“ It is not so important which of the following results set I would have got, they all gave me the results that I would have liked to see I think. ”

Query term: Wine

Expected information need:
Red wine
White wine
alcohol
Vineyards
Wine district
Wine grapes
Rose wine

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
TCSearch2 QAUSPR	TCSearch2 QAUSPR	TCSearch2 QAUSPR
Wikipedia	Wikipedia	
TCSearch2 QAUS	TCSearch2 QAUS	
TCSearch2 QA	TCSearch2 QA	
TCSearch2 TT	TCSearch2 TT	

Query augmentation preferred:
Query augmentation without user statistics

Comments or opinions given during this query:
“I don't know what viticulture means but since it is on the list(an extended term in the augmented query) I think it is relevant to wine”
“It is unnatural that Wine software appears in the result list, nobody that is searching for just “wine” wants that, they would have written “wine software or something” ”
“But I think that Wine software can for a few people be what they are searching for”
“Compared to my expectations I feel the results in general are too preoccupied with geographical locations. Except list E (TCSearch2 QAUSPR)”
“Most lists seems to be very americanish results”

Query term: Soccer

Expected information need:
Football
Football World Cup
Premier League
Football Rules

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
-	Wikipedia	
-	TCSearch2 QAUSPR + TCSearch2 QAUS + TCSearch2 QA+ TCSearch2 TT	
-		
-		
-		

Query augmentation preferred:
Query augmentation without statistics

Comments or opinions given during this query:
“I expected FIFA, it should be included to be a usable result”

Query term: Citrus

Expected information need:
Citrus fruits
Lemon
Orange
Citrus tastes
Citrus trees

Search result ranked on users expectation	Search result ranked on the result	Results classified as good results
TCSearch 2 QAUSPR	TCSearch 2 QAUSPR	TCSearch 2 QAUSPR
-	Wikipedia	
-	TCSearch 2 QAUSPR	
-	TCSearch 2 QA	
-	TCSearch 2 TT	

Query augmentation preferred:
Query augmentation without user statistics

Comments or opinions given during this query:
None

Step 6)

Query provided by the test subject: Beer

Search result ranked on the result
TCSearch2 QAUSPR
TCSearch2 TT
Wikipedia
TCSearch2 QAUS
TCSearch2 QA

Query provided by the test subject: Stiklestad

Search result ranked on the result
TCSearch2 QAUSPR
TCSearch2 QAUS
TCSearch2 QA
TCSearch2 TT
Wikipedia

Step 7)

"Yes, I think that would sometimes help in improving the result quality. I have sometimes experienced that I have got documents in my result list, that only contained the search term in the advertisement of the page, and not the content of the page. I think this approach can sometimes remove such results."

Appendix B Quantitative search simulation results

This was the list of query input used in the quantitative search process; the information need was given by the sources from where the terms were extracted. The explanation of the information needs is in parentheses after the search input terms. Due to the amount of data only a subset of the data is presented in the thesis while all the data used in the evaluation is publicly accessible online. A total of 2 of the 30 queries will be included. Below is the list of all the terms used in the quantitative experiment:

1 iPhone 4 (Smartphone developed by apple computer)
2 Indochine (Area in southeast Asia)
3 Depression (Mental disorder)
4 AK-47 (assault rifle)
5 Sushi (Food dish)
6 Star Wars (Entertainment franchise)
7 Deepwater Horizon (Offshore platform that exploded in the bay of Mexico with a following natural disaster)
8 Führer (German term for leader, used as title for the leader of nazi Germany closely attached to Adolf Hitler)
9 007 (James Bond)
10 USSR (Soviet Union)
11 Newton (Isaac Newton scientist)
12 Naples (Area in south Italy)
13 Muammar Gaddafi (Former leader in Libya)
14 Steve Jobs (Former CEO of several companies but most attached to Apple Computer)
15 Toyota (Car manufacturer)
16 Tiger Woods (Golfer)
17 Kate Middleton (princess in the UK)
18 Vuvuzela (musical instrument that became known during the 2010 World Cup in South Africa)
19 Wikileaks (NGO that publicise confidential information)
20 Aung San Suu Kyi (Burmese pro human right activist)
21 Pyongyang (North Korean Capital)
22 Aspergers (Autistic syndrome)
23 Richard Nixon (Former president in the USA)
24 Espionage (The act or practice of spying to collect secret information)
25 SS (A paramilitary force in Nazi Germany that grew to be an elite military force)
26 NATO (A defence organisation of several european and north american countries)
27 EU (The European Union)
28 TEPCO (Tokyo Electric Power Company)
29 The Simpsons (Cartoon)
30 The Beatles (British pop band that grew popular in the 1960s)

Query: iPhone 4

TCSearch2 TT

TCSearch2 TT result	TCSearch TT relevant results
IPhone	IPhone
IOS_version_history	IOS_version_history
IPhone_4	IPhone_4
IPhone_4S	IPhone_4S
IPhone_(original)	IPhone_(original)
IPhone_3G	IPhone_3G
IPhone_3GS	IPhone_3GS
Linksys_iPhone	-
300-page_iPhone_bill	300-page_iPhone_bill
History_of_the_iPhone	History_of_the_iPhone
IPhone_Dev_Team	IPhone_Dev_Team
IPhone_art	IPhone_art
IPhone_(disambiguation)	-
List_of_iOS_devices	List_of_iOS_devices
Smartphone	Smartphone
ITunes	ITunes
IOS_jailbreaking	IOS_jailbreaking
Iphone_Sessions	Iphone_Sessions
Greenpois0n	-
Apple_Inc.	Apple_Inc.

TCSearch QA

TCSearch2 QA result	TCSearch QA relevant results
IPhone	IPhone
IOS_version_history	IOS_version_history
IPhone_4	IPhone_4
IPhone_4S	IPhone_4S
IPhone_3G	IPhone_3G
IPhone_3GS	IPhone_3GS
IPhone_(original)	IPhone_(original)
Linksys_iPhone	-
History_of_the_iPhone	History_of_the_iPhone
List_of_iOS_devices	List_of_iOS_devices
ITunes	ITunes
IOS	IOS
IOS_jailbreaking	IOS_jailbreaking
IPad	IPad
Apple_Inc.	Apple_Inc.
300-page_iPhone_bill	300-page_iPhone_bill
Smartphone	Smartphone
IPhone_Dev_Team	IPhone_Dev_Team
IPod	IPod
Greenpois0n	-

TCSearch QAUS

TCSearch2 QAUS result	TCSearch QAUS relevant results
IPhone	IPhone
IOS_version_history	IOS_version_history
IPhone_4	IPhone_4
IPhone_4S	IPhone_4S
IPhone_3G	IPhone_3G
IPhone_3GS	IPhone_3GS
IPhone_(original)	IPhone_(original)
List_of_iOS_devices	List_of_iOS_devices
History_of_the_iPhone	History_of_the_iPhone
Linksys_iPhone	-
ITunes	ITunes
Apple_Inc.	Apple_Inc.
IOS	IOS
IPad	IPad
IOS_jailbreaking	IOS_jailbreaking
IPod	IPod
300-page_iPhone_bill	300-page_iPhone_bill
Smartphone	Smartphone
Phone_Dev_Team	Phone_Dev_Team
Greenpois0n	-

TCSearch QAUSPR

TCSearch2 QAUSPR result	TCSearch QAUSPR relevant results
IPhone	IPhone
Apple_Inc.	Apple_Inc
IPad	IPad
IPhone_4	IPhone_4
Pod_Touch	Pod_Touch
IOS	IOS
IPhone_4S	IPhone_4S
IPhone_3GS	IPhone_3GS
IPhone_3G	IPhone_3G
IPhone_(original)	IPhone_(original)
IOS_version_history	IOS_version_history
ITunes	ITunes
IPod	IPod
IOS_jailbreaking	IOS_jailbreaking
App_Store_(iOS)	App_Store_(iOS)
List_of_iOS_devices	List_of_iOS_devices
Steve_Jobs	Steve_Jobs
ITunes_Store	ITunes_Store
Smartphone	Smartphone
Macintosh	Macintosh

Wikipedia’s search engine

Wikipedia’s search engine result	Wikipedia’s search engine relevant results
IPhone_4	IPhone_4
IPhone	IPhone
IPhone_4S	IPhone_4S
FaceTime	FaceTime
Apple_Inc.	Apple_Inc
IPhone_5	IPhone_5
Siri_(software)	Siri_(software)
OS_version_history	OS_version_history
IPhone_3GS	IPhone_3GS
IOS	IOS
Smartphone	Smartphone
IPod_Touch	IPod_Touch
List_of_iOS_devices	List_of_iOS_devices
IPad	IPad
Apple_A5	Apple_A5
Steve_Jobs	Steve_Jobs
Verizon_Wireless	-
Gizmodo	-
Apple_A4	Apple_A4
App_Store_(iOS)	App_Store_(iOS)

Google’s search engine

Google’s search engine result	Google’s search engine relevant results
iPhone 4	iPhone 4
iPhone	iPhone
iPhone 4S	iPhone 4S
iOS	iOS
iOS jailbreaking	iOS jailbreaking
Retina Display	Retina Display
iOS version history	iOS version history
Apple A4	Apple A4
List of displays by pixel density	-
iPhone (original)	iPhone (original)
History of iOS jailbreaking	History of iOS jailbreaking
iPhone Dev Team	iPhone Dev Team
List of Apple Inc. slogans	-
iPhone 3GS	iPhone 3GS
Apple A5	Apple A5
iPhone 5	iPhone 5
George Hotz	-
Gizmodo	-
Siri (software)	Siri (software)
Greenpois0n	-

Query: Indochina

TCSearch2 TT

TCSearch2 TT result	TCSearch2 TT relevant results
First_Indochina_War	First_Indochina_War
French_Indochina	French_Indochina
Indochina	Indochina
Indochina_Wars	Indochina_Wars
Indochina_Airlines	Indochina_Airlines
Invasion_of_French_Indochina	Invasion_of_French_Indochina
Second_French_Indochina_Campaign	Second_French_Indochina_Campaign
Geneva_Conference_(1954)	-
Northern_Indochina_subtropical_forests	Northern_Indochina_subtropical_forests
Battle_of_Dien_Bien_Phu	Battle_of_Dien_Bien_Phu
War_in_Vietnam_(1954–1959)	War_in_Vietnam_(1954–1959)
List_of_Governors-General_of_French_Indochina	List_of_Governors-General_of_French_Indochina
Operation_Camargue	Operation_Camargue
Vietnam_War	Vietnam_War
Military_history_of_Cambodia	Military_history_of_Cambodia
Indochina_Expeditionary_Army	Indochina_Expeditionary_Army
History_of_Vietnam	History_of_Vietnam
Organization_of_Japanese_forces_in_Southeast_Asia	Organization_of_Japanese_forces_in_Southeast_Asia
Indochina_War_timeline	Indochina_War_timeline
Southeastern_Indochina_dry_evergreen_forests	Southeastern_Indochina_dry_evergreen_forests

TCSearch2 QA

TCSearch2 QA result	TCSearch2 QA relevant results
First_Indochina_War	First_Indochina_War
French_Indochina	French_Indochina
Indochina_Wars	Indochina_Wars
History_of_Vietnam	History_of_Vietnam
Vietnam_War	Vietnam_War
Vietnam	Vietnam
Indochina	Indochina
War_in_Vietnam_(1954–1959)	War_in_Vietnam_(1954–1959)
Indochina_Airlines	Indochina_Airlines
Battle_of_Dien_Bien_Phu	Battle_of_Dien_Bien_Phu
Hanoi	Hanoi
Vietnamese_National_Army	Vietnamese_National_Army
Operation_Camargue	Operation_Camargue
Geneva_Conference_(1954)	-
Second_French_Indochina_Campaign	Second_French_Indochina_Campaign
Military_history_of_Cambodia	Military_history_of_Cambodia
Vietnam_People's_Army	Vietnam_People's_Army
France–Vietnam_relations	France–Vietnam_relations
North_Vietnam	North_Vietnam
Invasion_of_French_Indochina	Invasion_of_French_Indochina

TCSearch2 QAUS

TCSearch2 QAUS result	TCSearch2 QAUS relevant results
First_Indochina_War	First_Indochina_War
French_Indochina	French_Indochina
Indochina_Wars	Indochina_Wars
History_of_Vietnam	History_of_Vietnam
Vietnam	Vietnam
Vietnam_War	Vietnam_War
Indochina	Indochina
War_in_Vietnam_(1954–1959)	War_in_Vietnam_(1954–1959)
Battle_of_Dien_Bien_Phu	Battle_of_Dien_Bien_Phu
Hanoi	Hanoi
Indochina_Airlines	Indochina_Airlines
Vietnamese_National_Army	Vietnamese_National_Army
Operation_Camargue	Operation_Camargue
Vietnam_People's_Army	Vietnam_People's_Army
Geneva_Conference_(1954)	-
Second_French_Indochina_Campaign	Second_French_Indochina_Campaign
Military_history_of_Cambodia	Military_history_of_Cambodia
France–Vietnam_relations	France–Vietnam_relations
North_Vietnam	North_Vietnam
Laotian_Civil_War	Laotian_Civil_War

TCSearch2 QAUSPR

TCSearch2 QAUSPR result	TCSearch2 QAUSPR relevant results
Vietnam	Vietnam
French_Indochina	French_Indochina
First_Indochina_War	First_Indochina_War
Vietnam_War	Vietnam_War
Hanoi	Hanoi
South_Vietnam	South_Vietnam
North_Vietnam	North_Vietnam
Viet_Minh	Viet_Minh
Ho_Chi_Minh_City	Ho_Chi_Minh_City
Ho_Chi_Minh	Ho_Chi_Minh
Laos	Laos
Indochina	Indochina
Cambodia	Cambodia
Geneva_Conference_(1954)	-
Battle_of_Dien_Bien_Phu	Battle_of_Dien_Bien_Phu
Bao_Dai	Bao_Dai
Vietnam_People's_Army	Vietnam_People's_Army
Viet_Cong	Viet_Cong
Ngo_Dinh_Diem	Ngo_Dinh_Diem
China	China

Wikipedia

Wikipedia result	Wikipedia relevant results
Indochina	Indochina
French_Indochina	French_Indochina
First_Indochina_War	First_Indochina_War
Vietnam	Vietnam
Vietnam_War	Vietnam_War
Indochina_Wars	Indochina_Wars
Indochine_(film)	-
Franco-Thai_War	-
Postage_stamps_and_postal_history_of_Indochina	Postage_stamps_and_postal_history_of_Indochina
Compendium_of_postage_stamp_issuers_(Ia_â€“_In)	-
Indochine_(band)	-
China_Records	-
Names_of_Cambodia	Names_of_Cambodia
Japanese_invasion_of_French_Indochina	Japanese_invasion_of_French_Indochina
Second_French_Indochina_Campaign	Second_French_Indochina_Campaign
Indochina_Airlines	Indochina_Airlines
Indochina_mangroves	Indochina_mangroves
French_Indochinese_piastre	French_Indochinese_piastre
Northern_Indochina_subtropical_forests	Northern_Indochina_subtropical_forests
Indochina_War_timeline	Indochina_War_timeline

Google

Google result	Google relevant results
Indochina	Indochina
French Indochina	French Indochina
Postage stamps and postal history of Indochina	Postage stamps and postal history of Indochina
Indochina Wars	Indochina Wars
Japanese invasion of French Indochina	Japanese invasion of French Indochina
First Indochina War	First Indochina War
List of Governors-General of French Indochina	List of Governors-General of French Indochina
Political administration of French Indochina	Political administration of French Indochina
Second French Indochina Campaign	Second French Indochina Campaign
Indochina Airlines	Indochina Airlines
Banque de l'Indochine	Banque de l'Indochine
Northern Indochina subtropical forests	Northern Indochina subtropical forests
French Indochinese piastre	French Indochinese piastre
Indochina Migration and Refugee Assistance Act	Indochina Migration and Refugee Assistance Act
Central Indochina dry forests	Central Indochina dry forests
Communist Party of Indochina	Communist Party of Indochina
Indochina Media Memorial Foundation	Indochina Media Memorial Foundation
Indochinese leopard	-
Indochina War timeline	Indochina War timeline
Indochinese tiger	-