NTNU – Trondheim
Norwegian University of
Science and Technology

Improving Search in Social Media Images with External Information

Mathilde Ødegård Oftedal
Marte Johansen Sæther

Master of Science in Informatics
Submission date: June 2014
Supervisor: Herindrasana Ramampiaro, IDI

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

Improving Search in Social Media with External Information

Marte Johansen Sæther and Mathilde Ødegård Oftedal

June 5, 2014

Preface

This Master thesis is delivered to department of Computer and Information Science, at the Norwegian University of Science and Technology. The thesis is part of the master program in Informatics, and done with guidance from Associate Professor Heri Ramampiaro.

June, 2014

Abstract

The use of social media has increased considerably the recent years, and users share a lot of their daily life in social media. Many of the users upload images to photo-sharing applications, and categorize their images with textual tags. Users do not always use the best tags to describe the images, but add tags to get ”likes” or use tags as a status update. For this reason, searching on tags are unpredictable, and does not necessary return the result the user expected.

This thesis studies the impact of expanding queries in image searches with terms from knowledge bases, such as DBpedia. We study the methods TF-IDF, Mutual Information and Chi-square to find related candidates for query expansion. The thesis reports on how we implemented and applied these methods in a query expansion setting. Our experiments show that Chi-square is the method that yields the best result with the best average precision, and was slightly better than a search without query expansion. TF-IDF gave the second best result with query expansion, and Mutual information was the method that gave the worst average precision. Query expansion with related terms is an exiting field, and the information from this thesis gives a good indication that this is a field that should be more explored in the future.

Keywords: Image search, social media, tagging, Flickr, DBpedia, query expansion, TF-IDF, Chi-square, Mutual information

Sammendrag

Bruk av sosiale medier har økt drastisk de siste årene, og mennesker deler mye av hverdagen sin ved hjelp av sosiale medier. Det er blitt vanlig å legge ut bilder på fotodelingssider der man kategoriserer bildene ved hjelp av tekstlige tagger. Brukerne benytter ikke alltid de beste taggene når de skal beskrive bilder, og legger til tagger for å få "likes", eller bruker det som en statusoppdatering. Søk etter bilder med tags kan derfor gi et noe uforutsigbart resultat, og er ikke alltid det brukeren søker etter.

Denne oppgaven ser nærmere på hvordan spørreutvidelse med termer fra kunnskapsdatabaser, som DBpedia, kan forbedre bildesøk. Vi undersøker metodene TF-IDF, Mutual information og Chi-square for å finne kandidater til spørreutvidelsen. Chi-square var metoden som returnerte det beste bilderesultatet når det kommer til gjennomsnittlig presisjon, og var kun litt bedre enn et søk uten spørreutvidelse. TF-IDF var metoden som ga det nest beste resultatet med spørreutvidelse, og Mutual information var den metoden som hadde den dårligste gjennomsnittlige presisjonen.
Spørreutvidelse med relaterte termer fra kunnskapsdatabaser er et spennende område, og informasjonen fra denne oppgaven gir en god indikasjon på at dette er et felt som burde bli utforsket mer i fremtiden.

Nøkkelord: Bildesøk, sosiale media, Flickr, DBpedia, spørreutvidelse, TF-IDF, Chi-square, Mutual information

Acknowledgements

We would like to give our sincere thanks to our supervisor, Heri Ramampiaro, for helping us choose this thesis, for programming help, and guidance through the process. Thanks to our families and boyfriends, for support and help through this year. And last we would like to give our thanks to everybody at Sule, for laughter, help with coding, and general procrastination.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and motivation

1.1 Photo-sharing in social media

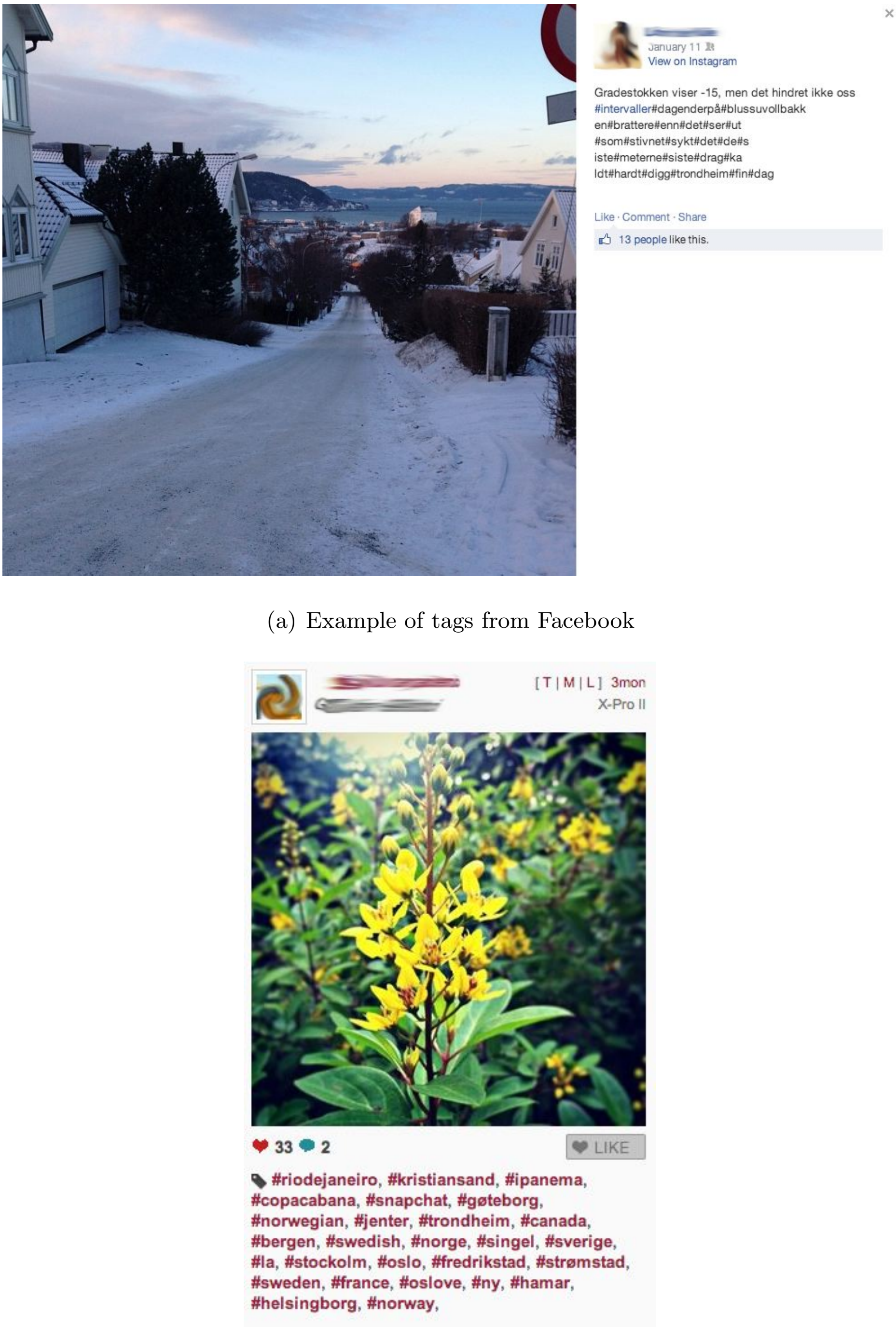
The use of social media has escalated the last years. According to “The Social Media Report, 2012”, by Nielsen [25], people in the US spend 20 % of their time on PCs, and 30 % of their mobile time, on social media. The time used on social media had increased with 24 % from 2011 to 2012. Social media, like Facebook and Twitter, as well as online photo-sharing apps like Instagram, Pinterest and Flickr, has facilitated sharing of pictures online. Sharing personal information is a large part of being active on social media. According to the Instagram blog, the amount of pictures shared on Instagram went from 5 billion [16] to 16 billion [17] between 2012 and 2013, and the total number of pictures shared on Flickr has passed 8 billion [8].

Users categorize pictures through tagging (or hashtagging), and if the user wants to find a picture, he or she must base their search on tags. This can, however, cause several challenges. An increasing trend is that users tag pictures just to get ”likes”, this includes using tags that not necessarily are relevant to the picture, or just use the tags as a status update. The two examples in Figure 1.1, are from Facebook and Instagram, and are, based on experience, typical examples of how users tag their pictures.

In Figure 1.1(a) the hashtags create a sentence which translated to English are: intervals "the day after" Blussuvollbakken steeper than it looks stiffened crazy the last meters last round cold hard delicious Trondheim nice day. In Figure 1.1(b), an example from Instagram, shows how users add tags just to get "likes". Both are good examples of irrelevant tag usage. Approximately 3-4 of 20 tags are relevant in the first example, wheras none of the tags in the second are.

Several articles mention problems with user-generated tags. Collaborative tagging environments and folksonomies[1] are known for tag spamming [13], and bad quality. In fact, only about 40-50 % of the tags are relevant to the image [9]. Tag ambiguity and tag synonyms are challenging as well. Imprecise and incomplete tags result in poor results when searching for images. The search results are based on the relationship between the tag and the image, and when users choose tags this implies that the result is based on the users relationship with the image [9]. Working with tags are challenging because they are user-generated. We want to get past the problem with tagging without changing the tags, and without input from the user.

[1] A folksonomy is a classification system that is created in collaboration with all the users



[redacted]
January 11
View on Instagram

Gradestokken viser -15, men det hindret ikke oss #intervaller#dagenderpå#blussuvollbakken#brattere#enn#det#ser#ut#som#stivnet#sykt#det#de#siste#meterne#siste#drag#kaldt#hardt#digg#trondheim#fin#dag

Like · Comment · Share

13 people like this.

(a) Example of tags from Facebook

[redacted]

[T | M | L] 3mon
X-Pro II

33 2 LIKE

#riodejaneiro, #kristiansand, #ipanema, #copacabana, #snapchat, #gøteborg, #norwegian, #jenter, #trondheim, #canada, #bergen, #swedish, #norge, #singel, #sverige, #la, #stockolm, #oslo, #fredrikstad, #strømstad, #sweden, #france, #oslove, #ny, #hamar, #helsingborg, #norway,

(b) Example of tags from Instagram

Figure 1.1: Examles of tagging on social media

1.2 Information retrieval

Information retrieval is about providing the user with easy access to the information they search for. In 1991 the first website was created, and by June 2014 it will be approaching 1 billion websites [18]. To be able to handle all this information, information retrieval systems and search engines are needed. As stated in [2]: *"Information retrieval deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects"*.

The problem with information retrieval

According to [2], the problem with information retrieval can be defined as *"The primary goal of an IR system is to retrieve all the relevant documents to a user query while retrieving as few non-relevant documents as possible"*. But which documents are relevant? And are the same documents relevant to all users and in every situation? The answers to these questions are subjective, users find different documents relevant. With some user input, information retrieval (IR) systems are able to detect relevant documents, and give the user a satisfying result. This user input is usually query terms, and the IR system uses these words as an indication to what information the user is interested in.

1.3 Research question

The user expect relevant documents and images, without using too much effort to find them. Users often just write the first word that comes to mind

as the query, and expect the system to understand the context. The extreme amount of sharing on social media creates challenges for information retrieval, as well as the low quality of the tags associated with the images. A good source for information is Wikipedia[2], which is an open editable information source where users can edit the information. The structured information on Wikipedia is extracted and added to DBpedia[3], and can be used on the web. We want to give the user a relevant result, without them having to specify a complete query with both query terms and the setting. We are suggesting an approach that exploits information connected to the users queries, and with this information generating a search after pictures based on their tags.
In view of this, the main research question can be formulated as follows:

RQ: *Is it possible to improve search results in social media, such as Flickr, by using additional metadata from a structured database, like DBpedia?*

This principal research question can be divided into the following subquestions:

RQ1: *Does a system like this already exist?*

RQ2: *Can this be done with query expansion, without feedback from the user?*

RQ3: *What method gives the most relevant result?*

[2] http://en.wikipedia.org/wiki/Wikipedia:About

[3] http://dbpedia.org/About

1.4 Outline

In Chapter 2 we will go through relevant theories for both social media and information retrieval, and related work to our approach will be discussed. Using this theory, we have made a proposed approach, this approach is described in Chapter 3. In Chapter 4 the evaluation of the approach is described, and in Chapter 5 we present the results from the evaluation. In Chapter 6 the results are being evaluated and discussed, at the end of the chapter we will answers the research quetions. In Chapter 7 a final conclusion is drawn, and future work explained.

Chapter 2

State of the art

This chapter will describe different theory and information used for this thesis. First photo-sharing applications will be described, and then basic theory from Information retrieval and methods used in this approach are discussed and described. Finally an overview and evaluation of related work.

2.1 Photo-sharing through social media

This section will present the different social media applications for photo-sharing.

2.1.1 Facebook

The most popular social media platform is **Facebook**, with 1,31 billion active users [37]. The consumers spend 17% of the time spent on computers on Facebook [25]. Facebook is 10 years in 2014, and has managed to stay popular and inventive all these years. In 2013, Facebook also included hashtags so that the users can tag pictures and status updates [1]. Today Facebook is used to communicate with people, create groups for different interests, sharing of images and advertisment. Facebook's search function varies from which language the users use on Facebook. Users having english US, have a search function called graph search. The graph search function is used to find people and interests that are connected in some way, for example they support the same soccer team, or have gone to the same university.

Figure 2.1: The Facebook logo

2.1.2 Instagram

Instagram is a picture sharing media, first launched as an application for smartphone in 2010. Currently there are 150 million monthly active users, and over 16 billion shared photos [26]. It is only possible to search after one hashtag when searching on Instagram. This leads to less relevant results. The pictures returned are only sorted by most recent.

Figure 2.2: The Instagram logo

[1] http://www.bbc.com/news/technology-22882119

2.1.3 Twitter

Twitter was launched in 2006, and is a medium where the users can express their meanings through tweets of 140 characters. The hashtag was introduced in 2007, and was used as a way of categorizing tweets. Twitter was the first social media with hashtags, and now almost every social media uses them. In 2011 Twitter integrated a photo-sharing service that made it possible for users to include a picture to their tweet [39].

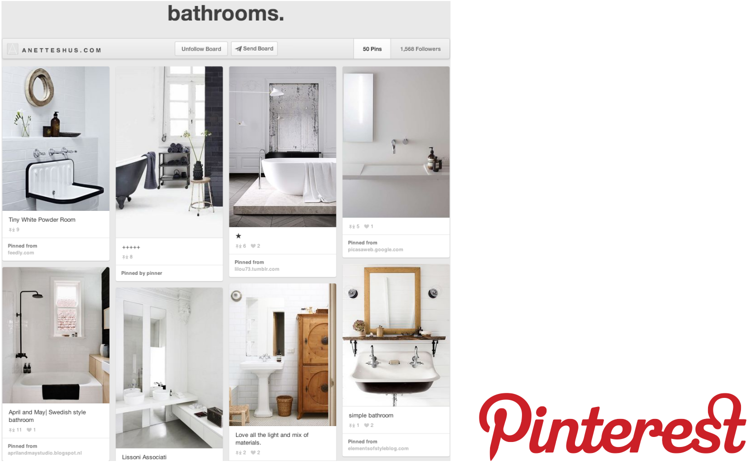
Figure 2.3: The Twitter logo

2.1.4 Pinterest

Pinterest is an application where it is possible to collect and organize things that interest the user. The user finds pictures on the Internet, and pins them to a Pinterest board. A board on Pinterest can therefore contain pictures of, for example, different attractions the user wants to visit, pictures from an inspiring blog, or gathering inspiration for a wedding [34]. An example of a board on Pinterest can be viewed Figure 2.4(a). Pinterest experienced an extreme grow the first year, and in 2012 they had the largest year-over-year increase of any social network in audience and time spent [25].

2.1.5 Flickr

Flickr is one of the most used photo-sharing application, and was launched in 2004. Both the owner of the photos, and others can tag the pictures on



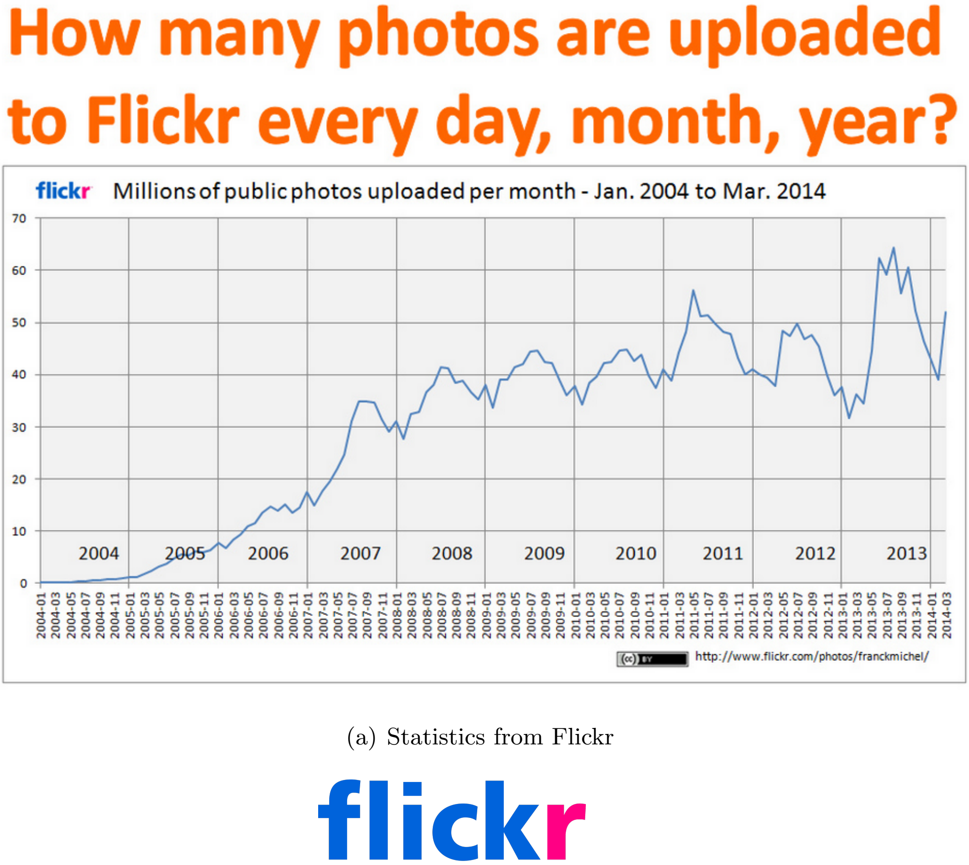
(a) Example of board on Pinterest (b) Pinterest logo

Figure 2.4: Pinterest

Flickr. Flickr is one of the first examples of a folksonomy, or collaborative tagging environments. They have, according to themselves, the best photo sharing community. Users can both share their photos, and can use 1000GB free for storing them in the Flickr cloud, it is also possible to pay for more space. Many professional photographers and bloggers store their photos on Flickr. In 2011 they announced that more than 6 billion photos had been uploaded to their site [2], and in 2012 and 2013, over 1100 million photos was uploaded [12]. Picture Figure 2.5(a) shows some statistics from Flickr. Most of the photos are public, but it is also possible to have private photos. [11]

For our approach we chose to use Flickr as the social media to search after photos in. We chose Flickr because it is the most used photo-sharing application. It contains large amounts of images, from different years, and many of these images contain hash tags of different quality. Another reason

[2] http://blog.flickr.net/en/2011/08/04/6000000000/



(a) Statistics from Flickr

(b) Flickr logo

Figure 2.5: Flickr

why we chose Flickr was that we could use a large dataset that already had been processed, and was ready to use.

2.2 Information retrieval

Information retrieval provides a way to store, represent and organize information so that the user easy can get access to data of their interest. To give users a good experience and meet their needs, interfaces are provided so that a user can query for information. The goal is to provide users with the most relevant documents from their point of view, build fast ranking algorithms and make effective indexes to do a search effective. It is important to distinguish the difference between data retrieval and information retrieval. Data retrieval concerns mainly with finding which documents that contains some keywords, where information retrieval focus on finding information about a specific topic or subject.

2.2.1 Preprocessing documents

"An index is a data structure built on the text to speed up the search"([2] pp.338). This means that words in a document are structured in a way that makes it easy to find specific words or sentences fast, and makes the ranking better. Indexing documents also requires less space than storing large amount of information. Below the process of the document preprocessing procedure will be explained. This process happens before the docuements can be indexed.

1. Lexical analysis: Lexicals analysis transforms a stream of characters to words. This means that all the words can be converted to lowercase

characters, removal of punctuations marks and hyphens. For example "Paris is the capital of France, France is in Europe". After lexical analysis: "paris is the capital of france france is in europe".

2. Stopwords: Stopwords are common words that are being used widely in a document, and can make a document more relevant than it is. To avoid that irrelevant document are returned, removal of stopwords can be done before indexing a document. There can be various stopwords for documents, so before stopword removal, it is important to understand the content of the document. Example of stopwords are: the, is, he, she.

3. Stemming: In a document there will be words in different conjugation forms, when doing a search the returned result might be different than expected even if you know that the documents contains a specific word. To avoid this problem, stemming can be applied on words, this means that the ending on word is removed, and only the stem of the word is left. There have been a discussion on how effective really stemming is, and many search engines do not use a stemming algorithm. An example: walking, walked → walk.

4. Index terms: Depending on how a text is represented, the choice of index terms can either be all the words in a document, this is a full text representation of the text. Keyword term index is an index where specific words in a document can be used as index terms. These words are specifically selected for the document, either from taxonomy or a vocabulary. A index term in a computer science paper can be Java, database and MySQL.

5. Thesaurus: *"The word thesaurus has Greek and Latin origins and is used as a reference to treasury of words"*[2]. Thesaurus is a collection of words that relates to each other either in similarity or meaning. An example of a thesaurus: connect → relate, join, associate.

Depending on the document being indexed, and which information the system will search for, these method must be evaluated to find which suites the system best. [38][22][20]

In Figure 2.6 the process of indexing, retrieval and ranking can be seen. The query from the user is modified first by removing stopwords. Further, more information can be given to the query, to give the user a better result. The retrieved documents are then ranked after relevance. Finally, the results are formatted before they are presented to the user. In addition to the query process, documents have to be indexed. This process is seen in the right part of the figure. Here text transformation is done before index terms are chosen.

The dataset that were downloaded from DBpedia needed some document preprocessing before it could be indexed, and used in our approach. We found it important to remove stopwords, in order to reduce the amount of data being indexed, and remove unimportant words. We also used lexical analysis, so that all the words were in lowercase, removed punctation and commas.

2.2.2 Searching

There are several available query methods, and the most used one is keyword-based queries. Here the user give some keywords to the system and in return a list of relevant documents are provided. Keyword-based querying is popu-

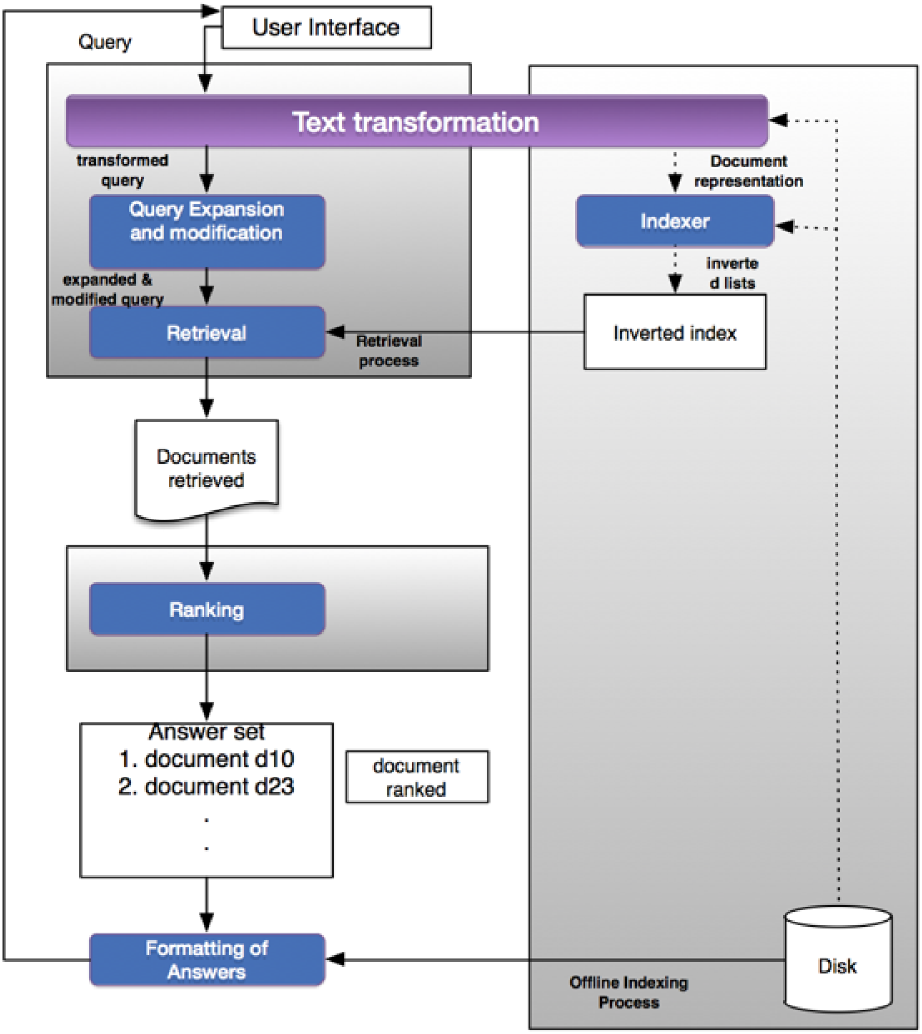


Figure 2.6: Indexing, retrieval and ranking process of documents.

lar in web search engines. The method provides an easy way to do a search for information and the ranking is done effectively. Other more complex query methods are pattern matching and fixed structures.

Query languages

Single word

Single word queries are the simplest form of queries. A single word is given to the search system, and all documents that contain this word are retrieved. To rank the retrieved documents, the occurrence of the word in each document can be counted, and documents are ranked in order of the occurrence.

Boolean queries

In Boolean queries, keywords are expressed as a Boolean expression, and operators are used to work on these operands. The Boolean operators used are OR, AND and NOT. Drawbacks of Boolean queries are that a document is either relevant or not, it does not tell how many times a word occurs in document. This means that partial matching is not an alternative, and no ranking is used.

Pattern matching

Pattern matching can be used if a query matches one of the words in a sentence/pattern. The use of pattern matching is often used in linguistics, text statistics and data extraction to form basic queries. There exist several patterns, some of them are word-, prefix- and substring-patterns. Word is basic pattern, a word consist of characters that must be found in search data. Prefix pattern is when only the beginning of a word is used in a search. An example is the prefix "info", then a search only look after word beginning with "info", like "information" and "informational".

Substring patterns are a string in a word, this means all words that contains this substring is retrieved. For instance if the query is "ting", all words containing "ting" will be retrieved, this can be "marketing", "acting", "skating" and so on.

For our approach we have chosen to use Boolean queries for the search in Flickr. With Boolean queries we can decide how the search will be executed. Either to include all words, or exclude some of the words. For our method the AND operator would be best suited, because we are expanding the search word with another. The returned result must therefore contain both the query typed, and the expansion found. Searching in DBpedia is done with Lucene, more about this in Section 2.2.7. [2][22]

2.2.3 Evaluation of an information retrieval system

To evaluate how good an information retrieval system is, the returned results must be analyzed. As referred in ([2] pp.131), *"Retrieval evaluation is a process of systematically associating a quantitative metric to the results produced by an IR system in response to a set if user queries. This metric should be directly associated with the relevance of the results to the users. A common approach to compute such a metric is to compare the results produced by the system with results suggested by humans for the same set of queries"*. It is important to distinguish between the evaluation of the performance of an information retrieval system, and the quality of the retrieved results. In the evaluation, only the results retrieved by the system is of relevance. A good information system is one that satisfies the users needs, therefore are the returned results important. Testing an information retrieval system evaluates the retrieval of relevant and non-relevant documents. A document is relevant if it contains the information that the user searched for. When evaluating information systems, not only the words used for searching are important, but that the content of the documents as well. There exists several methods for evaluating IR, some of them are precision, recall and P@n. [2] [22]

2.2.3.1 TF-IDF

TF (term frequency) is how many times a term k occurs in a documents d. In Equation 2.1 the term frequency $f_{i,j}$ is used to compute TF. The more often a term occurs in document d, the higher the term frequency is.

$$tf_{i,f} = 1 + log f_{i,f} \tag{2.1}$$

IDF, or inverse document frequency, tells if a term occurs often or rare in documents. Equation 2.2 uses the total number of documents in a collection divided by number of documents where a specific term exists. To find the IDF, the number of documents in the collection are divided on number of documents that contains a specific term. [2][22]

$$IDF_i = log\frac{N}{n_i} \tag{2.2}$$

2.2.3.2 Precision

Precison is a measure that finds the fraction of retrieved documents in a collection that are relevant. The equation $p = \frac{|R \cap A|}{|A|}$ uses the numbers of relevant retrieved documents diveded by the numbers of retrieved documents. [2][20]

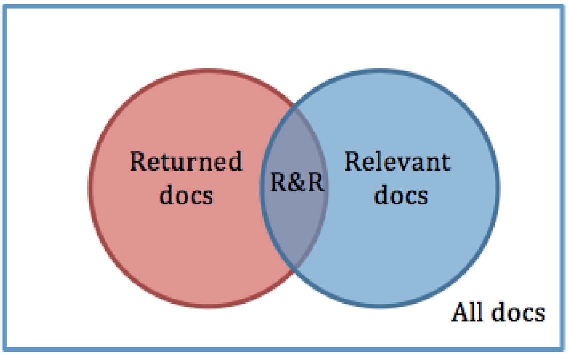


Figure 2.7: The relationship between presicion and recall

2.2.3.3 Recall

Recall measures the fractions of relevant documents that are retrieved from a collection. The equation $r = \frac{|R \cap A|}{|R|}$ uses the numbers of relevant documents retrieved diveded by numbers of relevant documents in the collections.

The relationship between precision and recall are illustrated in Figure 2.7. Precision is the part of the returned documents that is relevant (R&R). Recall is the part of the relevant documents that is returned. [22][2]

2.2.3.4 P@n

P@n, or precison at n, is a list of ranked terms were the top-n documents are the first n ranked. Precision at n's equation is, $P@n = \frac{r}{n}$, where r is the number of relevant documents, and n are documents number. This measure describes how pleased the user is with the results, often the user a more satisfied if the first n documents are the relevant ones. [2][20][22]

2.2.4 MAP

MAP(mean average precision) is a measurement that calculates the average precision for n queries. To calculate MAP, Equation 2.3 is used. The precision for each query (AP_n) is divided by the total number of queries(n). This measurement has been used a the "gold standard" [3] to test if the system is working as planned. [22] [20]

$$MAP = \frac{1}{n} \sum_{n} AP_n \tag{2.3}$$

2.2.5 Query expansion

A technique for giving the user a better search result is to expand the query. The first way this was done was by a thesaurus. The thesaurus keeps information about synonyms and related words of phrases from the document. These can be used to expand the initial query. Now it is more common to use a semi-automatic query expansion technique, where the user chooses suggested terms from a list [5].

2.2.5.1 Global analysis

Query expansion can be classified into two main classes: global – and local analysis. Global analysis was one of the first techniques that produced consistent and effective results with query expansion. To do global analysis it is necessary to have corpus-wide statistics, which can result in a similarity matrix. The words that are most similar to the query are added to the query. This is a robust technique, but it requires a lot of data resources.

[3] Gold standard - binary classification as relevant or nonrelevant on a document

Examples of global analysis are co-occurrence of pair of terms, Latent Semantic Indexing, and, similarity thesauri.

Co-occurrence

Term clustering and co-occurrence are measures that measure the number of times groups of words (usually pairs) occur together in document. A collocation is a pair of words that occur together more often that would be expected by chance. Term association measures are used to find collocations [5].

Mutual information

Mutual information is one of the measures used in collocations, and it measures the extent to which the words occur independently, if the value returned is zero they are completly independent. Mutual information computes the relative entropy between two terms Equation 2.4 , the higher the value, the more relevant the terms are to each other. Research has shown that this measure tends to favor low-frequency words, and this can be a problem. The expected mutual information measure try to solve this problem by using probability to weight the mutual information value [5].

$$MI(k_i, C) = \frac{n_{ab}}{n_a \times n_b} \tag{2.4}$$

Pearsons's Chi-square

Pearson's Chi-squared measure is also a popular association measure. Chi-square compares the expected co-occurrence, if the terms are independent, with the number of co-occurrence of two words. Two terms are independent if

$$P(AB) = P(A)P(B) or P(A|B) = P(A) and P(B|A) = P(B) \tag{2.5}$$

Then the measure is normalized by the expected measure [5]. Chi- square evaluates the different between the results retrieved and the expected result in a collection, Equation 2.6.

$$\chi^2(D,t,c) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} \left(\frac{N_{e_t e_c} - E_{e_t e_c}}{E_{e_t e_c}} \right)^2 \tag{2.6}$$

To figure out which methods that give the best term for query expansion, we have chosen to use TF-IDF, Mutual information and Chi-square in our approach. These methods are well known and established. [5][22]

2.2.5.2 Local analysis

Local analysis only uses the initial retrieved documents for query expansion.

Relevance feedback

A technique for local analysis is relevance feedback, which is based on what the user judge as relevant in the retrieved documents. Then, to do query expansion, the additional query terms are selected from the relevant documents. If the user provides correct and sufficient information, relevance feedback achieves good performance. In practice relevance feedback does not achieve as good results as wanted, because users are reluctant to provide this information [4]. Recommendation systems are examples of different use of relevance feedback, where the system gives the user alternatives that are similar to the initial query. It is widely used on online shopping sites like eBay [4] and Amazon [5]. [5]

Pseudo-relevance feedback

Pseudo-relevance feedback is used to overcome the difficulties from relevance feedback. This involves local feedback to mimic relevance feedback by assuming the top-ranked documents to be relevant. The expanded query adds terms by choosing the words that are most frequent in the assumed relevant documents. Pseudo-relevance only works if the retrieved documents are relevant, and therefore, the quality of the result strongly depends on the initial retrieval. This is why the technique is not incorporated in operational search applications; the results can be unpredictable as a cause of the automatic process [4] [5].

Our approach will use a sort of pseudo-relevance feedback; in addition to the high frequency words will we also use Mutual information and Chi-square to calculate candidates for query expansion. We do not know how relevant the returned documents from DBpedia are, and this will be a factor for the result and affect the choice of terms for query expansion.

Combining global and local analysis

There are also examples on systems that combine global and local analysis. An example is provided with the article "Probabilistic Query Expansion Using Query Logs" [4]. They describe how it is possible to exploit the accumulated information on user interactions, to do query expansion. Through the search log, they find out which queries that led to which documents, and use data mining to find a relationship between the terms in the query, and

[4] www.ebay.com

[5] www.amazon.com

in the document. Each session contains one query, and a set of documents that the user clicked on. The assumption is that a document is relevant if it is clicked on. This is a valid claim because the clicked documents are the top-ranked documents, as well as the user has made a selection among these documents. The experiments done on the method showed that the log-based method could achieve substantial performance improvements.

2.2.6 Web Retrieval

As stated in ([22] pp.2) *"the system has to provide search over billions of documents stored on millions of computers"*, this means that the system has to find a method to handle the various ways data are stored in. These data can be unstructured, redundant, heterogeneous or of bad quality. In order to overcome these challenges, a good system architecture and fast algorithm has to be used. Architectures used for search engines are centralized and distributed. Centralized architecture uses centralized crawler-indexer architecture. Crawlers are programs that traverse the Internet in search of new or updated information from websites to the server where they were indexed. The problem with centralized architecture is that the web is dynamic, and content are changing every minute. Distributed architecture uses harvesters to gather and distribute data, this make it more effective than centralized architecture. A harvester consists of two parts, gatherers and brokers. Gatherer collects and extracts information, while brokers retrieve this information. The drawback of it is that it requires coordination of many web servers; this can be problematic when a web server receives requests from several crawlers. Search engines uses two types of interfaces, one for the user query, and one for the answer. These must be user friendly, and give the user relevant ranked results. Like information retrieval, web retrieval also uses ranking algorithm to rank the given result. They both uses Boolean and Vector Space model but in different variations, developed to suite their needs. [5][1][10][42]

2.2.7 Lucene

Lucene is a free open source full-text search engine written in java. Lucene offers an API that is suitable for indexing and searching through lager amount of data. Many websites that needs a good search engine use Lucenes API to build this.

When doing a search with Lucene, the retrieved results are ranked by using vector space model and Boolean model. The Boolean model is first used to narrow down the retrieved result. The Boolean model uses set theory to find relevant documents. Each query is viewed as a Boolean expression, that use the operators AND, NOT or OR to solve this expression. In the Boolean model a word either exists or not, so there are no use of partial matching. All of the terms are equally weighted, so it is difficult to rank by term occurrence. But despite this, the model is effective and finds documents containing relevant terms, and it is easy to use and understand.

The Vector Space model on the other hand represent text as vectors, this means that index terms and documents gets a value that tells how relevant they are. If a word does not exist in the text it will have the value zero. To compute the value between the documents and the search query, TF-IDF is used. Figure 2.8 shows the similarity between a document and a query. It measures the degree of similarity between the document and the query. After finding TF-IDF the similarity between a document and a query can be

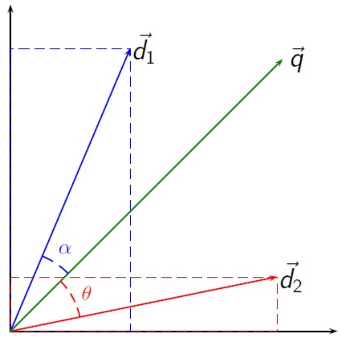


Figure 2.8: The similarity between documents and query

found, by using the similarity formula Equation 2.7. The similarity formula produces a weighting schema for all the terms. And the results produced are then returned to the user. [21][14][40]

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \cdot \overrightarrow{q}}{|\overrightarrow{d_j}| \times |\overrightarrow{q}|} \tag{2.7}$$

Solr

Solr is an open source based search engine built on Lucene. It provides full-text search, dynamic clustering and use of rich documents as some of its features. Solr is written in Java, just like Lucene, and uses API from Lucene to do search. [29][30]

2.2.8 Disambiguation

A word can have different meanings, one example is “wave”. “Wave” can either a wave at sea, or the verb wave. The problem that occurs here is ambiguity. Ambiguity is present in both query search and in the document retrieved. A user can avoid ambiguity by including additional words to the query, and then the system might understand the context. Ambiguity can be a problem for a computer only reading the word, and not the context of the content. To avoid that the wrong word is used, disambiguation is implemented to find which context the word is used in. Disambiguation uses dictionaries to find the meaning of a word, WordNet [6] is a popular online dictionary used for disambiguation. [7][28][43]

2.2.9 Synonyms

In social media people are using hash tags to describe an image and one problem that might occur in tagging is synonyms. Synonyms are described in [32]*" A word or phrase that means exactly or nearly the same as another word or phrase in the same language"* . Users are using different words for describing similar images. A user can tag an image with the word film, when another user is doing a query for the word movie, the image might not be in the returned result.[3]

2.3 Related work

The use of photos on web is widely used. Therefore is the need for an effective image retrieval system important. There are many research projects with different thesis about how the retrieval can be improved.

2.3.1 Improvement of TF-IDF

Popescu and Grefensette [35] suggest a system that uses Wikipedia and Flickr content to improve image retrieval. An improvement of TF-IDF that measured social relation was made. This improvement found which users that were associated to a tag in Flickr. They used the initial query

[6] http://wordnet.princeton.edu

in Flickr to compute co-occurring terms, and find nearby concepts through Wikipedia. The information that was retrieved was used to expand the query and compared the co-occurrence model. The proposed approach was tested on a noisy image collection, and the results were good. The new approach worked better than co-occurrence model, only the run time got worse.

Min et. al [23] use weighted TF-IDF with text-based image retrieval to find relevant images. The images metadata expanded with DBpedia were used to improve the images search. In order to find which important words that were connected to the images metadata, document expansion combined with document reduction were used. For the document expansion they used pseudo-relevance feedback method combined with the Okapi feedback algorithm. For the document reduction they used BM25, and removed terms under given cut-off value. The approach were tested on metadata of several languages, the best result were in English. The perfomance were improved when they used document expansion, and the findings showed that the combination of content-based image retrieval and text-based image retrieval methods performed better, than using single methods.

Both of these approaches uses TF-IDF to improve image retrieval, and their findings are interesting for our approach because the retrieved results are good. For our approach we will combine TF-IDF with other methods like mutual information and Chi-square, to see if the retrieved results will be improved.

2.3.2 Clustering

Clustering is a possible method to use for improving search results. Moëllic et al. describes a system that can exploit the relationships between tag and image, and the visual, if the images on Flickr are properly tagged. The proposed solution is based on shared nearest neighbor (SNN) clustering, but SNN consider only the tags associated with the image, not the visual. When a large amount of photos are associated to the same tag, these are sorted in a cluster for an effective overview. The algorithm resulted in good clusters that were more focused than the cluster that Flickr provides. [24]

A two-step approach for clustering on multimedia resources based on social, semantic and content features are also suggested [13]. The first step is to cluster based on tags. Then use the tags to analyze the semantic and social aspects. The second step employ content-based analysis of the resources, and does a cluster refinement. Based on social, semantic and content similarities, a similarity score is calculated. The experiments were done on WordNet and Flickr, with different clustering algorithms. For evaluation was all the clusters manually annotated. The conclusion was that the clustering method was robust, and that tag clusters can be used for semantic extraction and knowledge mining.

In our approach we will not use the method of clustering to improve image retrieval. We found these articles interesting because they focus on improvement of image retrieval, and gave some insight in what other have done earlier.

2.3.3 Improvement of tags

As mentioned in Section 1.1, if the photos on social media are properly tagged, the search results will become much more relevant. Liu et al. [19] has proposed a tag improvement solution that is based on consistency between visual similarity and semantic similarity of images. To test the system they did a query on the ten most popular searches and the search results were displayed by using "Ranking by interestingness". They obtained a dataset of 10 000 images, and 38335 unique tags, where many of the tags were misspelled or meaningless. After this they compared each tags to Wikipedia's thesaurus, and only tags in the thesaurus were kept. Then they computed the similarity between tags based on the co-occurrence. To evaluate the performance of the tag quality improvement they calculated the recall and precision to tags in each image, and then found the average. Users evaluated the relationship between tags and images, to decide if the tags and images were related. The results were good, but they used a lot of time deciding if the images contained the right tags.

Du et al. [9] propose an algorithm called Walking and Sleeping (WaS) to overcome the obstacle of bad-quality tags. This algorithm has several steps to find out if a tag is important to the image or not. The tags that have the most clicks are considered important (”walking” state), and the ones that are not are set in a ”sleeping” state. Only the tags that are in the ”walking” state are kept. Testing of the algorithm shows that it works, and top 25 tag recommendation results were about 40 % higher than using the test system.

A possible solution to overcome the problem of bad user tags are to automatically generate reliable and useful tags for multimedia content on social networks. Piatrick et al. [33] wants to exploit the full range of information available online to create user tags automatically. They want to predict user tags by using the associated metadata, expanded query terms, complementary resources and the photos visual features. Testing showed that the system benefitted from the complementary textual resources. When tested in an open-set annotation almost 40 % of the generated tags were considered relevant by manual annotators.

Improvements of tags are interesting field in image improval aspect. The tags are important, because they say something about an image. Misspelled tags, and the use of wrong tags in an image might contribute to a less good image retrieval. We are not going to improve the tags in our approach, but it is important to be aware of the problems that tags can might lead to. In this aspect, use of wrong terms and structured data.

2.3.4 Other relevant work

Vallet et. al. [41] suggest an approach to improve video retrieval by using content that satisfy personal interests. The user provides a query, and the system use the external collaborations Flickr and DBpedia in order to collect a set of images potentially relevant for query. These images were compared to key frames of videos available in the system; the videos with keyword similar to the images were retrieved. The result given by this approach shows that exploiting the semantics available in knowledge sources leads to sensible performance improvements compared to basics approach results. When they used external knowledge applied to manual query examples, the precision was improved. This showed that the use of knowledge sources could be successfully exploited to complement visual examples provided by the user.

An algorithm proposed by Qi et. al. [36] use both content and context-specific information to improve multimedia retrieval. To create a rich multimedia information network, they use multimedia objects like Flickr and YouTube, with content objects like tags and related attributes. Their assumption is that semantic concepts for annotations are correlated, and that a latent structure exists. The algorithm uses content links to enrich the multimedia information network, and geometric structures are created to show multimedia objects content. The results from the testing showed that the proposed algorithm was effective when it came to integrating content and contexts links.

The proposed approach these articles contains, shows interesting methods for improving multimedia retrieval by using external knowledge. For our approach we would use DBpedia as an external knowledge base, and it is interesting to see how other have used sources like this to improve retrieval with a good result. In our approach DBpedia will be used to find a term to expand the user query with, this in order to retrieve more relevant images.

Chapter 3

Our approach

This chapter will describe our approach. We begin with describing how the datasets were processed, and then we describe how the different methods perform calculations for query expansion. We also describe how the approach communicates with the user.

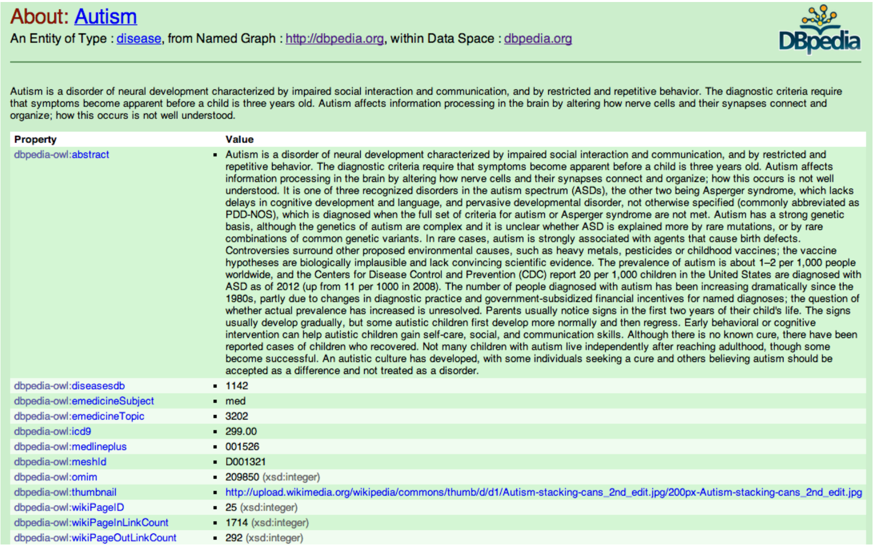
3.1 Datasets

The hypothesis is that by adding external information to a query, the result can be more relevant to the user. The idea is to choose a second term that is associated with the query term, and use it for query expansion.The second term is found by calculating the relevance to the query, by using one of the methods TF-IDF, Mutual Information or Chi-square, described in Chapter 2. DBpedia is used as the external data source, and the dataset is downloaded from [6]. Titles (label), short abstract and extended (long) abstracts from DBpedia were the initial information that we needed.

3.2 Pre-processing and Indexing

In order to use the datasets from DBpedia, they had to be indexed in a way that information easily could be accessed. Lucene was used for the indexing, and it was easy to specify which data that were interesting for our approach. After testing the datasets, the conclusion was that only the extended abstract dataset was of interest. The reason for this was that short abstract only were a short summary of the long abstract, and the titles only consisted of labels. For the query expansion to work most effective, the words associated to the query must be chosen. It is also important to have a pool of words to choose from. If the term chosen for query expansion was from the ”Label” field in DBpedia, it would most likely be the same as the query, or a stopword. The word would have a low term frequency, and also low Chi-square. This is a result of ”Label” only containing the label of the article from Wikipedia. This also applies to the ”Short abstract” from DBpedia. It is more likely to find relevant terms the longer the index is. Then the related measures will be more accurate, and there will be more terms to choose from. Long abstract gave good information about famous people, countries, historical events and other information that Wikipedia contains.

The long abstract dataset contains several fields, but the most interesting field for our approach was resource and the abstract. The reason for this, are that they contains the information that are needed to find a second term that is relevant to the search query, so that a query expansion can be performed. Figure 3.1 is an excerpt from DBpedia.org. One can see that an entity in DBpedia has several fields. This example is about autism



About: Autism

An Entity of Type : disease, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Autism is a disorder of neural development characterized by impaired social interaction and communication, and by restricted and repetitive behavior. The diagnostic criteria require that symptoms become apparent before a child is three years old. Autism affects information processing in the brain by altering how nerve cells and their synapses connect and organize; how this occurs is not well understood.

Property	Value
dbpedia-owl:abstract	▪ Autism is a disorder of neural development characterized by impaired social interaction and communication, and by restricted and repetitive behavior. The diagnostic criteria require that symptoms become apparent before a child is three years old. Autism affects information processing in the brain by altering how nerve cells and their synapses connect and organize; how this occurs is not well understood. It is one of three recognized disorders in the autism spectrum (ASDs), the other two being Asperger syndrome, which lacks delays in cognitive development and language, and pervasive developmental disorder, not otherwise specified (commonly abbreviated as PDD-NOS), which is diagnosed when the full set of criteria for autism or Asperger syndrome are not met. Autism has a strong genetic basis, although the genetics of autism are complex and it is unclear whether ASD is explained more by rare mutations, or by rare combinations of common genetic variants. In rare cases, autism is strongly associated with agents that cause birth defects. Controversies surround other proposed environmental causes, such as heavy metals, pesticides or childhood vaccines; the vaccine hypotheses are biologically implausible and lack convincing scientific evidence. The prevalence of autism is about 1–2 per 1,000 people worldwide, and the Centers for Disease Control and Prevention (CDC) report 20 per 1,000 children in the United States are diagnosed with ASD as of 2012 (up from 11 per 1000 in 2008). The number of people diagnosed with autism has been increasing dramatically since the 1980s, partly due to changes in diagnostic practice and government-subsidized financial incentives for named diagnoses; the question of whether actual prevalence has increased is unresolved. Parents usually notice signs in the first two years of their child's life. The signs usually develop gradually, but some autistic children first develop more normally and then regress. Early behavioral or cognitive intervention can help autistic children gain self-care, social, and communication skills. Although there is no known cure, there have been reported cases of children who recovered. Not many children with autism live independently after reaching adulthood, though some become successful. An autistic culture has developed, with some individuals seeking a cure and others believing autism should be accepted as a difference and not treated as a disorder.
dbpedia-owl:diseasesdb	▪ 1142
dbpedia-owl:emedicineSubject	▪ med
dbpedia-owl:emedicineTopic	▪ 3202
dbpedia-owl:icd9	▪ 299.00
dbpedia-owl:medlineplus	▪ 001526
dbpedia-owl:meshId	▪ D001321
dbpedia-owl:omim	▪ 209850 (xsd:integer)
dbpedia-owl:thumbnail	▪ http://upload.wikimedia.org/wikipedia/commons/thumb/d/d1/Autism-stacking-cans_2nd_edit.jpg/200px-Autism-stacking-cans_2nd_edit.jpg
dbpedia-owl:wikiPageID	▪ 25 (xsd:integer)
dbpedia-owl:wikiPageInLinkCount	▪ 1714 (xsd:integer)
dbpedia-owl:wikiPageOutLinkCount	▪ 292 (xsd:integer)

Figure 3.1: Example from DBpedia, Autism

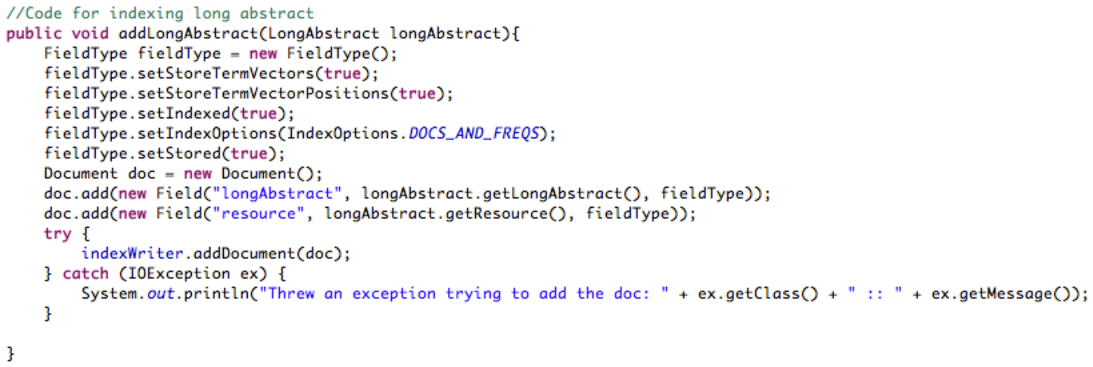
[1]. Before the indexing process, the long abstract consisted of sections of sentences like this:

<http://dbpedia.org/resource/Autism>

<http://dbpedia.org/ontology/abstract>*"Autism is a disorder of neural development characterized by impaired social interaction and communication, and by restricted and repetitive behavior."* This is taken from the dataset we downloaded.

Before the dataset were indexed, stopword removal was conducted on the datasets, and common words were removed from the dataset. A standardized stopword list that was found on the Internet [31] was used. This was because the DBpedia dataset consisted of genres of all kinds. Had the dataset for instance been a medical one, the stopword list had to be custom made. This is because medical terms are used in a different way than

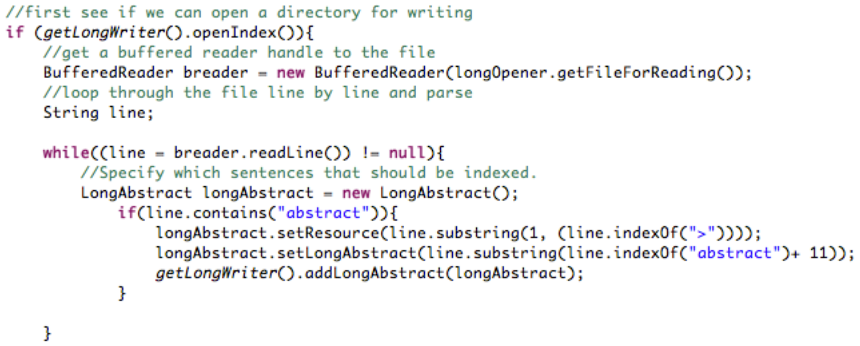
[1] http://dbpedia.org/page/Autism



```java
//Code for indexing long abstract
public void addLongAbstract(LongAbstract longAbstract){
    FieldType fieldType = new FieldType();
    fieldType.setStoreTermVectors(true);
    fieldType.setStoreTermVectorPositions(true);
    fieldType.setIndexed(true);
    fieldType.setIndexOptions(IndexOptions.DOCS_AND_FREQS);
    fieldType.setStored(true);
    Document doc = new Document();
    doc.add(new Field("longAbstract", longAbstract.getLongAbstract(), fieldType));
    doc.add(new Field("resource", longAbstract.getResource(), fieldType));
    try {
        indexWriter.addDocument(doc);
    } catch (IOException ex) {
        System.out.println("Threw an exception trying to add the doc: " + ex.getClass() + " :: " + ex.getMessage());
    }

}
```

Figure 3.2: Example from the code, method for indexing long abstract



```java
//first see if we can open a directory for writing
if (getLongWriter().openIndex()){
    //get a buffered reader handle to the file
    BufferedReader breader = new BufferedReader(longOpener.getFileForReading());
    //loop through the file line by line and parse
    String line;

    while((line = breader.readLine()) != null){
        //Specify which sentences that should be indexed.
        LongAbstract longAbstract = new LongAbstract();
            if(line.contains("abstract")){
                longAbstract.setResource(line.substring(1, (line.indexOf(">"))));
                longAbstract.setLongAbstract(line.substring(line.indexOf("abstract")+ 11));
                getLongWriter().addLongAbstract(longAbstract);
            }

    }
```

Figure 3.3: Example from the code, method setting long abstract

”normal” terms. If a standardized stop word list is used when indexing medical documents, the returned result would not have been as good as expected. The information in the URL and the abstract were stored in a LongAbstract object that was created during the processing of the dataset, Figure 3.2 show how this was done. This object was then used to create the index, this is viewd in Figure 3.3.

3.3 Searching DBpedia

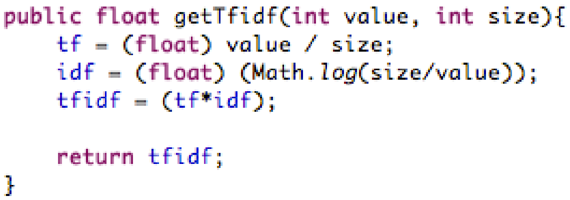
After the index is created it is possible execute queries against it. For our approach it was interesting to know how many terms that were associated to a query term. When the user enters a query, this query will be used to search through the index. All the retrieved documents are documents that contain the query. Then, to choose a second term to use for query expansion, different methods will be used to find candidates.

3.3.1 TF-IDF

Term frequency and inverted document frequency is one of the measurements that are used to select a second term. The approach calculates the TF-IDF of all second term candidates, and chooses the term with the highest TF-IDF. If a word have a high TF-IDF it is likely to be a word that is strongly connected to the query. In order to accomplish this it was necessary to get the frequency of all the candidate terms in the returned DBpedia result. This measure was used to calculate term frequency by dividing it on the amount of terms in the index. Figure 3.4 shows our method for calculating TF-IDF. The method receives a value, which is the term frequency of the desired term, and a size, that is the number of terms in the index. The method then makes the calculations, and returns the TF-IDF of the desired term.

3.3.2 Mutual information

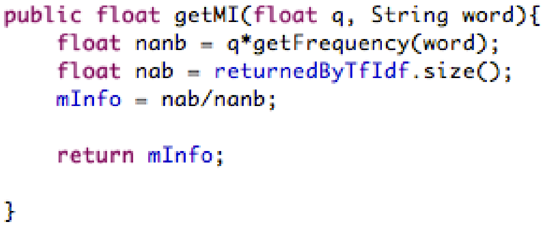
Mutual information is the second measurement that is used to choose the second term. The methods calculates alternatives for query expansion from a list of relevant terms generated from TF-IDF. For each relevant term



```
public float getTfidf(int value, int size){
    tf = (float) value / size;
    idf = (float) (Math.log(size/value));
    tfidf = (tf*idf);

    return tfidf;
}
```

Figure 3.4: Example from the code, method for calculating TF-IDF



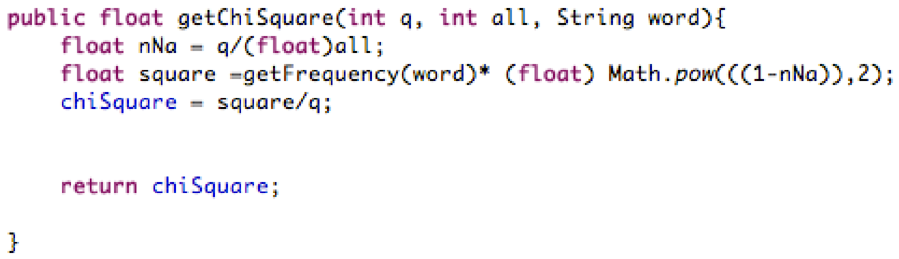
```
public float getMI(float q, String word){
    float nanb = q*getFrequency(word);
    float nab = returnedByTfIdf.size();
    mInfo = nab/nanb;

    return mInfo;

}
```

Figure 3.5: Example from the code, method for calculating mutual information

mutual information is calculated, only the terms with the highest values are returned. Figure 3.5 shows how the mutual information is calculated. The method receives the number of occurrences of the query term, and the candidate term for query expansion. Then the method calculates mutual information by dividing the number of documents the candidate term occurs in by the number of documents the query term occurs in multiplied with the number of documents the candidate term occurs in. We are not using the entire collection in the results, only the number returned from the second term.



```
public float getChiSquare(int q, int all, String word){
    float nNa = q/(float)all;
    float square =getFrequency(word)* (float) Math.pow(((1-nNa)),2);
    chiSquare = square/q;


    return chiSquare;

}
```

Figure 3.6: Example from the code, method for calculating Chi square

3.3.3 Chi-square

Chi-square is the last measurement that is used to calculate alternatives for the query expansion. Like mutual information it uses the list of relevant terms generated by the TF-IDF method. For the approach, a simplified Chi-square equation was used 3.1. The reason for simplifying the equation is because in the approach, we only use the 100 first returned documents, not the entire collection.

$$\chi^2 = \frac{n_b \left(1 - \frac{n_a}{N}\right)^2}{n_a} \tag{3.1}$$

Like the two other methods the value for each term is calculated, and only the terms with a high value are returned. Figure 3.6 shows how Chi square is calculated in the approach.

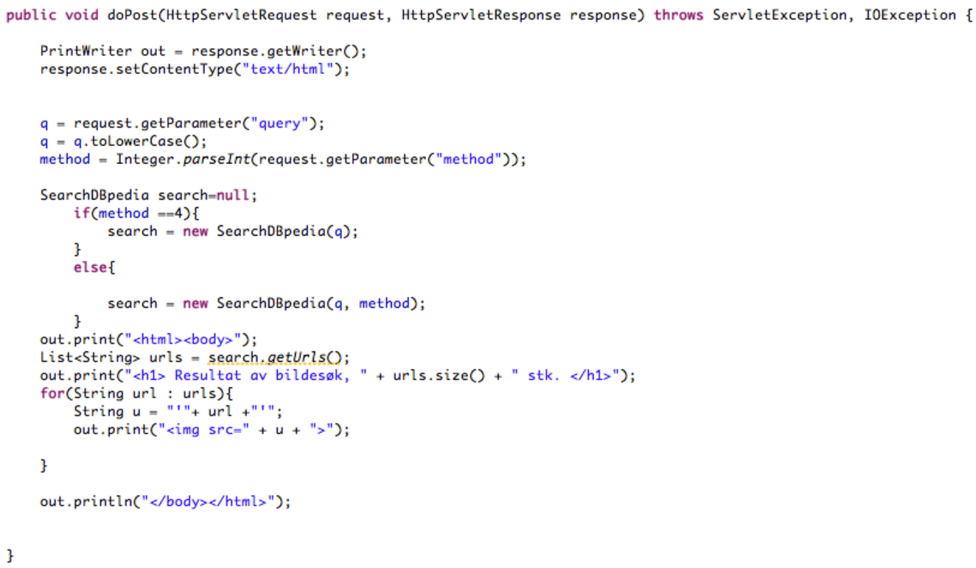
3.4 Searching Flickr

Searching the Flickr dataset is done through Solr. This was pre-processed, indexed and posted on a Solr server, and made available by our supervisor. To access the dataset a connection to the HttpSolrServer is made, and the expanded query is entered. The images the Solr server returns are put in a list.

3.5 Servlet

The communication with the user is done through the browser. An HTML site shows the different methods used to calculate alternatives for query expansion, and an input field for the query term. The HTML site makes a POST call to the servlet, and here the query term is used to create a SearchDBpedia object. The result from Solr is retrieved through the SearchDBpedia object. The URL from each image is used to view them in the browser by putting it into a <img>tag. Figure 3.7 shows how the servlet receives the information from the index.html file. This information is used to create the SearchDBpedia object. If the user is searching without query expansion, method 4 is chosen. Then the SearchDBpedia object created skips the calculations, and go directly to searching in Flickr.

We have used a Jetty server to run the approach, and an overview of the approach is described in Figure 3.8.



```
public void doPost(HttpServletRequest request, HttpServletResponse response) throws ServletException, IOException {

    PrintWriter out = response.getWriter();
    response.setContentType("text/html");


    q = request.getParameter("query");
    q = q.toLowerCase();
    method = Integer.parseInt(request.getParameter("method"));

    SearchDBpedia search=null;
        if(method ==4){
            search = new SearchDBpedia(q);
        }
        else{

            search = new SearchDBpedia(q, method);
        }
    out.print("<html><body>");
    List<String> urls = search.getUrls();
    out.print("<h1> Resultat av bildesøk, " + urls.size() + " stk. </h1>");
    for(String url : urls){
        String u = "'"+ url +"'";
        out.print("<img src=" + u + ">");

    }

    out.println("</body></html>");


}
```

Figure 3.7: Example from the code, method for the servlet

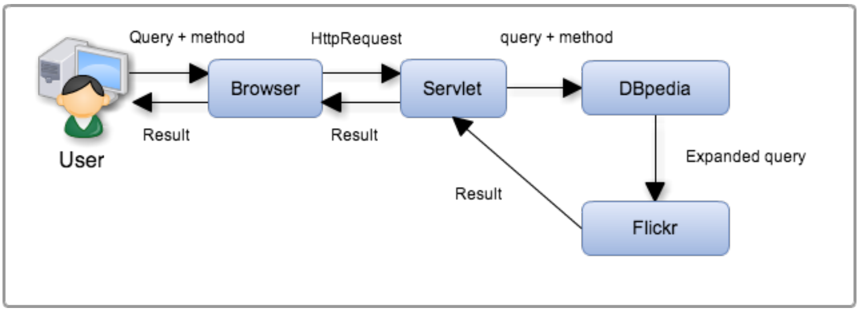


Figure 3.8: An overview of the suggested approach

Chapter 4

Evaluation

This chapter presents the test-method used for evaluating the approach. First we will describe the quantitative evaluation methodology, then we describe the evaluation we have used, and, how the evaluation was conducted.

4.1 Test methodology

Quantitative data analysis is the study of numbers. When doing a quantitative analysis the researcher are looking for patterns in the collected data, and find conclusions based on these patterns. There exist several types of quantitative data, each one of them suited for different analysis methods [27]. In the following we describe the different types of data that can be produced with quantitative testing.

4.1.1 Quantitative data

Nominal data is data that is not numerical, but questions are made numerical by adding a value to them. For example, “do you walk to work?

1. Yes, 2. No". You can not do any calculations on nominal data, only see how many answers the alternatives have. [27]

Ordinal data is data that are assigned to a quantitative scale. For instance: "How good do you like this movie (ranked from 1 → 6, where 6 is the best)?" Ordinal data are often used when someone only is interested in responses, based on number like the example before. Data like this are ranked, but there cannot be found any intervals between the different data. [27]

Interval data is similar to ordinal data, but in this case the data are measured against a quantitative scale. Two numbers are proportional against each other, like the difference between 9-10, and 2-3. [27]

Ratio data is similar to interval data, but here there is a definition of a zero. All types of calculations can be made on ratio data, since there can be stated a true zero. The test we conducted generated this type of data. There exist different approaches to present data found in quantitative analysis these are tables, bar charts, pie charts, line graphs and other graphs that present the relationship between variables. [27]

In order to draw conclusion from the data found, statistics are usually used to get a better insight in data. The main idea is to see if there really are any links between the variables. Some of these statistics are median, mean, range, standard deviation, t-test and correlation coefficients. These aids are also used to give the reader a better insight to the research.

In the evaluation part, the researchers' own meaning comes to light, they try to find out what the results imply, and if there exist similarities to other studies. [27]

4.2 Planning of the evaluation

We needed 10 persons to evaluate the approach; this number would give us a good variation of answers. The approach had to be tested with at least 20 queries, to. The queries were chosen in advance, and were a collection of different terms. Some words was chosen because they can have more than one meaning, some were wide terms, and some very narrow. In this way the system could be tested for different aspects that could occur during a search. More about the selected queries in Section 4.4.

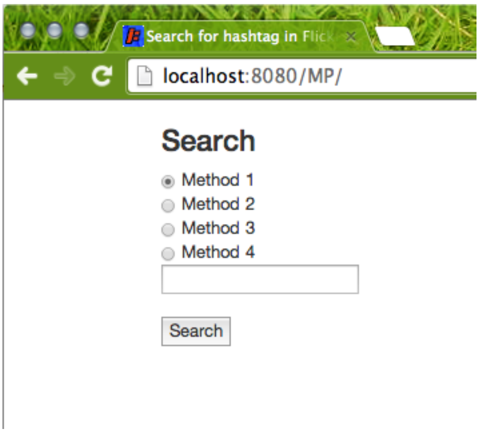
While we planned the conduction of the testing, we came across a time issue problem. It would take too much time if the testers should count the relevant, slightly relevant, and the not relevant photos, and in addition note the order they came in. A test that last over 2 hours per person would be difficult to conduct, and there is a great probability that the testers would become unfocused. This would affect the results, and would probably not be as good as the result of a shorter test. We therefore decided that the test should only focus on finding how many images that were relevant, slightly relevant and not relevant.

We used the quantitative method for testing the system, because this would give us the results that we could continue to work with. With the quantitative method, we look after patterns in the terms tested, and see if there are any methods that stands out. We decided that the setting the test persons should familiarize themself with was that they would create an (hypothetical) image collage for each search term. Before the search, they had to decide what would be relevant for them. In this way they had to have their own opinion on the query. The reason why the tester should have their own opinion, are that people don not share the same perception. In order to develop a good search system, different people opinions on what is relevant, should be considered.

4.3 Conduction of the evaluation

The test were conducted like we planned, and there were no problems during the test. The time estimate we made during the planning were good, all testers used between one and two hours to complete the test. Some of the search terms did not have any meaning to the test persons, and therefore they did not know what to expect when the result were returned. We observed all of the persons testing the system, and took notes if they commented on any specific things, or situations during the testing. The testers did not know what the spesific search methods did, or why the results were different. When they were finished with the test, we explained what they had done. Most of the testers were more impressed over the term the different methods used for query expansion, than the result from Flickr.

Figure 4.1 shows the search interface the testers used. The testers wrote the query in the text-field, and selected one of the methods. The search returned up to 50 images that were tagged with the term from the search query, and the additional term calculated from the methods. Except method 4, that searched without an additional term.

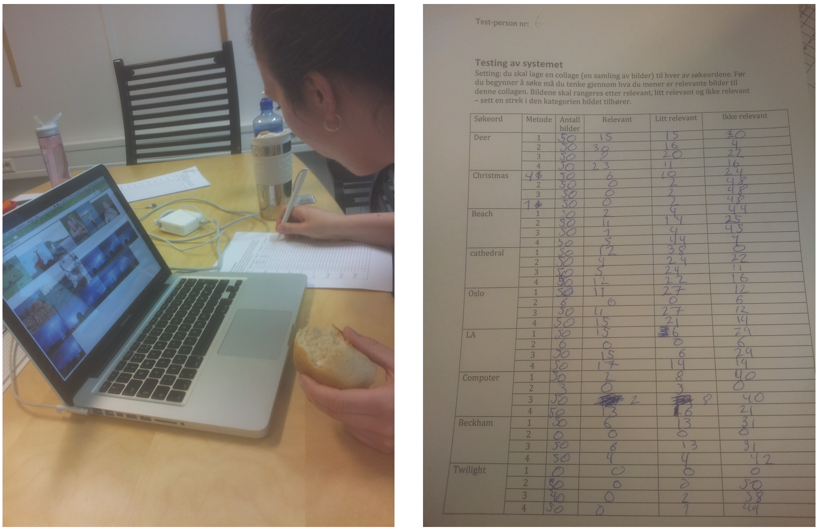


Search

● Method 1
○ Method 2
○ Method 3
○ Method 4

Search

Figure 4.1: The search interface



Test-person nr: 6

Testing av systemet

Setting: du skal lage en collage (en samling av bilder) til hver av søkeordene. Før du begynner å søke må du tenke gjennom hva du mener er relevante bilder til denne collagen. Bildene skal rangeres etter relevant, litt relevant og ikke relevant – sett en strek i den kategorien bildet tilhører.

Søkeord	Metode	Antall bilder	Relevant	Litt relevant	Ikke relevant
Deer	1	50	15	15	30
	2	50	30	16	4
	3	50	8	20	22
	4	50	23	11	16
Christmas	4 ~~1~~	50	6	10	24
	2	50	0	2	48
	3	50	0	2	48
	1 ~~4~~	50	0	2	48
Beach	1	50	2	4	44
	2	50	11	14	25
	3	50	1	4	45
	4	50	5	44	1
cathedral	1	50	12	38	0
	2	50	4	24	22
	3	50	5	24	11
	4	50	12	22	16
Oslo	1	50	11	27	12
	2	6	0	0	6
	3	50	11	27	12
	4	50	15	21	14
LA	1	50	15	6	29
	2	6	0	0	6
	3	50	15	6	29
	4	50	17	14	19
Computer	1	50	2	8	40
	2	3	0	3	0
	3	50	2	8	40
	4	50	13	16	21
Beckham	1	50	6	13	31
	2	0	0	0	0
	3	50	6	13	31
	4	50	4	4	42
Twilight	1	0	0	0	0
	2	50	0	0	50
	3	40	0	2	38
	4	50	0	1	49

(a) Testing

(b) Categorizing

Figure 4.2: Testing

4.4 Query terms

Choosing query terms were not easy. Deer was chosen because we have seen that this is a difficult tag in social media. Deer and dear have the same pronunciation, and many users tag their pictures with the wrong word. We also chose to use wide terms like christmas, Oslo, beach, cloud and tree. Names were chosen mostly to test if the first/last name was chosen for the query expansion, these are also narrow words. Abbreviations like NTNU and LA were chosen both to see if DBpedia used these, and if there existed tags with them. We also wanted to include words with multiple meanings like Twilight, bun and nail. Other words that are spesific or narrow are cathedral, computer, bonfire, field, tulip and nurse.
To summerize was the chosen words for the search:
deer, christmas, beach, cathedral, oslo, la, computer, beckham, twilight, jolie, ntnu, bonfire, field, tulip, nurse, cloud, bun, nail, tree, miley

Chapter 5

Results

This chapter presents the results from the evaluation, the discussion of the results will be presented in Chapter 6. The approach was, as mentioned in Chapter 4, tested with a quantitative method. For the evaluation, we have calculated the precision, the average precision and the standard deviation.

5.1 Results from DBpedia

We have extracted the top five alternatives for each query. The first in each list was the term that was chosen for query expansion, none of the other words were used for searching. For some of the queries, the word that is the first alternative have a much higher value than the second. For other queries there are also lists with many words with the same value. The reason for this is the method for choosing the word for query expansion, will selects the first word with the highest value. The list of words that are used to find the word for query expansion in Mutual information and Chi-square, uses a sorted list based on TF-IDF. So the first word with the highest value is also a word that had a high value when using TF-IDF. This implies that the word was relevant for TF-IDF. We have chosen to present some of the results from DBpedia, the rest can be viewed in Appendix A.

5.1.1 "Cathedral"

"Cathedral" is a narrow term, and it is unlikely that this word can be mistaken as anything else.

TF-IDF

There is not a big difference between the values in this method, in Table 5.1, but it states that the word with the highest TF-IDF score in all the documents with "cathedral" is "church".

Table 5.1: Cathedral, TF-IDF

church	0.18757975
catholic	0.18460815
st	0.17221306
diocese	0.16750138
roman	0.14288071

Mutual information

"Portsmouth" is the word that was chosen with mutual information. Top five have in this case the same value, see Table 5.2. But as mentioned is "Portsmouth" chosen based on the relevance from TF-IDF.

Chi-square

Chi-square chose "diocese" for query expansion, and has only slightly better value than "catholic". Chi-square and TF-IDF share four of the top five words in this case, see Table 5.3.

Table 5.2: Cathedral, Mutual information

portsmouth	6.29
matera	6.29
hamar	6.29
puebla	6.29
cambrai	6.29

Table 5.3: Cathedral, Chi-square

diocese	0.039999995
catholic	0.03777777
church	0.035555553
st	0.03222222
roman	0.026666664

5.1.2 ”Oslo”

”Oslo” is a wide term. Since ”Oslo” is a city a lot of different photos are tagged with this. Whether there is an image of a building, or of a statue. As a consequence the returned result can contains images of different quality.

TF-IDF

”Norway” was chosen as the most relevant term with TF-IDF for ”Oslo”. The other alternatives are displayed in Table 5.4.

Mutual information

With Mutual information ”nye” was the word that were chosen for query expansion, see Table 5.5. The situation is the same here as it was with Mutual information in the previous query, when it comes to several words with the same value.

Table 5.4: Oslo, TF-IDF

norway	0.22311872
station	0.16057621
norwegian	0.10746468
located	0.103901386
railway	0.09722726

Table 5.5: Oslo, Mutual information

nye	7.74
flights	7.74
lambertseter	7.74
cup	7.74
grønland	7.74

Chi-square

Chi-square had the same word with best value as TF-IDF, and also shared four of five top words as it did with ”cathedral”, see Table 5.6.

Table 5.6: Oslo, Chi-square

norway	0.077777766
located	0.02333333
norwegian	0.021111108
competed	0.021111108
station	0.018888885

5.1.3 ”Jolie”

The term ”Jolie” is a relatively narrow term, and is mostly known as the last name of an actor. In French, though, it also means beautiful.

TF-IDF

”Film” is the word with the highest value associated to ”Jolie” according to TF-IDF. The rest of the results can be viewed in Table 5.7.

Table 5.7: Jolie, TF-IDF

film	0.1033999
angelina	0.07925571
la	0.06433213
american	0.061430503
mantes	0.05991963

Mutual information

With Mutual information is the word with the highest score ”lazare”, see Table 5.8.

Table 5.8: Jolie, Mutual information

lazare	21.98
breaux	21.98
laide	21.98
brise	21.98
saint	10.99

Chi-square

Chi-square gave ”Angelina” the highest value, and this was used for query expansion. The rest of the results can be viewed in Table 5.9.

Table 5.9: Jolie, Chi-square

angelina	0.04666666
american	0.035555553
film	0.03333333
directed	0.024444442
la	0.024444442

5.1.4 "Deer"

"Deer" is a narrow word, and refers to the animal deer. One problem with "deer", is that it is often mistaken with the word "dear", and this can lead to less relevant results.

TF-IDF

As Table 5.10 shows, was the word with the best TF-IDF score for "deer", "red".

Table 5.10: Deer, TF-IDF

red	0.18458982
species	0 .111514665
subspecies	0.0954121
alberta	0.08928807
county	0.08320574

Mutual information

With Mutual information was "axis" the word with the highest value. See Table 5.11 for the rest of the results.

Table 5.11: Deer, Mutual information

axis	9.15
school	2.2875
creek	1.3071429
alberta	0.5382353
subspecies	0.5083333

Chi-square

"Red" was also the word with the highest value for Chi-square, see Table 5.12.

Table 5.12: Deer, Chi-square

red	0.03222222
species	0.021111108
located	0.021111108
north	0.015555553
central	0.012222221

5.1.5 "Tulip"

The word "tulip" is a special kind of flower, and therefore a narrow term.

TF-IDF

"Snails", see Table 5.13, was the term with the highest value calculated by TF-IDF.

Mutual information

With Mutual information, "Virus" was the word with the highest value, and as we see in Table 5.14, all the words in the list have the same value.

Table 5.13: Tulip, TF-IDF

snails	0.29719922
gastropod	0.26865178
sea	0.26865178
fasciolariidae	0.26865178
marine	0.26865178

Table 5.14: Tulip, Mutual information

virus	5.36
breaking	5.36
tulips	5.36
cygnus	5.36
nebula	5.36

Chi-square

”Family” is the term returned by Chi-square, and was used for query expansion together with ”tulip”. The resuls can be viewed in Table 5.15.

Table 5.15: Tulip, Chi-square

family	0.085555546
species	0.08444443
sea	0.082222216
fasciolariidae	0.082222216
marine	0.082222216

5.1.6 "Bun"

"Bun" is a wide word, and do have different meanings. People might associate the word with different things, for example a cinnamon bun or a hair bun.

TF-IDF

"Kong" had the highest value in TF-IDF, see Table 5.16, and was returned for the term "bun".

Table 5.16: Bun, TF-IDF

kong	0.067697495
hong	0.06576405
sweet	0.05490708
buns	0.045686215
album	0.045686215

Mutual information

For Mutual information, "penny" was returned. As we see from Table 5.17 there were several with the same value, but TF-IDF ranked "penny" highest.

Table 5.17: Bun, Mutual information

penny	15.12
franks	15.12
bars	15.12
schnecken	15.12
chop	15.12

Chi-square

"Sweet" was returned with the highest value, the results can be viewed in 5.18.

Table 5.18: Bun, Chi-square

sweet	0.019999998
kong	0.015555553
hong	0.014444443
type	0.012222221
dough	0.009999999

5.1.7 "Miley"

The search term "Miley" is a narrow term, because it is a last name for the singer "Miley Cyrus". They who have heard about her, will expect to see images of her.

TF-IDF

For the term "Miley" the word chosen for query expansion was "Cyrus", see Table 5.19.

Table 5.19: Miley, TF-IDF

cyrus	0.17484209
hannah	0.15812285
montana	0.1552738
song	0.14404838
album	0.10339598

Mutual information

"Deepdale" was returned for the term "Miley". Two terms had the same value, but "Deepdale" was favored because it was higher ranked in TF-IDF, see Table 5.20.

Table 5.20: Miley, Mutual information

deepdale	20.29
dyer	20.29
ellington%27s	10.145
methodist	10.145
jesse	6.7633333

Chi-square

"Cyrus" was returned with the highest value in the Chi-square method, see Table 5.21, this word is the same as in TF-IDF method.

Table 5.21: Miley, Chi-square

cyrus	0.05666666
american	0.04333333
hannah	0.03777777
series	0.03777777
montana	0.03666666

5.1.8 "Beckham"

"Beckham" is a narrow word, and a search one the word the expected result will be with the soccer player "David Beckham".

TF-IDF For the method TF-IDF, the term with the highest value returned was "David". The results can be viewed in Table 5.22.

Table 5.22: Beckham, TF-IDF

david	0.074879766
victoria	0.063213564
released	0.061895546
county	0.05772426
united	0.054938573

Mutual information

"McCreary" was returned in Mutual information, it was higher ranked in TF-IDF than the other two terms with the same value, see Table 5.23.

Table 5.23: Beckham, Mutual information

mccreary	23.86
molloy	23.86
oregon	23.86
register	11.93
entertainment	11.93

Chi-square

Like in TF-IDF, the word with the highest value was "David", see Table 5.24.

Table 5.24: Beckham, Chi-square

david	0.04333333
victoria	0.03222222
united	0.028888887
born	0.027777774
released	0.022222219

5.1.9 "Beach"

The term "beach" is a wide word. Users may expect different results when they search on this.

TF-IDF The word with the highest value was "located", the results can be viewed in Table 5.25.

Table 5.25: Beach, TF-IDF

located	0.11350435
long	0.10312733
south	0.09635876
north	0.09261813
saskatchewan	0.08503797

Mutual information

"Palm" was returned in Mutual information, and was higher ranked in TF-IDF than the other words with the same value. See Table 5.26.

Chi-square

In the method Chi-square, "located" was the word with the highest value, see Table 5.27.

We can see, according to these results, that Chi-square and TF-IDF are

Table 5.26: Beach, Mutual information

palm	7.12
barbara	7.12
santa	7.12
school	7.12
huntington	7.12

Table 5.27: Beach, Chi-square

located	0.019999998
south	0.016666666
saskatchewan	0.016666666
north	0.013333332
long	0.009999999

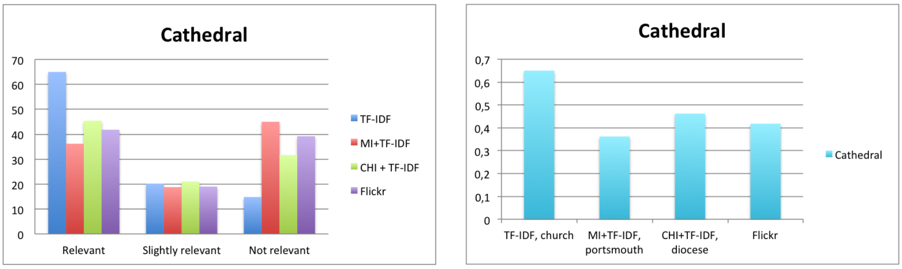
more related than Mutual information.

5.2 Results from Flickr

We have chosen to show the results from Flickr as bar charts, and with the precision for each method. The limit for each search was 50 photos, but not every search returned this amount. We will discuss the reason for this in Chapter 6. As we can see, some of the methods were more successful than others. The rest of the results are listed in the Appendix C. The results presented in this section are the terms with the highest value from the previous section. These terms are found using the methods described in Chapter 3.

5.2.1 TF-IDF

The query that returned the highest number of relevant results when using TF-IDF for choosing the term for query expansion was ”cathedral”. The term chosen for query expansion was ”church”. 65 % of the returned images were relevant to the test persons, and only 15% of the returned images are categorized as not relevant. The result is presented in Figure 5.1.



(a) Cathedral, percent of relevance

(b) Cathedral, precision

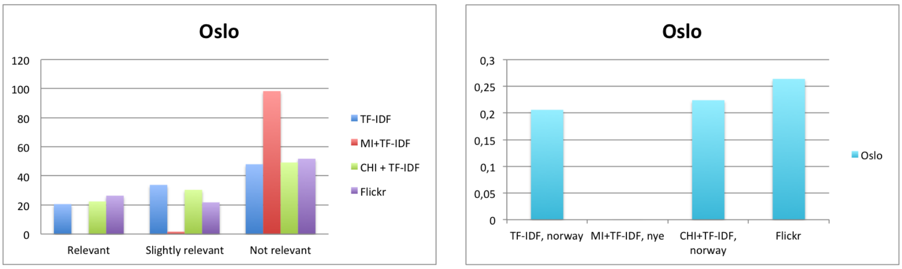
Figure 5.1: Cathedral, Distribution of relevant pictures, and precision.

The query term that achieved the best ”slightly relevant” result with TF-IDF was ”Oslo” with 33.8%. The result on ”slightly relevant” was actually better than the ”relevant” (with 20 %), and is presented in Figure 5.2. The testers categorized most images as ”slightly relevant”. This indicates that the testers are not satisfied with them, but that they have some relevance to the query.

The query with the most ”not relevant” pictures for TF-IDF was ”Jolie” with 92.7%. The result is presented in Figure 5.3.

5.2.2 Mutual information

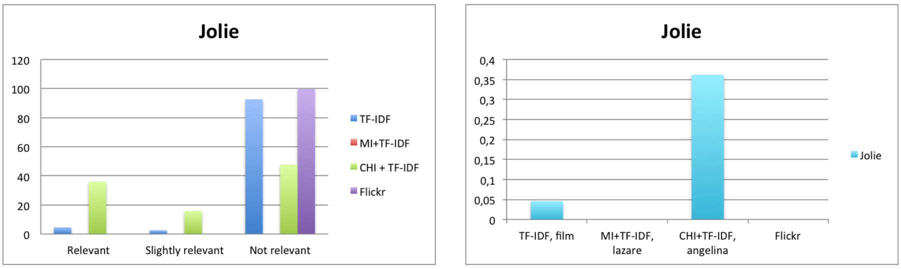
The query with the most relevant images when searching with Mutual information was ”deer”, Figure 5.4. 77% of the images returned was categorized



(a) Oslo, percent of relevance

(b) Oslo, precision

Figure 5.2: Oslo, Distribution of relevant pictures, and precision.



(a) Jolie, percent of relevance

(b) Jolie, precision

Figure 5.3: Jolie, Distribution of relevant pictures, and precision.

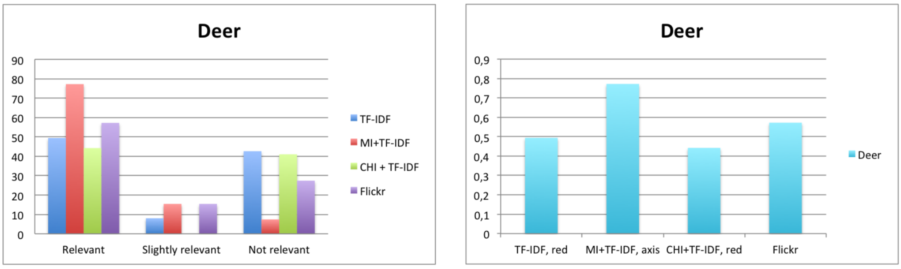
as relevant, and only 7.4 % as not relevant.

"Tulip", Figure 5.5, is the query with the best "slightly relevant" result for Mutual information, with 35%.

"Bun" was the term with the worst result with Mutual information, none of the returned images was considered relevant. The result is presented in Figure 5.6.

5.2.3 Chi-square

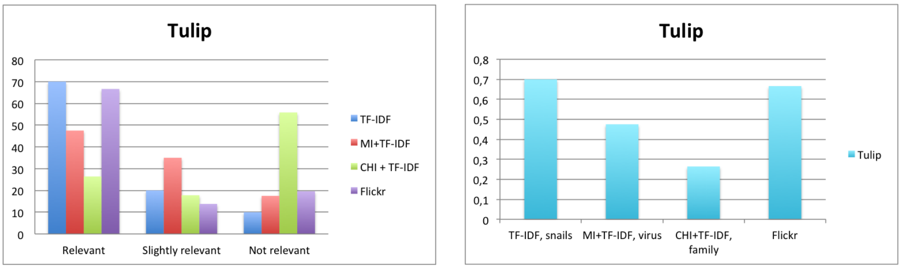
For the method Chi-square, the term "Miley" returned one of the best results, Figure 5.7. There were in total 50 returned images per query, and



(a) Deer, percent of relevance

(b) Deer, precision

Figure 5.4: Deer, Distribution of relevant pictures, and precision.



(a) Tulip, percent of relevance

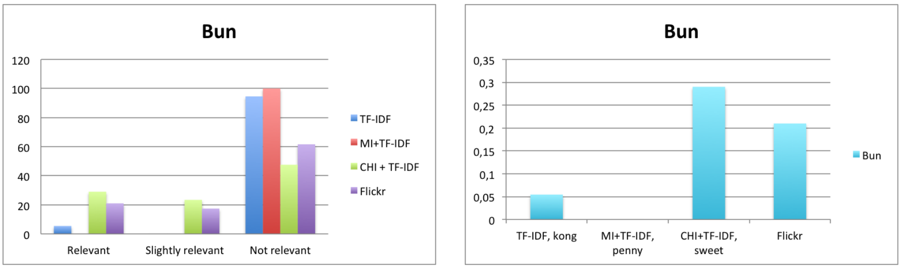
(b) Tulip, precision

Figure 5.5: Tulip, Distribution of relevant pictures and precision.

of these the testers found 46,8% relevant. Only 23,6% was not relevant, whereas 29,6% was slightly relevant.

The term that returned one of the best results in slightly relevant with Chi-square was ”Beckham”. Of 50 returned pictures, 20,4% were slightly relevant. This was lower than the relevant, with 24,6%. The result is presented in Figure 5.8.

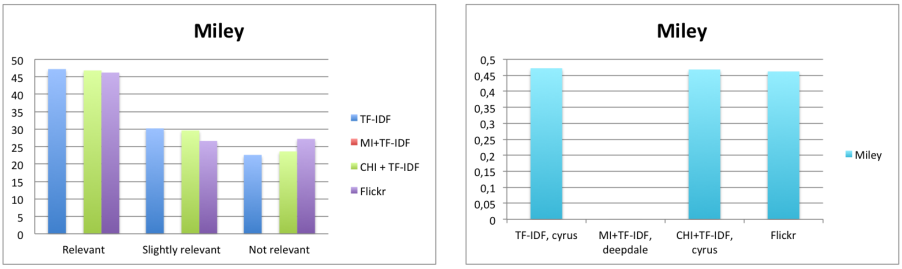
The term that received the poorest result on not relevant documents was ”beach” when using Chi-square. 85.6% of the pictures was not relevant, and only 5.2% were relevant, see Figure 5.9 for the results.



(a) Bun, percent of relevance

(b) Bun, precision

Figure 5.6: Bun, Distribution of relevant pictures and precision.



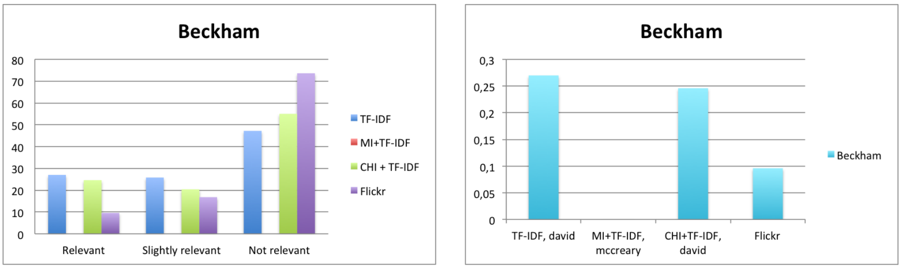
(a) Miley, percent of relevance

(b) Miley, precision

Figure 5.7: Miley, Distribution of relevant pictures and precision.

5.2.4 Average precision

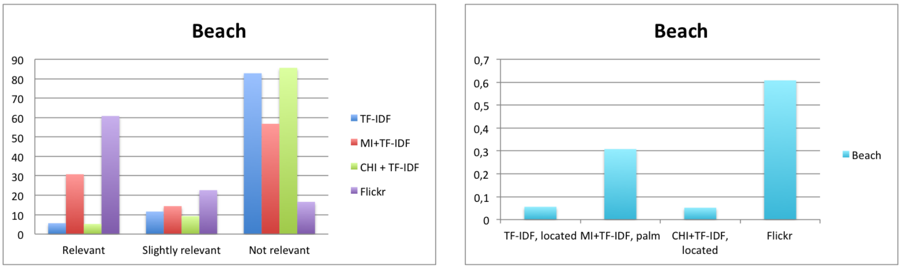
For the evaluation, we calculated the precision for each method on each term. To get a better understanding on how these methods worked overall on all terms, we found the average precision for each method. In Figure 5.10, we can see that the method that had the poorest precision was Mutual information, with an average score of 0.22727777. TF-IDF is next with an average precision of 0.24728. Only a precision of 0.00045 distinguish Chi-square from a search without query expansion. Chi-squares' average precision was 0.40195, and the search without query expansion had a precision of 0.4015.



(a) Beckham, percent of relevance

(b) Beckham, precision

Figure 5.8: Beckham, Distribution of relevant pictures and precision.



(a) Beach, percent of relevance

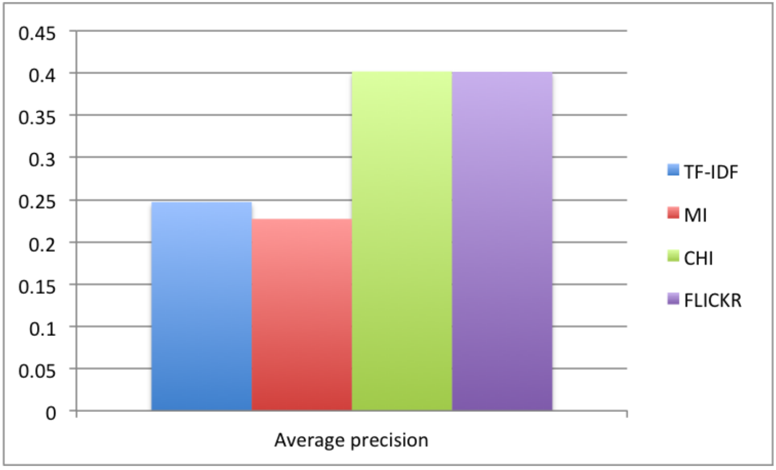
(b) Beach, precision

Figure 5.9: Beach, Distribution of relevant pictures and precision.

5.3 Standard deviation

To evaluate how well the methods performed during the evaluation, standard deviation was calculated for each term. Below are the results from the terms that got the best and poorest precision, are presented. The result from the other terms can be found in Appendix D. If the tester agreed on the relevance of a term, in this case all are either relevant, slightly relevant or not relevant, the standard deviation should approach zero. A higher standard deviation tells that the term is difficult to place in one specific relevance section.

Figure 5.10: Average precision, all methods



5.3.1 TF-IDF

Cathedral	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Church	2.502	0.545	0.738	Relevant

Table 5.28: Standard deviation, ”cathedral”

In Table 5.28 the results from standard deviation for the term ”cathedral” can be found. ”Cathedral” was the term that got the highest score on precision with TF-IDF, the variation in the answer was lagre, see Table 5.28. Neither relevant, slightly relevant or not relevant got the majority of the answers, as we can see from the result of the standard deviation.

Jolie	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Film	1.11923	0.19	0.444	Not relevant

Table 5.29: Standard deviation, ”Jolie”

For the term ”Jolie”, which had the poorest precision with TF-IDF, the varians of the answers were not so various. Most of the answers were not relevant, so the standard deviation for "Jolie" was closer to zero, which is expected, since the testers agreed on the relevance, see Table 5.29.

5.3.2 Mutual information

Deer	Second term	AVG	Varians	SD	Highest relevance
MI	Axis	2.698	0.358	0.598	Relevant

Table 5.30: Standard deviation, "deer"

"Deer" was the term that got one of the best precision with mutual information, the varians for the answer were low, see Table 5.30. As a result of this, the standard derivation for the term was low and the testers agreed on that most of the images were relevant.

Bun	Second term	AVG	Varians	SD	Highest relevance
MI	Penny	1.00	0.00	0.00	Not relevant

Table 5.31: Standard deviation, "bun"

For the term "Bun" which got the poorest precision with mutual information, all agreed on that the returned result was not relevant. Therefore the values of seen in Table 5.31, are all zero.

5.3.3 Chi-square

Miley	Second term	AVG	Varians	SD	Highest relevance
CHI	Cyrus	2.232	0.65	0.806	Relevant

Table 5.32: Standard deviation, "Miley"

The term "Miley" got the best precision for Chi-square, the varians in answers were quite lagre, see Table 5.32. The returned result was either good or bad, something that can be seen on the standard deviation that was 0.8063.

Beach	Second term	AVG	Varians	SD	Highest relevance
CHI	Located	1.196	0.261	0.51	Not relevant

Table 5.33: Standard deviation, "beach"

"Beach" was the term, that got the poorest precision with Chi-square. The variation in answers were not large, see Table 5.33, and the majority found the returned result as not relevant.

Chapter 6

Discussion

In this chapter, we evaluate the methods we have studied in this work, and discuss the results from Chapter 5.

6.1 Discussion of the methods

To try to find out which method that gives the best results, and why it does so, will we take a deeper look at the results from the different methods.

6.1.1 Review of TF-IDF

The first method we will evaluate is TF-IDF. As presented in Chapter 5 TF-IDF was the method that performed second best according to average precision.

6.1.1.1 Pictures jugded as relevant

"Cathedral" had the best result with TF-IDF. "Church" was chosen for query expansion. This term is strongly connected to the query since a cathedral is a church. The only thing that differs a church from a cathedral is that it holds the seat of a bishop [1]. The term "cathedral" is narrow, but not as narrow that pictures in Flickr are not tagged with it. The term "cathedral" does not have several meanings, so there is no ambiguity that can affect the retrieval result. The method returns both churches and cathedrals, and users find both of them relevant. This is not necessarily negative, if the test person finds the image relevant – it is indeed relevant. If we were to find out if the image really contains a cathedral the testing had to be done in a different way. However, this is beyond the scope of this thesis.

6.1.1.2 Pictures jugded as slightly relevant

"Oslo" gave the best "slightly relevant" result. The term chosen for query expansion was "norway". The result is presented in Figure 5.2. "Oslo" is a difficult term when it comes to categorizing relevant photos because it is a very wide term. With the term "Norway" as an addition, the query becomes even wider. The desirable situation had been query expansion with a term that narrows the query. Oslo is a big city, and even if the photo is of something in Oslo, the test persons were most likely looking for images that were of landmarks and tourist attractions.

6.1.1.3 Pictures jugded as not relevant

The query with the highest "not relevant" percent was "Jolie". The term chosen for query expansion was "film". The result is presented in Figure 5.3. These terms are actually relevant to each other, the actor Angelina

[1] http://www.oxforddictionaries.com/definition/english/cathedral?q=cathedral

Jolie plays in films. However, it seems that the relevant images on Flickr are not tagged with this term.
"Film" is a word with many synonyms and meanings. One description of film is *"a thin flexible strip of plastic or other material coated with light-sensitive emulsion for exposure in a camera, used to produce photographs or motion pictures"*. Another is *"a story or event recorded by a camera as a set of moving images and shown in a cinema or on television"*. This description can have synonyms like cinema, motion picture and movie [2].

6.1.2 Review of Mutual information

The overall performance of this method was not satisfactory. 50% of the search with mutual information did not return any images. This must either be because the terms selected for query expansion not really is relevant to the query, or because pictures in Flickr were not tagged with the terms at all. It seems that no images with the tags we are searching for exist in our dataset from Flickr.

In Chapter 5 we saw that many of the words that were using Mutual information had the same value. Although they were sorted after TF-IDF, we could not avoid the fact that the "wrong" word may have been chosen for query expansion. Since we do not execute any analysis of the alternative words for query expansion, we have no control over what may be the most correct term to use.

[2] http://www.oxforddictionaries.com/definition/english/film?q=film

6.1.2.1 Pictures jugded as relevant

"Deer" gave the best result when searching with Mutual information. The reason why we chose this term as one of the queries was because many users write "deer" instead of "dear", and we wanted to test if we could make the system understand that we wanted pictures with deer's, and not someone's dear. The term selected for query expansion was "axis". Axis deer is a specific type of deer, and makes the query narrower. See Figure 5.4, in Chapter 5, for the results from searching with the term "deer".

6.1.2.2 Pictures jugded as slightly relevant

The query with the best "slightly relevant" result with Mutual information was "tulip". The term added with query expansion was "virus". This method only returned 4 pictures. Figure 5.5 displays the results from "tulip". This was one of the queries that had several words with the same value. Although there were few returned images, most of the pictures were relevant, or slightly relevant, to the testers. Therefore we wondered what was most important; to get a result with many images, where most of them are slightly relevant, or a result that returned few images, but all of these were relevant. We believe that it is more important to retrieve relevant images, of high quality, even if the returned result contained few images. The precision is higher when the result returns few images, and all of these are relevant.

6.1.2.3 Pictures jugded as not relevant

"Bun" only returned 6 photos with mutual information, and none of them were relevant. "Bun" alone is a difficult word for Norwegian test persons, and almost everyone expected different pictures. The results from the query "bun" can be viewed in Figure 5.6. The word chosen for query expansion was "penny". A Google search on "bun" and "penny" leads to eBay [3], where these two words together are associated with Queen Victoria bun head pennies. We observed the test persons comment *"hm, I don't know what bun is, but it's almost bunny so I'll just choose photos with bunnies"*, or *"when I think of bun I associate it with some food we were served in China"*. Another test person expressed that she expected hair buns. "Bun" also had several alternative words for query expansion with the same value.

6.1.3 Review of Chi-square

Chi-square was the method with the best average precision.

6.1.3.1 Pictures jugded as relevant

"Miley" returned one of the best results with Chi-square. In Figure 5.7, the results for the term "Miley" can be viewed. There might be several reasons why this term gave good results. For the query "Miley" was "Cyrus" calculated with the highest Chi-square, and used for query expansion. "Cyrus" is the last name of the singer and actor Miley Cyrus, and images tagged with these will give a good result. Another aspect of why the results were good are that both "Miley" and "Cyrus" are narrow terms, and they do not have synonyms or ambiguity related to them. The combination of these two will probably always give results that are related to the singer and actor. A third reasons for the good result is that Miley Cyrus is a well-known person, and many people have heard about her. Then it is easy to make an

[3] http://www.ebay.co.uk/bhp/bun-penny

opinion if the images retrieved are of her or related to her.

"NTNU" also give one of the best results in Chi-square. The word chosen for query expansion for "NTNU" was "Norwegian", in Figure C.5 the results for "NTNU" can be viewed. "NTNU" is a Norwegian abbreviation for Norwegian University of Science and Technology, and "Norwegian" is a good description for "NTNU" despite that the search is in English. The abbreviation could have another meaning in English. Like the case with "tulip" did "NTNU" return few images, but these were relevant. The same aspect as discussed with "tulip", play a part with the term "NTNU".

6.1.3.2 Pictures jugded as slightly relevant

"Beckham" had the best "slightly relevant" result with Chi-square. The results can be viewed in Figure 5.8. The term given for query expansion was "David", which is the forename of the soccer player David Beckham. When including "David", the query becomes very specific, and we believe this affects the result.

6.1.3.3 Pictures jugded as not relevant

The worst query for Chi-square was "beach". There might be several reasons why the result of this query is poor. The first reason might be that the term for query expansion was "located", which is not a good description of a beach. Even though "located" can be a part of a beach description, for example "#Beach #located #losangeles", this is probably not the tag used on most beach pictures. A second reason why "beach" got a poor result can be that beach is a wide term, it might not only display pictures of a sandy beach, but also activities at the beach, or buildings close to the beach. The testers might not find pictures like that relevant, and therefore mark them as not relevant.

6.1.4 Summary

As we have seen in the previous section, narrow terms gives better precision when searching with query expansion. In the cases where a search in Flickr has outperformed the search with query expansion, it is clearly because of the word chosen for query expansion.

TF-IDF finds the words that are frequently mentioned in the documents that also contain the query word. These words are likely to be related to the query, but there is no guarantee that this is the case. TF-IDF gives good results for some of the queries, but not every time.
In this approach we have altered the equation for Mutual information to use the number of documents that contains B (the term chosen for query expansion), that is returned by the search in DBpedia. This means that all these documents also contain A (the query term). Originally the number should be the number of documents that contain B, from the whole index. This might be a reason why some of the words returned with Mutual information are a less descriptive than for others. Another reason can be that, as mentioned in Section 2.2.5.1, Mutual information tend to favor low frequency words. We also experienced this when we tested the approach with the ranged word list returned from TF-IDF; the words that got the highest value with Mutual information were lower on the TF-IDF list.

From Figure 5.10 we saw that query expansion with Chi-square was slightly better than a search without. Chi-square especially stands out when it comes to searching after names. Unlike Mutual information, Chi-square calculated a higher value for the words high up on the TF-IDF list. Therefore, the relevant pictures from TF-IDF and Chi-square, are also (almost) the same. They only differ in 9/20 queries. Yet, many of the results from these methods differ in number of relevant images. We believe that the testers became more strict during the testing, and therefore it could be interesting to compare the results based on the number of not relevant images. When the testers categorized images as slightly relevant, they expected that they could have gotten a better result for other images. If they had no other images to choose, the images with slightly relevant judgement would have been chosen. Still, it would be best to give the user the most relevant result.

As explained in Chapter 4, our evaluation focused on finding the number of relevant pictures, not in which order they came. As a result, we could not calculate the MAP for the terms. If we had performed an evaluation with the order of relevant images, and calculated the MAP the end result would have been different. With the end result, we do not mean the returned images, but the values calculated. The testers, who performed the evaluation, would still have found the same numbers of relevant images as they did in the evaluation, and the average for each method would have been the same since there are no changes in the amount of returned images. We could have evaluated the average precision at 50, which are maximum number of images returned for the interface. For our approach it was important to find out if the result became more relevant with query expansion, and this we accomplished with the evaluation method we used, regardless of the order the relevant images occurred in.

6.2 Discussion of the datasets

The results are strongly connected to the datasets because we perform the search in them. In this section will we discuss how the datasets may have impacted the results in this approach.

6.2.1 Flickr

The dataset from Flickr were last modified in 2012, and thus it is not entirely up to date. The result of this is that some of the returned images have been removed from Flickr, and could not be shown. This led to more "not relevant" photos than it could have been if the dataset were more recent. An example of this can be viewed in Figure 6.1.
Due to this, the result from the testing might have been different. Hypothetically these images could have been relevant, and affected the result in a positive way.

6.2.1.1 Tagging

As mentioned in Section 1.1, the use of social media have increased the latest years, and the users are now more familiar with using hash-tags. This can imply that images on Flickr are tagged with more relevant tags now, than in 2012. We did a directly search at Flickr.com with some of the expanded queries, and the result seemed more relevant. We believe that if we updated the dataset from Flickr, the results would be more relevant to the user. This can apply to both a larger amount of pictures that are

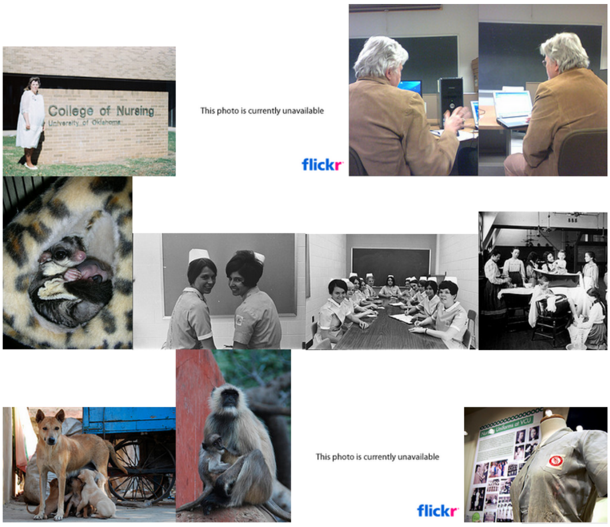
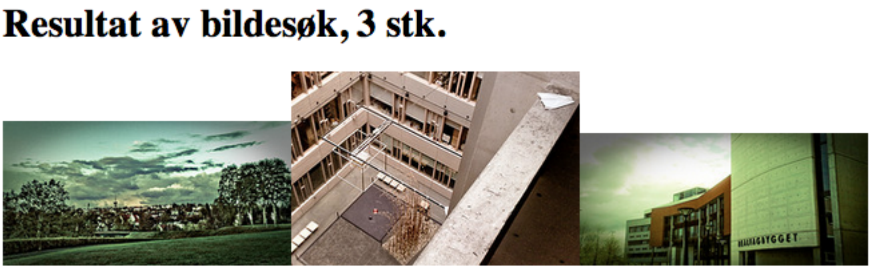

College of Nursing
University of Oklahoma
This photo is currently unavailable
flickr
This photo is currently unavailable
flickr

Figure 6.1: Photo from Flickr unavailable, search on nurse, method 1

returned, and that there exist pictures that are tagged with the expanded query we are searching with. An example of this is presented in Figure 6.2, where we can see the result from searching on "ntnu" and "norwegian" in our approach. Figure 6.3 shows the result when searching on the same tags in Flickr (search performed 30.05.14).



Resultat av bildesøk, 3 stk.

Figure 6.2: Result from searching on NTNU and norwegian in our approach

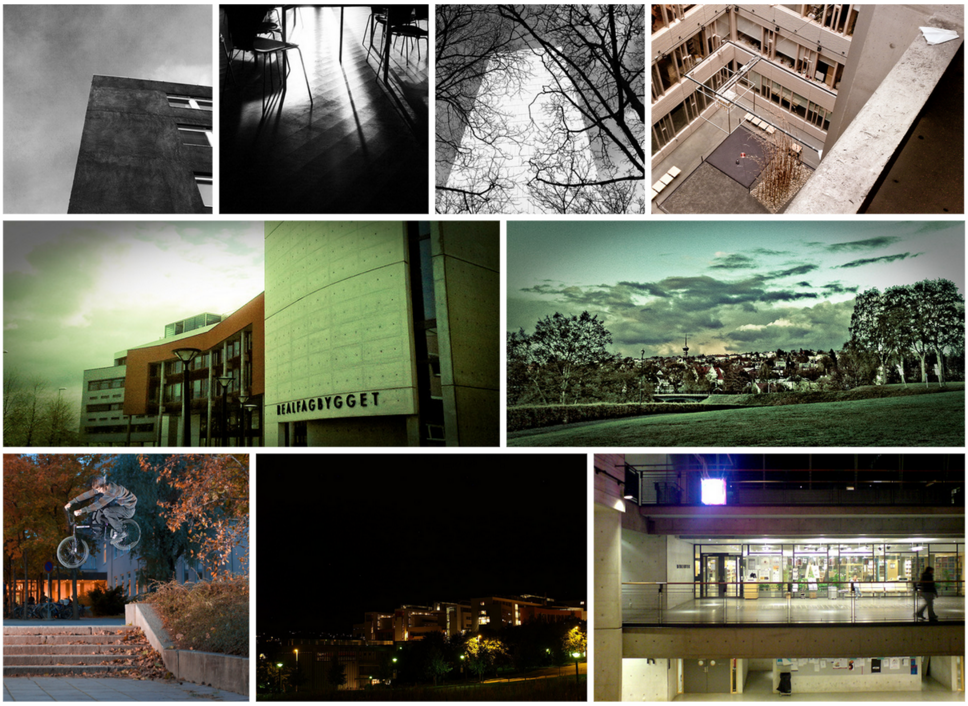


Figure 6.3: Result from searching on NTNU and norwegian in Flickr

6.2.2 DBpedia

The dataset from DBpedia is from 2014, and this can be a mismatch with the dataset from Flickr. The information on DBpedia, and Wikipedia, are in constant change since users edit the information in each article. This implies that new trends quickly are written about. An example is the relatively new word "twerking". When searching on "twerking" in DBpedia, the word "dance" is chosen for all the methods, but none of the methods returned any photos. In fact, no photos in the dataset were tagged with either "twerking", or "twerking" and "dance". However, if we do a search on "twerking" and "dance" directly in Flickr many photos are returned [4]. If some of the queries we chose to test with were to "new", this could have an impact on the result in terms of less relevant pictures.

6.3 Research question revisited

In the beginning of this thesis we asked the research question *Is it possible to improve search results in social media, such as Flickr, by using additional metadata from a structured database, like DBpedia?* The answer to this question is yes, but only by a slight margin. As described in Section 6.1.4, it is not possible to make any concrete conclusions based on this, and we suggest testing the approach with more queries to draw the final conclusions. In RQ1 we asked *Does a system like this already exists?*, and in RQ2 we wondered *Can this be done with query expansion, without feedback from the user?* To our knowledge there are no other systems that use query expansion in the same way as we have suggested. For our approach query

[4] https://www.flickr.com/search/?q=twerking+dance&m=tags&ct=&mt=photos&adv=1

expansion can give the user more relevant results. For the last question asked, RQ3, *What method gives the most relevant result?*, Chi-square chose the best terms for query expansion.

6.4 Limitations

No approach is perfect, and every system has its limitations. In our case, one limitation is that many images are removed from the Flickr dataset, as described in Section 6.2.1. The approach is also limited by the fact that images in Flickr today are better tagged than the ones in our dataset (Section 6.2.1.1). This could have been avoided if we had chosen to use a newer dataset, though this dataset had to be preprocessed like the one we got from our supervisor.

Another limitation can be that any kind of words can be chosen for query expansion. However, analyzing the words are not a part of this thesis. Some words without any meaning could have removed by the stopword list, but it is difficult to know which words that really are associated to the different queries, and which word can be removed.

We only evaluated the approach with testers that were between 22-26 years, and are studying IT. This is a poor representation of the population, and implies that the approach might have no external validity. The results might have been different for some of the queries. An example is the query "Twilight", older users might associate this with the time of day that the sun is below the horizon, and not the books and movie that younger users think about.

Chapter 7

Conclusion

7.1 Summary

The extreme amount of badly tagged photos on social media is challenging for the information retrieval process. The motivation behind this thesis has been to give the users results that are more relevant. Our hypothesis was that adding related words to queries can give the users more relevant results. We have suggested an approach that uses DBpedia for additional metadata, and Flickr as the source for the images. The dataset from DBpedia has been preprocessed and indexed, and then used for searching. We used the three methods: TF-IDF, Mutual information and Chi-square for calculating alternative words for query expansion. With TF-IDF, the chosen word for query expansion was based on the term frequency in the returned documents, normalized by the document frequency. With Mutual information, the word was chosen based on the dependence to the search term. The higher the value, the more relevant, or dependent, the words are to each other. This method, however, tend to favor low frequency words. Chi-square was the final method, and compares the expected co-occurrence of the query term and the second term, with the actual number of co-occurrences of these terms. The resulting images were showed in the browser. The approach has been evaluated by 10 test persons, were they decided which method that gave them the most relevant result. Our evaluation has shown that Chi-square has been the method with the best average precision, but was only been slightly better than a search without query expansion. There are no indications on that the use of photo-sharing applications will decreasing in the years to come. The amount of images on the Internet will continue to grow, and to accomplish this problem it is important to develop a good search function. Our evaluation have shown that the approach we have proposed in this thesis, has great potentials be a good beginning.

7.2 Future work

In this work, we did not analyze the terms chosen for query expansion. This implies that any other word associated with the query could be chosen. It might be more useful to choose a word that either describes the query, or makes the query narrower. For example, if the query were "tree", suitable alternatives for query expansion would be "green", "tall" or "foliage". To be able to do this, the system most likely has to understand the context of the query, and this is already a challenge for information retrieval.

There could also be an alternative to use different datasets for this approach. This could be metadata sources like dictionaries and datasets from other photo-sharing communities. As we discussed in Chapter 6 other methods should also be evaluated, to see if there are any other methods that returns a better result.

Appendices

Appendix A

Query expansion alternatives

Here are the results from TF-IDF, mutual information and Chi-square.

Table A.1: Christmas, TF-IDF

album	0.26640436
released	0.16503504
songs	0.14497727
music	0.09267939
song	0.09267939

Table A.2: Christmas, Mutual information

metal	6.72
tree	3.36
chipmunks	2.24
list	1.344
spirit	1.12

Table A.3: Christmas, Chi-square

album	0.06888888
released	0.041111108
songs	0.024444442
music	0.016666666
song	0.013333332

Table A.4: LA, TF-IDF

film	0.24179554
argentine	0.23359197
airport	0.17094004
municipality	0.16001536
province	0.1416057

Table A.5: LA, Mutual information

colle	3.01
florida	3.01
?	3.01
soto	3.01
düsseldorf	3.01

Table A.6: LA, Chi-square

film	0.038888883
argentine	0.035555553
municipality	0.018888885
province	0.015555553
airport	0.0111111095

Table A.7: Computer, TF-IDF

software	0.08566075
game	0.07401343
system	0.07401343
scientist	0.07401343
society	0.06594622

Table A.8: Computer, Mutual information

fraud	7.51
chemical	7.51
optimization	7.51
help	7.51
online	7.51

Table A.9: Computer, Chi-square

software	0.016666666
game	0.015555553
scientist	0.014444443
science	0.012222221
list	0.0111111095

Table A.10: Twilight, TF-IDF

anthology	0.36020076
series	0.26419133
zone	0.26135057
episode	0.25566906
television	0.25282827

Table A.11: Twilight, Mutual information

theater	2.44
shoreliner	2.44
train	2.44
fans	2.44
steve	2.44

Table A.12: Twilight, Chi-square

series	0.10222221
zone	0.098888874
television	0.098888874
episode	0.09666665
american	0.095555544

Table A.13: NTNU, TF-IDF

university	0.15280305
norwegian	0.14333647
science	0.095688656
technology	0.09371771
railroad	0.08495972

Table A.14: NTNU, Mutual information

railroad	41.8
corporation	41.8
norfolk	41.8
transportation	41.8
library	41.8

Table A.15: NTNU, Chi-square

norwegian	0.27664956
university	0.2717094
science	0.22230768
technology	0.21242735
trondheim	0.1877265

Table A.16: Bonfire, TF-IDF

album	0.10988759
released	0.09281023
band	0.085774966
rock	0.07688904
live	0.06732922

Table A.17: Bonfire, Mutual information

corpusty	22.5
toffee	22.5
trees	22.5
hobbits	22.5
seurasaari	22.5

Table A.18: Bonfire, Chi-square

released	0.03777777
band	0.03777777
album	0.03222222
rock	0.03222222
hard	0.021111108

Table A.19: Cloud, TF-IDF

computing	0.10341233
clouds	0.06803896
software	0.061869614
onapp	0.05536385
service	0.05536385

Table A.20: Cloud, Mutual information

onapp	10.23
funnel	10.23
pileus	10.23
printing	10.23
cdn	10.23

Table A.21: Cloud, Chi-square

computing	0.021111108
based	0.013333332
software	0.0111111095
service	0.0111111095
clouds	0.009999999

Table A.22: Nail, TF-IDF

nails	0.09057493
released	0.08206253
tooth	0.08011355
album	0.07592929
plate	0.073706284

Table A.23: Nail, Mutual information

salons	13.53
warts	13.53
technicians	13.53
ungual	13.53
water	13.53

Table A.24: Nail, Chi-square

tooth	0.028888887
released	0.026666664
album	0.025555553
records	0.02333333
plate	0.019999998

Table A.25: Tree, TF-IDF

family	0.14394906
frog	0.09156793
native	0.07137148
species	0.07137148
binary	0.05630977

Table A.26: Tree, Mutual information

suslin	8.46
automata	8.46
random	8.46
peony	8.46
deterministic	8.46

Table A.27: Tree, Chi-square

family	0.03333333
native	0.016666666
species	0.015555553
genus	0.0111111095
called	0.008888888

Table A.28: Nurse, TF-IDF

nursing	0.13188018
states	0.09913722
united	0.0924088
practice	0.08145145
registered	0.0791664

Table A.29: Nurse, Mutual information

flight	12.29
badge	12.29
midwives	12.29
bullying	12.29
teach	12.29

Table A.30: Nurse, Chi-square

united	0.026666664
states	0.025555553
nursing	0.024444442
registered	0.019999998
american	0.017777776

Table A.31: Field, TF-IDF

artillery	0.23263976
regiment	0.21048652
army	0.15919432
states	0.14437264
united	0.14437264

Table A.32: Field, Mutual information

electromagnetic	5.81
bowdoin	5.81
coils	5.81
compact	5.81
falcon	5.81

Table A.33: Field, Chi-square

artillery	0.035555553
army	0.035555553
regiment	0.028888887
states	0.028888887
united	0.028888887

Appendix B

DBpedia, query expansion

The three methods for choosing an additional term for the query expansion are based on different calculations. In this section the terms chosen for query expansion for each query are listed with the corresponding calculation.

Table B.1: Deer

Deer	Second term	Calculation
TF-IDF	Red	0.1845
MI	Axis	9.15
Chi	Red	0.03222222

Table B.2: Christmas

Christmas	Second term	Calculation
TF-IDF	Album	0.26640436
MI	Metal	6.72
Chi	Album	0.06888888

Table B.3: Beach

Beach	Second term	Calculation
TF-IDF	Located	0.11350435
MI	Palm	7.12
Chi	Located	0.019999998

Table B.4: Cathedral

Cathedral	Second term	Calculation
TF-IDF	Church	0.18757975
MI	Portsmouth	6.29
Chi	Diocese	0.039999995

Table B.5: Oslo

Oslo	Second term	Calculation
TF-IDF	Norway	8.593762E-5
MI	Nye	7.74
Chi	Norway	0.077777766

Table B.6: LA

LA	Second term	Calculation
TF-IDF	Film	0.24179554
MI	Colle	3.01
Chi	Film	0.038888883

Table B.7: Computer

Computer	Second term	Calculation
TF-IDF	Software	0.08566075
MI	Fraud	7.51
Chi	Software	0.016666666

Table B.8: Beckham

Beckham	Second term	Calculation
TF-IDF	David	0.074879766
MI	Mccreary	23.86
Chi	David	0.04333333

Table B.9: Twilight

Twilight	Second term	Calculation
TF-IDF	Anthology	0.36020076
MI	Theater	2.44
Chi	Series	0.10222221

Table B.10: Jolie

Jolie	Second term	Calculation
TF-IDF	Film	0.1033999
MI	Lazare	21.98
Chi	Angelina	0.046666

Table B.11: NTNU

NTNU	Second term	Calculation
TF-IDF	University	0.15280305
MI	Railroad	41.8
Chi	Norwegian	0.27664956

Table B.12: Bonfire

Bonfire	Second term	Calculation
TF-IDF	Album	0.10988759
MI	Corpusty	22.5
Chi	Released	0.0377777

Table B.13: Field

Field	Second term	Calculation
TF-IDF	Artillery	0.23263976
MI	Electromagnetic	5.81
Chi	Artillery	0.035555553

Table B.14: Tulip

Tulip	Second term	Calculation
TF-IDF	Snails	0.29719922
MI	Virus	5.36
Chi	Family	0.085555546

Table B.15: Nurse

Nurse	Second term	Calculation
TF-IDF	Nursing	0.13188018
MI	Flight	12.29
Chi	United	0.026666664

Table B.16: Cloud

Cloud	Second term	Calculation
TF-IDF	Computing	0.10341233
MI	Onapp	10.23
Chi	Computing	0.21111108

Table B.17: Bun

Bun	Second term	Calculation
TF-IDF	Kong	0.067697495
MI	Penny	15.12
Chi	Sweet	0.019999998

Table B.18: Nail

Nail	Second term	Calculation
TF-IDF	Nails	0.09057493
MI	Salons	13.53
Chi	Tooth	0.028888887

Table B.19: Tree

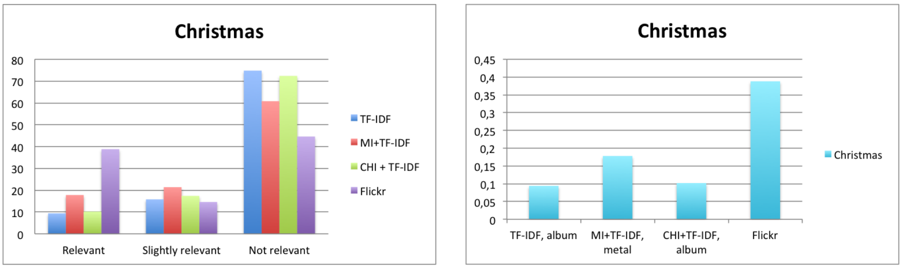
Tree	Second term	Calculation
TF-IDF	Family	0.14394906
MI	Suslin	8.46
Chi	Family	0.0333333

Table B.20: Miley

Miley	Second term	Calculation
TF-IDF	Cyrus	0.17484209
MI	Deepdale	20.29
Chi	Cyrus	0.05666666

Appendix C

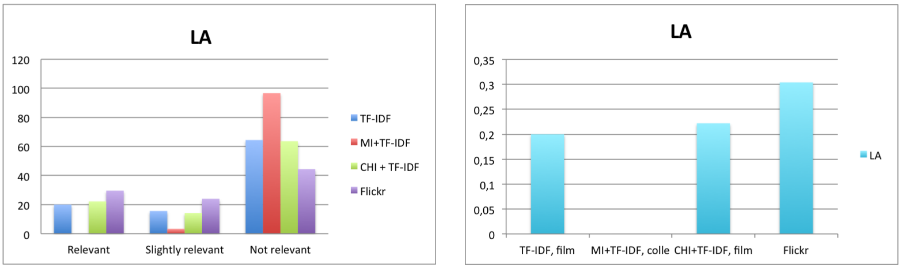
Results



(a) Christmas, percent of relevance

(b) Christmas, precision

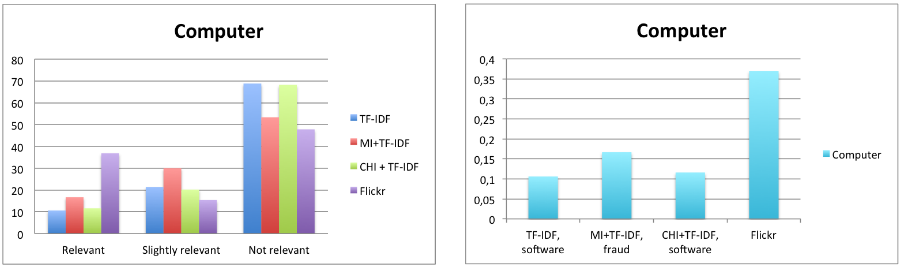
Figure C.1: Christmas, Distribution of relevant pictures, and precision.



(a) LA, percent of relevance

(b) LA, precision

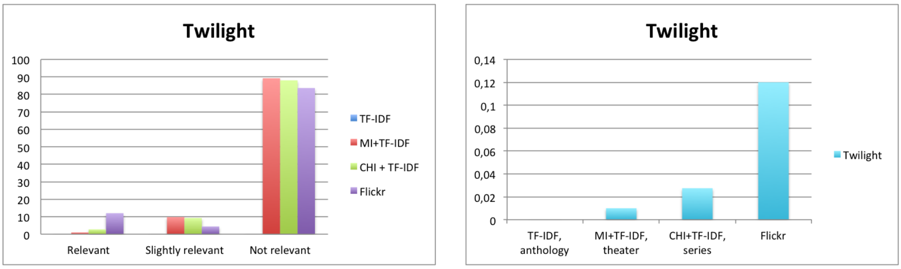
Figure C.2: LA, Distribution of relevant pictures, and precision.



(a) Computer, percent of relevance

(b) Computer, precision

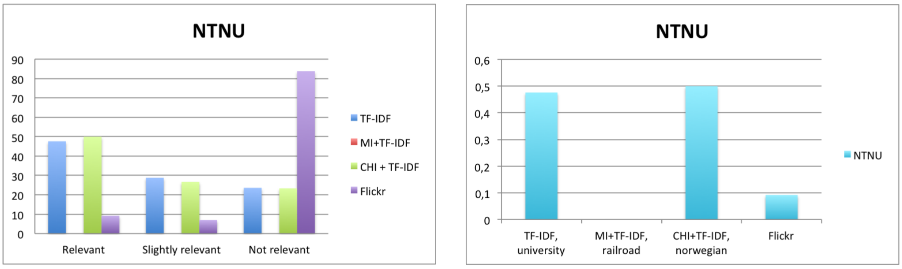
Figure C.3: Computer, Distribution of relevant pictures, and precision.



(a) Twilight, percent of relevance

(b) Twilight, precision

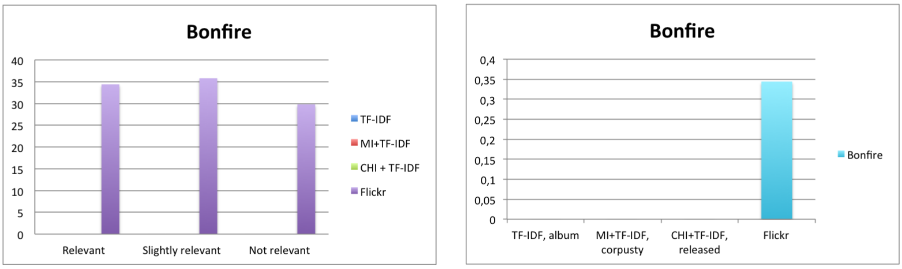
Figure C.4: Twilight, Distribution of relevant pictures, and precision.



(a) NTNU, percent of relevance

(b) NTNU, precision

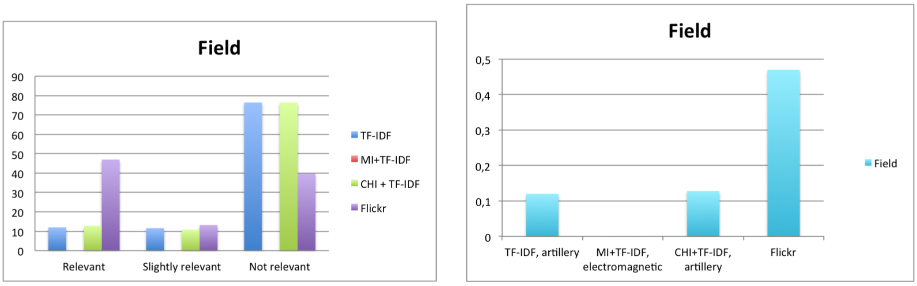
Figure C.5: NTNU, Distribution of relevant pictures, and precision.



(a) Bonfire, percent of relevance

(b) Bonfire, precision

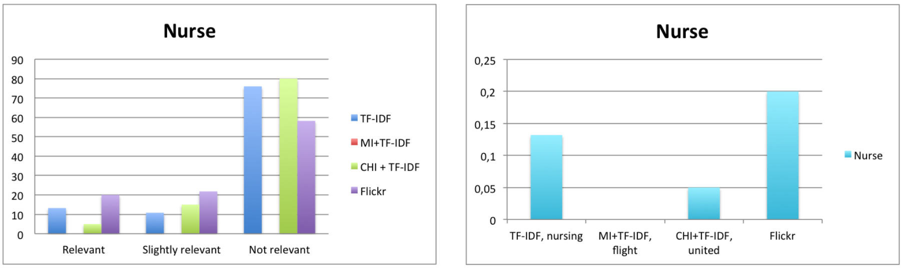
Figure C.6: Bonfire, Distribution of relevant pictures, and precision.



(a) Field, percent of relevance

(b) Field, precision

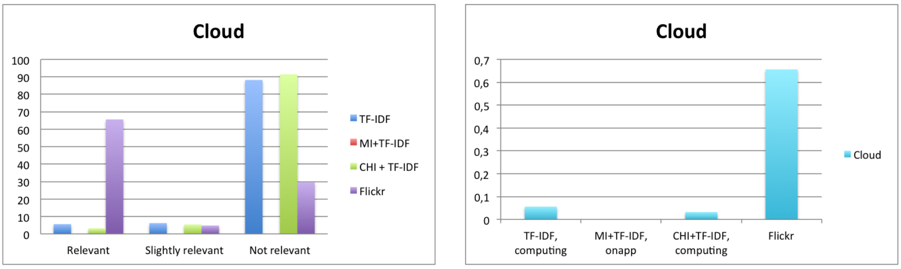
Figure C.7: Field, Distribution of relevant pictures, and precision.



(a) Nurse, percent of relevance

(b) Nurse, precision

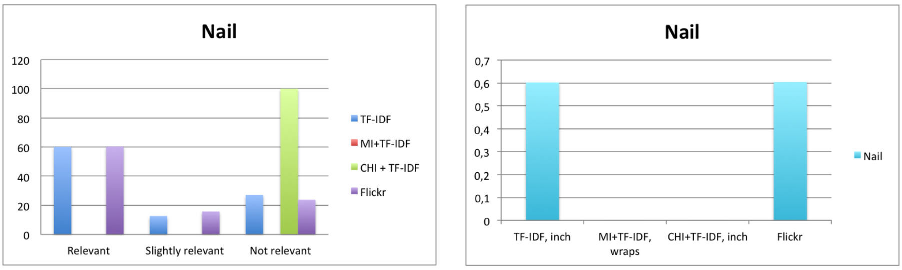
Figure C.8: Nurse, Distribution of relevant pictures, and precision.



(a) Cloud, percent of relevance

(b) Cloud, precision

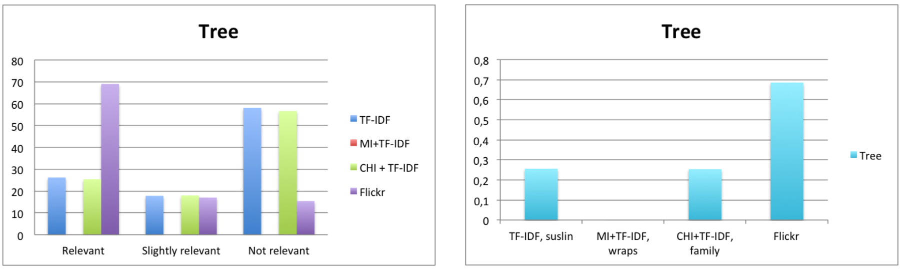
Figure C.9: Cloud, Distribution of relevant pictures, and precision.



(a) Nail, percent of relevance

(b) Nail, precision

Figure C.10: Nail, Distribution of relevant pictures, and precision.



(a) Tree, percent of relevance

(b) Tree, precision

Figure C.11: Tree, Distribution of relevant pictures, and precision.

Appendix D

Standard deviation

Table D.1: Deer

Deer	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Red	2.068	0.915	0.956	relevant
CHI	Red	2.032	0.850	0.922	relevant
Flickr	-	2.298	0.75	0.870	relevant

Table D.2: Christmas

Christmas	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Album	1.346	0.414	0.64	Not relevant
MI	Metal	1.57	0.601	0.775	Not relevant
CHI	Album	1.378	0.439	0.662	Not relevant
Flickr	-	1.942	0.830	0.911	Not relevant

Table D.3: Beach

Beach	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Located	1.228	0.288016	0.536	Not relevant
MI	Palm	1.76	0.798	0.893	Not relevant
Flickr	-	2.442	0.578	0.760	Relevant

Table D.4: Cathedral

Cathedral	Second term	AVG	Varians	SD	Highest relevance
MI	Portsmouth	1.912	2.075	1.440	Not relevant
CHI	Diocese	2.14	0.764	0.874	Relevant
Flickr	-	2.028	0.807	0.898	Relevant

Table D.5: Oslo

Oslo	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Norway	1.746	0.601	0.775	Not relevant
MI	Nye	1.016	0.016	0.128	Not relevant
CHI	Norway	1.752	0.634	0.796	Not relevant
Flickr	-	1.746	0.717	0.847	Not relevant

Table D.6: LA

LA	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Film	1.556	0.646	0.945	Not relevant
MI	Colle	1.033	0.032	0.179	Not relevant
CHI	Film	1.586	0.686	0.828	Not relevant
Flickr	-	1.856	0.731	0.855	Not relevant

Table D.7: Computer

Computer	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Software	1.424	0.456	0.675	Not relevant
MI	Fraud	1.633	0.565	0.752	Not relevant
CHI	Software	1.434	0.477	0.691	Not relevant
Flickr	-	1.892	0.836	0.914	Not relevant

Table D.8: Beckham

Beckham	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	David	1.798	0.701	0.837	Not relevant
MI	Mccreary	0.00	0.00	0.00	-
CHI	David	1.969	0.703	0.838	Not relevant
Flickr	-	1.36	0.422	0.649	Not relevant

Table D.9: Twilight

Twilight	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Anthology	0.00	0.00	0.00	-
MI	Theater	1.118	0.124	0.352	Not relevant
CHI	Series	1.147	0.180	0.425	Not relevant
Flickr	-	1.284	0.443	0.66	Not relevant

Table D.10: Jolie

Jolie	Second term	AVG	Varians	SD	Highest relevance
MI	Lazare	0.00	0.00	0.00	Not relevant
CHI	Angelina	1.884	0.197	0.90	Not relevant
Flickr	-	0.00	0.00	0.000	Not relevant

Table D.11: NTNU

NTNU	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	University	2.24	0.654	0.808	Relevant
MI	Railroad	0.00	0.00	0.00	Not relevant
CHI	Norwegian	2.267	0.662	0.813	Relevant
Flickr	-	1.254	0.373	0.611	Not relevant

Table D.12: Bonfire

Bonfire	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Album	0.00	0.00	0.00	-
MI	Corpusty	0.00	0.00	0.00	-
CHI	Released	0.00	0.00	0.00	-
Flickr	-	2.046	0.639	0.799	Slightly relevant

Table D.13: Field

Field	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Artillery	1.356	0.469	0.685	Not relevant
MI	Electromagnetic	0.00	0.00	0.00	-
CHI	Artillery	1.364	0.487	0.698	Not relevant
Flickr	-	2.072	0.862	0.928	Relevant

Table D.14: Tulip

Tulip	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Snails	2.6	0.44	0.663	Relevant
MI	Virus	2.3	0.641	0.748	Relevant
CHI	Family	1.706	0.735	0.857	Not relevant
Flickr	-	2.47	0.641	0.800	Relevant

Table D.15: Nurse

Nurse	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Nursing	1.372	0.497	0.705	Not relevant
MI	Flight	0.00	0.00	0.00	-
CHI	United	1.25	0.287	0.536	Not relevant
Flickr	-	1.618	0.636	0.797	Not relevant

Table D.16: Cloud

Cloud	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Computing	1.25	0.287	0.505	Not relevant
MI	Onapp	0.00	0.00	0.00	-
CHI	Computing	1.118	0.168	0.40	Not relevant
Flickr	-	2.36	0.822	0.90	Relevant

Table D.17: Bun

Bun	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Kong	1.109	0.206	0.454	Not relevant
CHI	Sweet	1.814	0.731	0.855	Not relevant
Flickr	-	1.594	0.661	0.813	Not relevant

Table D.18: Nail

Nail	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Nails	2.33	0.765	0.874	Relevant
MI	Salons	0.00	0.00	0.00	-
CHI	Tooth	1.00	0.00	0.00	Not relevant
Flickr	-	2.366	0.708	0.841	Relevant

Table D.19: Tree

Tree	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Family	1.686	0.727	0.852	Not relevant
MI	Suslin	0.00	0.00	0.00	-
CHI	Family	1.686	0.722	0.850	Not relevant
Flickr	-	2.536	0.548	0.74	Relevant

Table D.20: Miley

Miley	Second term	AVG	Varians	SD	Highest relevance
TF-IDF	Cyrus	2.246	0.637	0.79	Relevant
MI	Deepdale	0.00	0.00	0.00	-
Flickr	-	2.19	0.6979	0.835	Relevant

Bibliography

[1] Designing a Regional Crawler for Distributed and Centralized Search Engines.
http://ausweb.scu.edu.au/aw04/papers/refereed/shokouhi/paper.html
Retrieved 11.05.14

[2] Baeza-Yates, R. and Riberiro-Neto, B.
Modern Information Retrieval the concepts and technology behind search
Pearson Education Limited, Second edition 2011

[3] Clements, Maarten. de Vries, Arjen P. Reinders Marcel J.T.
Detecting Synonyms in Social Tagging Systems to Improve Content Retrieval
Published in SIGIR 08

[4] Cui, Hang. Wen, Ji-Rong. Nie, Jian-Yun. Ma, Wei-Ying
Probabilistic Query Expansion Using Query Logs
Published in Proceeding WWW '02 Proceedings of the 11th international conference on World Wide Web
Pages 325-332

[5] Croft, Metzler, Strohman
Search Engines, Information retrieval in practice.
Pearson Education, International Edition, 2010

[6] The dataset from DBpedia is downloaded from
http://wiki.dbpedia.org/Downloads39
Retrieved 25.11.13

[7] The importance of Word-Sense Disambiguation in Online Information Retrieval
http://lisnews.org/importance_wordsense_disambiguation_online_information_retrieval
Retrieved: 14.05.14

[8] Digital Marketing Ramblings, the latest digital marketing statis, tips, trends and technology
http://expandedramblings.com/index.php/digital-social-media-directory/8/
Retrieved 22.01.14

[9] Du, W.-H., Rau, J.-W., Huang, J.-W., and Chen, Y.-S.
Improving the Quality of Tags Using State Transition on Progressive Image Search and Recommendation System
2012 IEEE International Conference on Systems, Man, and Cybernetics

[10] Total number of Websites and size of the Internet as of 2013
http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html
Retrieved 12.05.14

[11] Flickr website
https://www.flickr.com/
Retrieved 01.05.14

[12] Flickr statistics from Frank Michel
https://www.flickr.com/photos/franckmichel/6855169886/
Retrieved 01.05.14

[13] Giannakudo, E., Kompatsiaris, I., and Vakali, A.
SEMSOC: SEMantic SOcial, and Content-based Clustering in Multimedia Collaborative Tagging Systems
2008 IEEE International Conference on Semantic Computing

[14] Gospodnetic, Otis. Hatcher, Erik
Lucene in Action
2005 Manning Publications Co

[15] What is the difference between ordinal, interval and ratio variables?
http://www.graphpad.com/support/faqid/1089/
Retrieved: 18.05.14

[16] Instagram blog
http://blog.instagram.com/post/39142353889/year-in-review-instagram-in-2012-2012-has-been
Retrieved 22.01.14

[17] Instagram website
http://instagram.com/press/
Retrieved 22.01.14

[18] Internet live stats
http://www.internetlivestats.com/total-number-of-websites/
Retrieved 19.05.14

[19] Liu, Dong. Wang, Meng. Yang,Yichen. Hua, Xian-Sheng and Zhang. Hong Jiang
Tag quality improvement for social images
Published in Multimedia and Expo, 2009. ICME 2009. IEEE International Conference
Pages 350 - 353

[20] Liu, Ling, Özsu, M Tamer,
Encyclopedia of Database Systems
Springer 2009

[21] Apache Lucene - Scoring
http://lucene.apache.org/core/3_6_2/scoring.html
Retrieved 17.02.14

[22] Manning, Christopher D. Raghavan, Prabhakar and Schütze, Hinrich
Introduction to Information Retrieval
Cambridge University Press, Online edition 2009

[23] Min, Jimming. Leveling, Johannes and Jones, Gareth J.F. *Document Expansion for Text-based Image Retrieval at WikipediaMM 2010*
Published in Proceeding RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information,
Pages 65-71

[24] Moëllic, Haugeard and Pittel
Image clustering based on a shared nearest neighbors approach for tagged collection
Published in Proceeding CIVR '08 Proceedings of the 2008 international conference on Content-based image and video retrieval,
Pages 269-278

[25] Nielsen Holdings N.V. and NM Incite
State of the media: the social media report 2012
Nielsen U.S. social media survey 2012

[26] Instagram statistics from Nitrogram
http://nitrogr.am/instagram-statistics/
Retrieved 10.02.14

[27] Oates, Briony J.
Researching Information Systems and Computing
Sage Publications Ltd 2011

[28] Sanderson, Mark.
Word Sense Disambiguation and Information Retrieval
Published in proceeding SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval,
pages 142 - 151

[29] Apache Solr
http://lucene.apache.org/solr/
Retrieved 11.02.13

[30] Solr Features

http://lucene.apache.org/solr/features.html

Retrieved 11.02.13

[31] Stop word list

http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

Retrieved 30.11.13

[32] Synonyms

http://www.oxforddictionaries.com/definition/english/synonym

Retrived 14.05.14

[33] Piatrik,Tomas. Zhang, Qianni. Sevillano, Xavier and Izquierdo, Ebroul

Predicting User Tags in Social Media Repositories Using Semantic Expansion and Visual Analysis

Pages 143-167, Springer London, 2013

[34] About Pinterest

http://about.pinterest.com/

Retrieved 11.02.14

[35] Popescu, Adrian and Grefenstette, Gregory

Social media driven image retrieval

Published in proceeding ICMR '11 Proceedings of the 1st ACM International Conference on Multimedia Retrieval

[36] Qi, Guo-Jun. Aggarwal, Charu. Tian, Qi. Ji, Heng and Huang, Thomas

Exploring Context and Content links in Social Media: A Latent Space Method
Published in Journal IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 34 Issue 5, May 2012
Pages 850-862

[37] Facebook statistics from Statistic Brain
http://www.statisticbrain.com/facebook-statistics/
Retrieved 10.02.14

[38] Chapter 09 Thesaurus construction
http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap09.htm
Retrieved 06.03.14

[39] Twitter, about
https://about.twitter.com/milestones
Retrieved 11.02.14

[40] Powered by
http://wiki.apache.org/lucene-java/PoweredBy
Retrieved 14.02.14

[41] Vallet, David. Cantador, Iván and Jose, Joemon M.
Exploiting external knowledge to improve video retrieval
Published in Proceeding MIR '10 Proceedings of the international conference on Multimedia information retrieval
Pages 101-110

[42] Web Informaton Retrieval

http://research.microsoft.com/en-us/people/hangli/webir-msra.pdf

Retrieved 12.05.14

[43] Word Sense Disambiguation

http://aclweb.org/aclwiki/index.php?title=Word_sense_disambiguation

Retrieved 14.05.14