T-75.4400 Information retrieval

Assignment 2

On the Efficacy of Alternate Similary Measures, Stemming and Stop Word Removal in Text-Based Information Retrieval Systems

Antti Partanen
antti.partanen@aalto.fi
295967

Vikram Kamath
vikram.kamath@aalto.fi
440819

Aalto University
School of Science

April 23, 2015

Abstract

As the amount of data available in electronic form increases continuously, efficient and effective information retrieval techniques are critical. It has been estimated that the amount of data on the Internet doubles every second year. Thus, it is no wonder, that most visible current information retrieval systems are web search engines.

In this paper, we compare compare two ranking methods: VSM (Vector Space Model) and BM25 for document collection from ACM digital library database. Furthermore, we study the effects of stemmers and stop words. For this study, we use Java with Apache Lucene.

1 Introduction

The task of Information retrieval is to obtain information resources, including books, journals and other documents, relevant to the imminent need from information resources. Searches can be based on meta-data or on full-text indexing.

Automated information retrieval systems are used to reduce the so called *information overload*, i.e. the difficulty of a person to understand an issue due to overwhelming amounts of information. Most of the existing text retrieval techniques rely on indexing key-words, although keywords and index terms alone do not sufficiently capture the documents content. Still, the computational difficulties of more advanced systems keeps the keyword indexing as the most viable way by far to process large amount of text.

The generic textual information retrieval process is show in Figure 1 and it has the following steps:

1. The system builds and index of the documents (*indexing*)

2. User describes the *information need* in a form of a query, which is parsed and transformed by the system with the same operations applied to the documents (*query formulation*)

3. The system retrieves, ranks and displays documents that are relevant to the query from the index

4. User may give relevance feedback to the system

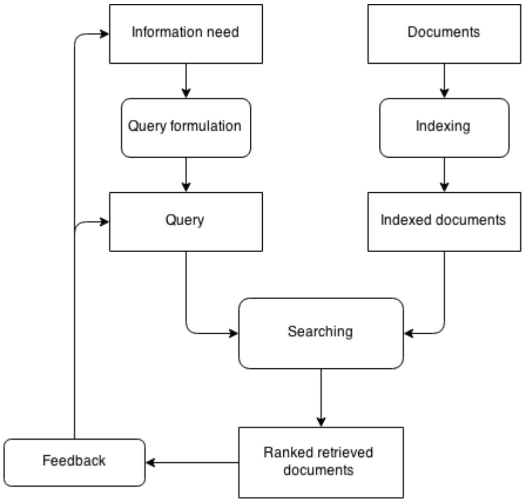

Figure 1: Information retrieval pipeline (Hiemstra, 2009)

Efficient and effective information retrieval techniques are critical in managing the increased amount of textual information available in electronic form. Therefore, it is no wonder that currently the most visible IR systems are web search engines.

1.1 Tools

We used Java with Apache Lucene, to carry out our experiments. Apache Lucene is a free open source information retrieval software library, developed originally by Doug Cutting in 1999. Lucene is capable of offering high-performance and includes full-featured text search engine.

1.2 Content

In this paper, we compare the two information retrieval techniques: Vector Space Model (VSM) and BM25. Furthermore, we analyze what effects does the usage of stemmers (porter stemmer) and stopwords have with the retrieval result. The evaluation of our experiments is visualized with eleven point precision recall curves.

2 Techniques

Ranking methods are used by search engines to rank matching documents according to their relevance to a given search query. Most common approaches to information retrieval are algebraic (e.g. VSM) and probabilistic (e.g. BM25).

2.1 VSM

VSM (Vector Space Model) is an algebraic representation model for objects as vectors of identifiers. Beyond information retrieval, VSM is also used in information filtering and indexing. The concept dates back to the early days of IR and it was first utilized in SMART system (Dubin, 2004).

The core idea of VSM is rather simple: represent documents (D_j) and queries (Q) as vectors of weights (Salton et al., 1975).

$$D_j = (w_{1,j}, w_{2,j}, ..., w_{t,j}) \tag{1}$$

$$Q = (w_{1,q}, w_{2,q}, ..., w_{n,q}) \tag{2}$$

There are several different ways of calculating these weights, but *tf-idf* is one on the best known. Documents are ranked according to their proximity to the query in this space, where proximity corresponds to the similarity of the vectors. Similarity is not calculated as an euclidean distance, but rather as an cosine between the vectors.

$$cos\theta = \frac{D_j \cdot Q}{\|D_j\|\|Q\|} \tag{3}$$

Where the norm or length $\|Q\|$ is calculated as

$$\|Q\| = \sqrt{\sum_{i=1}^{n} q_i} \tag{4}$$

2.2 BM25

BM25, often referred to as Okapi BM25 (where BM stands for Best Matching), is a probabilistic ranking function. It is based on probabilistic relevance model, devised by (Robertson et al., 1996).

BM25 uses a bag-of-words representation of a document, a representation that ignores the relative ordering of words in the document and the query. It isn't one stand-alone method, but is actually a collection of functions, each of which uses different hyper-parameters, the outputs of which are combined to give the final similarity score between a document and a query. BM25 uses both term frequencies and inverse document frequencies to compute the final similarity score. The BM25 retrieval formula is part of the BM family of retrieval models and according to the Encyclopedia of database systems ((Amati, 2009)), it defines the document-query matching function as:

$$\sum_{t \in q} (k+l) \frac{\boldsymbol{tf}}{k+\boldsymbol{tf}} \cdot (k_3 + ld) \frac{\boldsymbol{tf}_q}{k_3 + \boldsymbol{tf}_q} \cdot ln \frac{(r_t + 0.5) \cdot (N - R - n_t + r_t + 0.5)}{(n_t - r_t + 0.5) \cdot (R - r_t + 0.5)} \tag{5}$$

where

- R is the number of documents know to be relevant to a specific topic,
- r_t is the number of relevant documents containing the term,
- N is the number of documents of the collection,
- n_t is the document frequency of the term,
- $\boldsymbol{tf}$ is the frequency of the term in the document,
- q is the query,
- $\boldsymbol{tf}_q$ is the frequency of the term within the topic from which q was derived
- l is the document length
- k is $k_1((1-b) + b(\frac{l}{1}))$
- $k, 1$, b and k_3 are parameters which depend on the nature of the queries and possibly the database

2.3 Stemmers

In most languages and predominantly so in English, words occur in different forms even though they have the same meaning. For example ‘democracy’, ‘democratic’ , ‘democratize’ all have the same derivation and although they’re used in different contexts, they have the same ‘semantic’ meaning. More often than not, it would be useful to search for using just one of the forms and have documents retrieved that contain all semantically similar forms. Lemmatization is a crude heuristic process that aims to achieve this by removing the ends of the words so that different forms of the same word, after processing, lead to the same, shortened form. So in the example above, all three words would map to ’democraci’ after processing. We used the porter stemmer [TODO: Insert Citation] for our experiments (although there exist many other different stemming algorithms that aim to achieve a good ’lemmatization’)

2.4 Stopwords

Most search engines preprocess both indexed text and search queries so as to remove commonly occurring words that are of little value to the context of the document. Words like ‘A’, ‘and’,‘the’, ‘of’ etc are found abundantly in most English documents but add no overall value to the indexing, ranking or retrieval process. The average size of stopword collections is between 200 - 300 terms [TODO: Insert Citation]. We experimented with using the Lucene in-built stopword list, which although not exhaustive (relative to some of the stop-word corpora available), is sufficiently good.

3 Evaluation

To evaluate the results, we used precision, recall and *11-point precision recall curves*. Precision (a.k.a. positive predictive value) is the fraction of the *retrieved relevant documents to the number of retrieved documents*, whereas recall (a.k.a. sensitivity) is the fraction of the *retrieved relevant documents to all relevant documents* (Buckland and Gey, 1994). If we denote the amount of *relevant documents* with N_{rel} and the amount of *retrieved documents* with N_{ret}, we get the following formulas for precision and recall:

$$precision = \frac{|N_{rel} \cap N_{ret}|}{|N_{ret}|} \tag{6}$$

$$recall = \frac{|N_{rel} \cap N_{ret}|}{|N_{rel}|} \tag{7}$$

Eleven point precision-recall curve is a graph that plots recall and precision of an information retrieval system at 11 equally spaced recall levels $recall = (0.0, 0.1, 0.2, ..., 1.0)$ (Zhang and Zhang, 2009). Uninterpolated precision-recall graph has a saw-tooth like shape.

4 Conclusions

Add conclusions

5 Contributions

In the end both of us put equal amounts of effort into the project. We used GitHub [TODO: Citation?] for version control and although we did some independent work at the comfort of our own homes, a bulk of the work was done in a series of 'hackathon' style, caffeine-fueled meetings, where we just sat for 4-6 hours straight at Maari and woked on it. Although this style might be deemed by some as a sloppy, we'd like to think of it as pair programming on speed - in that the fact that we sat together and worked on it enabled almost seamless debugging and ideation.

A diagrammatic (albeit) rough division of tasks is shown below [TODO: Should we add this?][Probably not]

References

Giambattista Amati. Bm25. In *Encyclopedia of Database Systems*, pages 257–260. Springer, 2009.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Michael K. Buckland and Fredric C. Gey. The relationship between recall and precision. *JASIS*, 45(1):12–19, 1994.

David Dubin. The most influential paper gerard salton never wrote. 2004.

Djoerd Hiemstra. Information retrieval models. *Information Retrieval: searching in the 21st Century*, pages 2–19, 2009.

Stephen E Robertson, Steve Walker, MM Beaulieu, Mike Gatford, and Alison Payne. Okapi at trec-4. In *Proceedings of the fourth text retrieval conference*, volume 500, pages 73–97. NIST Special Publication, 1996.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL `http://doi.acm.org/10.1145/361219.361220`.

Ethan Zhang and Yi Zhang. Eleven point precision-recall curve. In *Encyclopedia of Database Systems*, pages 981–982. Springer, 2009.

A Installation instructions

Here is a brief systems installation and system configuration instructions.

1. **Preprocessing**
   We wrote a small script to parse the XML file and dump ‘items’ relevant to our search task in a document called ‘data.xml’. It is assumed that this file will be used as a command line argument to the program. To create a fresh dump, run ‘parseXML.py’ and ensure that the corpus file is in the same directory as the script.

2. **Java Indexing, Search and Ranking Implementation**
   The whole Java project ‘Assignment2’ once loaded into your Eclipse workspace will make available the file ‘assignment2.java’, that contains our implementation. Make sure to use all our imports as we use the Java HashMap implementation in our code.

3. **Plotting, Metrics**
   We dumped the output of ‘assignment2.java’ into csv files, one for each query-parameter combination (24 in total - 8 per query). The csv files, ordered according to query are in a folder called ‘PrecRec’, which further contains 3 folders, one per query containing csv files relevant to the query. Each of the query folders also contains a python script called ‘calcPrecRecall.py’that consumes the csv files in the folder and computes the 11-point non-interpolated precision recall curve, which is then saved into a ‘plot.png’ file in the same folder.