Aalto University

T-75.4400 Information retrieval

Assignment 2

Comparing ranking methods for information retrieval

Antti Partanen
antti.partanen@aalto.fi
295967

Vikram Kamath
vikram.kamath@aalto.fi
440819

April 21, 2015

Aalto University
School of Science

1 Introduction

The task of Information retrieval is to obtain information resources relevant to the imminent need from information resources. Resources include books, journals and other documents. 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 existing text retrieval techniques rely on indexing key-words. Unfortunately keywords or index terms alone cannot adequately capture the document contents, resulting in poor retrieval performance. Yet, keyword indexing is widely used in commercial systems because it is still 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

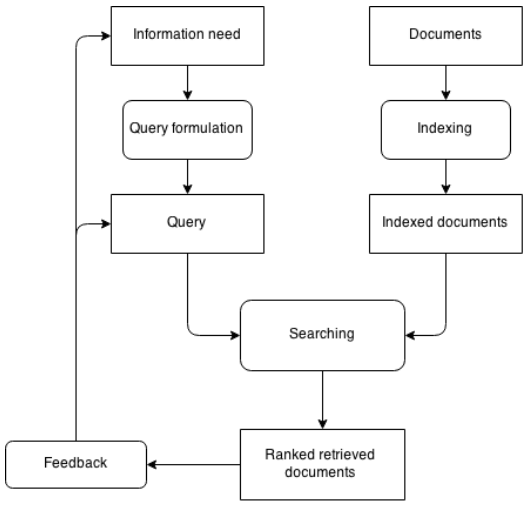

Figure 1: Information retrieval pipeline (Hiemstra, 2009)

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.

Apache Lucene

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.

Draft version: aka needs modifying + references

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. It was first introduced by (**?**) and it's first use were in SMART IR system (**?**).

The core idea of VSM is to represent documents (D_j) and queries (Q) as vectors in vectorspace.

$$D_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$$
$$Q = (w_{1,q}, w_{2,q}, ..., w_{n,q})$$

Documents are

2.2 BM25

BM25, often referred to as Okapi BM25 (where BM stands for Best Matching), is a probabilistic ranking function.

2.3 Stemmers

2.4 Stopwords

3 Evaluation

Add text and charts

Precision
Recall
Eleven point precision recall curve

4 Conclusions

Add conclusions

5 Contributions

Mind map style of work division might look neat?

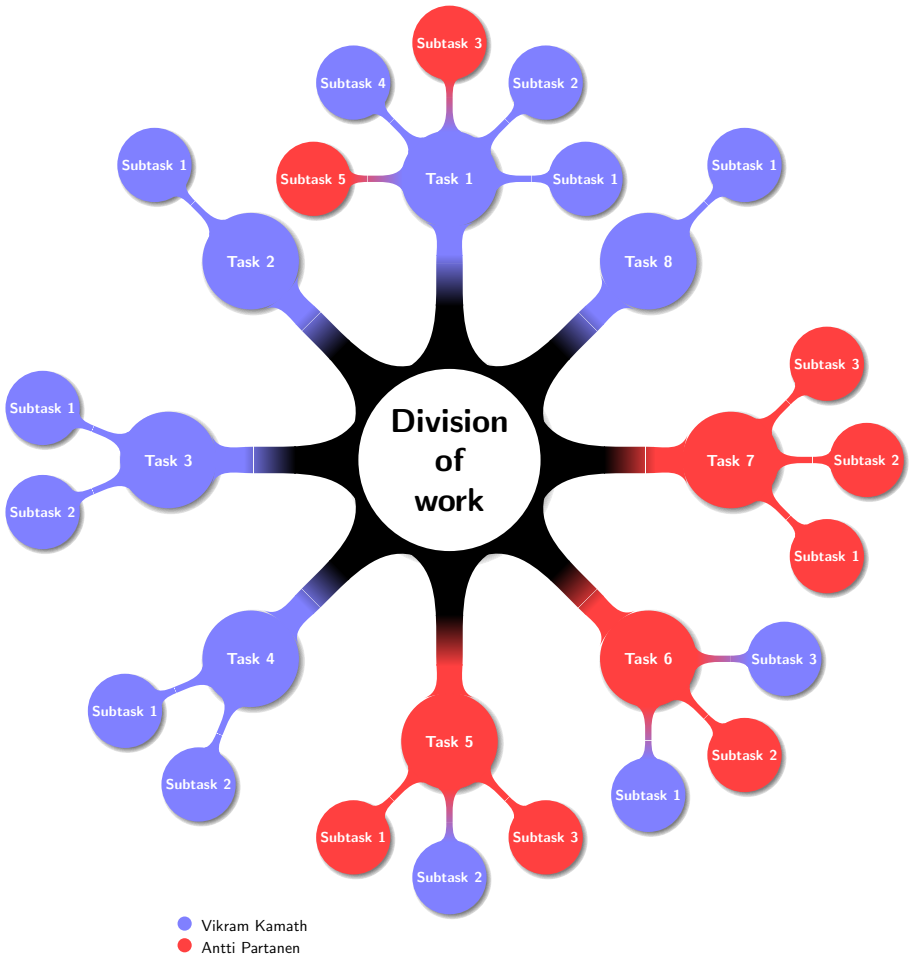

Figure 2: Contributions represented in a mindmap

References

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

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

Add more references

A Installation instructions

Here is a brief systems installation and system configuration instructions.

1. ...
2. ...