TDT4215 - 1 - Exercise

Magnus L Kirø

February 19, 2013

Task 1

The most important parts of IR is document retrieval and text mining.

Document retrieval is the part where we aquire the collection of text. It's the raw data that we can extract information from later.

Text minig is the process of analysing the retrieved documents. Tagging, lemmatization and otherwise adding meta data to the raw data to better be able to extract information.

Task 2

Boolean model: Considers the percense of terms. It's there or it's not. The advantage of the boolean model is that the formalism is clean and the model is simple. The downside is that there is no ranking in this model.

Vector model: It considers documents that only partially match the query. Despite it's simplicity the model gains good results. This is mostly due to the term veighting scheme and the lenght normalization.

Probabilistic model: It operates with an ideal answer set. This is the set of documents that accrately fits the description of the query. The query process can then be thought of as defining the properties of the ideal answer set.

Task 3

Information retrieval is: The act of finding and recovering specific information from data sources.

Task 4

The number of records, relevance and user satisfaction can be used to rate a given retrieval ersult.

Although the common tasks of document retrieval can be individually timed and rated. The main parts are "document indexing", "query interpretation", "ranking of retrieved documents", and "linguistics and statistics".

Document indexing is quite easy to get the time of and therefore measure the efficiency of the indexing. The quality of the indexing can also be measured but that is a bit harder.

Query interpretation is difficult but tim emeasurable. This is one of the parts that has been greatly improved over time. And would likely be improved in the future.

The ranking of the rerieved documents is a difficult task. How do we rank the documents? And what is the accuracy of the ranking? Also the relevance of a given document will vary depending on the recipient of the given document.

Statistics are not directly used in the measuring of results from a document retrieval. But all the retrievals combined becomes the statistics that says what works and what doesn't.

Task 5

Interpolated precision is the maximum known recall of all the levels above the given level.

Task 6

Task 7

We have different types of basic queries, some of them are:

Single-word queries, contex queries and boolean queries. Context queries consider proximity, the space between words. Single-word queries are ranked according to relevance.

"Shiing", "Trondheim", "Hybel" are query examples of singel-word queries, while "new york times" is an example of a context query.

Boolean queries are based on the precense of terms. There or not there. Ranking is not provided for boolean queries. The queries are like this: (e1 OR e2), (e1 AND e2)

Task 8

What charactherises structural queries are that the search is based mainly on the structure of a document. The content is unimportant in this kind of

search. A mix of structure and content in a search will allow richer and more expressive queries.

There are three main types of structural queries; form-lie, hypertext and hierarchical.

Task 9

Ranking is supported by single-word queries, context queries and natural language queries supports ranking. Boolean queries does not.

Task 10

Ther are many typical problems with web searches. Some of them are; ambiguity, to many results, bad queries, data indexing, data size, time constraints and bandwidt.

Didifferent problems are visible for different parties to a search. Bandwidt and delay are typically a network problem that affects the user the most. While the data size and complexity of the raw data are problems for the program/developer to deal with.

Task 11

By using all words in a dictionary we create lots of noise. The noise is the words that are very common. Like the word like. Or as, it, is, for, are.

Mainly not all words are relevant for the meaning of the text.

Task 12

An initial query can be improved by removing noise terms. Such as: is are it etc. Further improvement can come fro suggesting word improvements due to spelling mistakes. Or a change of search words can improve the query.

Task 13

Pre-processing of queries are common. It's used mostly to reduce the DB-access time. And to get a more accurate result back.

A pre-processing step can do things like indexing the query, splitting the query, distribute the query etc. It's quite common to have a query plan that deals with the execution of pre-processing and query execution.

Task 14

Document preprocessing:

Often starts with lexical analysis, treating charachters that is unimportant for a hindrance later.

Continuing with the elimination of stopwords. This filters out words with low discrimination values. Improving document retrieval.

Stemming. Taking a word and remoing affixes. Parts of words, at the start or end, that bears no meaning. Connect, connected, connecting are words that would be the same after stemming. This improves the indexing of documents.

Selecting keywords to determine the indexes of the document. Nouns are favored to be selected as index words.

Term categorisation follows as the last stage. Here structure is extracted from the text allowing the original query can be expanded.

Task 15

Linguistics are used to gather information from text and queries. It also gives insight into how and why information is stored. Linguistics is the science disipline that studies the human language.

Task 16

Inverted index (or file) is mechanism that speeds up searching in a text collection. It consists of two main parts: the vocabulary(lexicon/dictionary) and the occurences.

	vocabulary	n	d
	any	1	[1,1]
	fool	1	[1,1]
	can	1	[1,1]
	make	1	[1,1]
	things	1	[1,1]
	bigger	1	[1,1]
	more	2	[1,2]
	complex	1	[1,1]
	and	2	[1,2]
	violent	1	[1,1]
	it	1	[1,1]
conversion of text:	takes	1	[1,1]
	a	2	[1,2]
	touch	1	[1,1]
	of	2	[1,2]
	genious	1	[1,1]
	lot	1	[1,1]
	courage	1	[1,1]
	to	1	[1,1]
	move	1	[1,1]
	in	1	[1,1]
	the	1	[1,1]
	opposite	1	[1,1]
	direction	1	[1,1]

A compression of this inverted index list would be:

$$1, [1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1]$$

Task 17

Serching an inverted index follows three general steps:

1: Vocabulary search. Finding the occurences of search terms in the vocabulary.

2: Retrieving occurences. Compiling list of documents containing seach terms.

3: Manipulation of occurences. The compliance to the query is upheld. This part can go deeper into the documents to find the exact information wanted, not just the document wich contains the wanted information.

Task 18

Expressing queries and interpreting the results are two problems that has to be addressed with web searches.

Human input is difficult to interpret. The input is just a reflection of the information need and therefore imperfect.

If the user input is a perfect query we will stil have problems with the interpretation of the results. The result set might be to big. How do we handle that? And how do we rank the results? There are a lot of problems in regard to the presentation of the results.

Task 19

Centralsed web architecture and distributed web architecture. The main difference is that the centralised architecture has a master copy of the information in one location while the rest of the system has copies, while the distributed system has lots of main chunks of information with no master copy. In the distributed system the information you are looking for might not be in the nearest location. You might need to ask the first location first before being redirected to the next location.