Week 6 – Information Retrieval WS 15/16

Johannes Jasper(755292)
Norman Rzepka(754644)

November 24, 2015

Assignment 1

a)

When does relevance feedback not work?

Relevance feedback requires a user who defines which documents are relevant. If there are too few users only pseudo relevance feedback can be used which relies on the first ranking to be correct.

b)

What is the idea underlying Rocchios algorithm for pseudo relevance feedback? Do not use any formulas.

Rocchios algorithm moves the query vector in the direction of relevant documents, i.e. it tries to maximize the similarity with relevant documents and minimize similarity with irrelevant documents. It does so by summing and normalizing the feature vectors of documents. Relevant documents are added to the original query vector, irrelevant documents are subtracted. 3 weights are used that denote the influence of the original query vector, the relevant and the irrelevant document vectors.

c)

What are the main advantages and disadvantages of pseudo relevance feedback? Use your own words.

Pseudo relevance feedback does not require user input. It can therefore be applied as a local optimization before showing results to the user. Since it uses terms from the top ranked X documents, however, it relies on the assumption, that the initial ranking was valid. It might create a kind of "self-fulfilling prophecy" as documents that are considered relevant are never ranked down. Using pseudo-relevance feedback slows down the searching of documents by a large factor (in our case 10-20x).

Assignment 2

Compute the following measures for the terms "data" and "unit": Mutual information, Expected mutual information , Chi-square and Dice's coefficient. Consider that $n_a = 1000$, $n_b = 3000$, $n_{a,b} = 2000$ and $N = 15000$.

MI

$$\frac{n_{a,b}}{n_a \cdot n_b} = 0.000\overline{6}$$

EMI

$$n_{a,b} \cdot log\left(N \cdot \frac{n_{a,b}}{n_a \cdot n_b}\right) = 4605.17019$$

χ^2

$$\frac{\left(n_{a,b} - \frac{n_a \cdot n_b}{N}\right)^2}{n_a \cdot n_b} = 1.08$$

Dice

$$\frac{n_{a,b}}{n_a + n_b} = 0.5$$

Assignment 3

Another method for improving a search engine's results is to exploit the synonyms of the query terms. One can expand the query with the query terms' synonyms and attempt to retrieve more relevant documents to the original query. This technique can be helpful in order to retrieve documents with similar topic to the original query, but different vocabulary terms.

a)

Which are the pros and cons of using synonyms for query expansion?

If synonymes are used to expand the query the user gets documents that satisfy the information need but not the query (terms). This only works, however, if correct synonymes are chosen which is often not the case due to ambiguity. To reduce the ambiguity one would need to consider the context f the query term (i.e. preceeding and succeeding terms). This on the other hand makes precomutation of good expansions very cost inefficient.

b)

In general, lexicons can be used to find synonyms of a given word. Are you aware of any publicly available lexicons? Are they domain dependent or general lexicons?

Many generic lexica or thesauri have online access, Merriam-Webster for example has an API[1] that yields synonymes. Big Huge Thesaurus[2] has stronger focus on synonymes and clusters them e.g. into verbs, nouns or phonetically simmilar words. The free dictionary has specializations for certain domains such as the legal domain[3].

Assignment 4

As we used a query likelihood model to rank our results, we implemented pseudo relevance feedback using a relevance model. We computed a relevance model over the set of relevant documents, expanded the query based on the relavance model, searched again and ranked them again using the KL-divergence.

Assignment 5

For spelling correction we consider only terms that do not exist in the index. We keep the first letter and compute the normalized Levenshtein distance with all terms in our index that start with the same letter. We replace the term in our original query with a term that has minimal normalized Levenshtein distance.

[1] `http://www.dictionaryapi.com/products/api-collegiate-thesaurus.htm`
[2] `https://words.bighugelabs.com/api.php`
[3] `http://legal-dictionary.thefreedictionary.com/`