Information Retrieval

Chiel Kooijman (5743028), Michael Cabot (6047262)

December 2, 2012

1 Introduction

One way of improving the performance of retrieval systems is by query expansion, i.e. expanding a query with terms that are expected to be found in relevant documents. In this experiment two different approaches are compared: Pseudo-relevance feedback, and expansion with synonyms of the words in the original query. Pseudo-relevance feedback expands the query by adding terms that occur more often than average in the documents that are retrieved with the original query. In synonym expansion, synonyms are obtained from a dictionary and added to the original query with different weights. A study by Hersh et al. (2000) has assessed query expansion using thesaurus relationships. They conclude that thesaurus-based query expansion generally causes a decline in retrieval performance. Our goal is to verify their results, explore the effects of synonym expansion and see how its results compare to those of pseudo-relevance feedback.

2 Method

Our experiments used a data set containing four different types of indexable features: body, headers, metadata and titles. No stemming or removal of stop-words was applied to the data set before indexing. Dirichlet smoothing was used for document retrieval.

When applying synonym expansion, synonyms were obtained from `http://thesaurus.altervista.org/`, a web service providing search capability for synonyms in different languages. Synonyms of query terms were added to the query with a weight lower than the weight of the original query terms. All words in the original query were used to find synonyms. For example the query "genetic modification" results in two queries that yield the terms

- familial, hereditary, inherited, transmitted, transmissible, inheritable, genic, genetical, sequence, beginning, genetical, biology, biological science
- alteration, adjustment, change, copy, qualifying, limiting, grammatical relation, change, alteration, happening, occurrence, occurrent, natural event

It is possible for the web service to return multiple sets of synonyms for a single query term. Each synonym set belongs to a single semantic interpretation of the term. Take for example the query "climate change":

- clime, environmental condition, mood, condition, status
- alteration, wear, wearable, cash, difference, get dressed, replace (some terms omitted for readability)

The synonyms for "change" contains synonyms from different semantic interpretations. The synonym "alteration" is more applicable to the context of the query than "get dressed". This shows that the correct set of synonyms depends on the context in which the query term is used. We did not differentiate between these synonyms sets, and used all terms with equal weight. Each term in the original query would receive a weight of 1.0 while the synonyms all would receive the same weight between 0.0 and 0.5.

Indri's built-in query expansion feature was used to perform pseudo-relevance feedback. As opposed to synonym expansion, pseudo-relevance feedback first retrieves ranked documents using the original query. Words that correlate with the n highest ranked documents are then used to expand the query.

The code for synonym expansion can be found at `https://github.com/chielk/mc_ir/tree/master/assignment_2`.

3 Results

We have evaluated the MAP score of three different methods: Query expansion by synonyms, query expansion by pseudo-relevance feedback, and the original queries without expansion. The added synonyms were evaluated at weights 0.0, 0.1, 0.2, 0.3, 0.4, and 0.5 of the original terms. For pseudo-relevance feedback, the top 20 highest ranked documents were used of which 10 new terms were added to the original query. The original query received weight 0.5. The results are shown in figure 1.

The figure shows that the performance of synonym-expansion (blue line) decreases as the weight of the synonyms increases. Pseudo-relevance feedback (red line) performs slightly better compared to using no query expansion at all (green line).

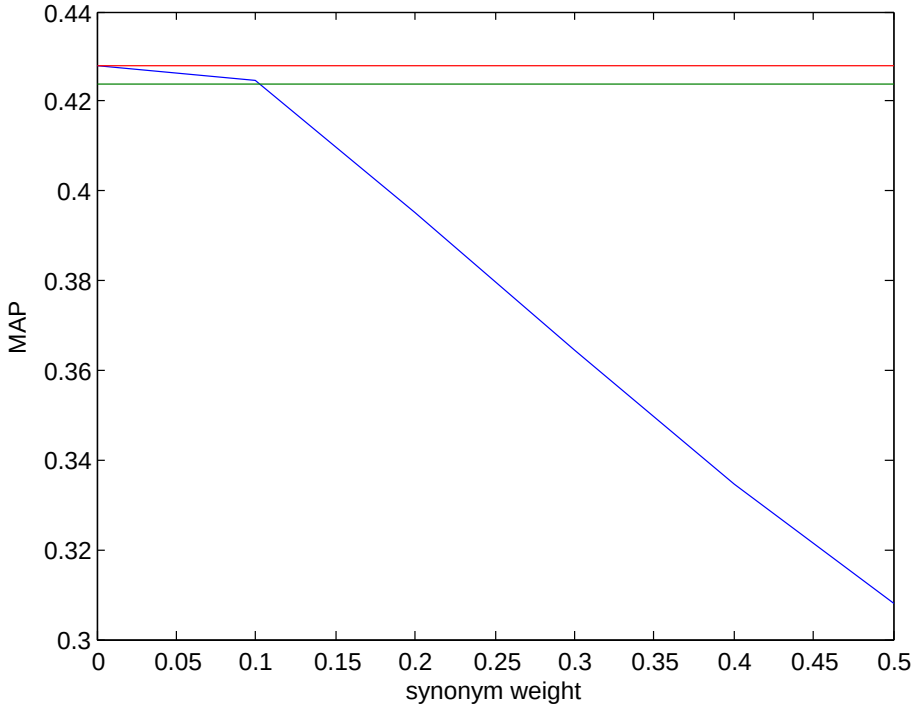

Figure 1: The green line denotes the MAP score without query expansions. The blue line shows the MAP scores for expanding the queries with synonyms for different weights. The red line denotes the MAP of pseudo-relevance feedback.

4 Conclusion and Discussion

We have used the built-in query expansion methods from Indri, and although this did improve the results, changing the parameters did not have any effect on the performance of the system. Using the synonyms with a weight of 0 also improved performance over the normal query where it was expected to perform exactly the same. As we were unable to explain these anomalies, we should be very careful to draw conclusions from this experiment.

The reason synonym expansion has a negative effect on the performance might be due to the fact that the data set came from a scientific organisation. The vocabulary used in these documents is possibly very specific, with very specific meanings and few synonyms. Therefore query expansion by synonyms may perform worse on this set compared to data from other sources. The low performance might also be caused by the uniform distribution of the weights over all the synonyms. It is possible that the performance of synonym-expansion would increase if each synonym was weighted by how relevant it is to the context of the query. Language models could be used to give "alteration" a higher weight than "get dressed" when the query is "climate change".

We have tried to explore the effects of synonym expansion and see how its results compare to those of pseudo-relevance feedback. Due to anomalies in our results the extents and limitations of synonym expansion should be explored further before any real conclusions can be made.

References

W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of the AMIA Symposium*, page 344. American Medical Informatics Association, 2000.