Rocchio's Relevance Feedback Algorithm in basic vector comparison and LSI models (ROCKNROLL)

Arnaud Fietzke* Renata Dividino†

January 21, 2005

1 Introduction

Relevance feedback is a query reformulation technique that improves the effectiveness of information retrieval.

The basic idea is to do an initial query, get feedback from the user as to what documents he or she considers relevant and then use this information to supplement and enrich the user's initial query, allowing greater retrieval performance [2]. Relevance feedback is an iterative process, which by adding more terms to the query, can help the user to more accurately specify his or her information need.

There are many variants of relevance feedback in information retrieval [1]. However, in this project we will use a variation of Rocchio's similarity-based relevance feedback algorithm [2].

The original algorithm operates on a vector space model in which both documents and queries are represented as vectors in term space. This motivates us to also experiment with Rocchio's algorithm in the context of Latent Semantic Indexing (LSI), in which terms and documents (as well as queries) are represented as vectors in a term-document-space of reduced dimension.

As the similarity between two vectors in the term-document space is used to compute similarities between documents and queries, the vector space model allows to return a ranked list of all documents in the collection w.r.t. to a given initial query. In practice, the user will then select the relevant documents from the set of the top-k ranked documents, for some (low) number k. These marked documents can then be passed to the algorithm, wich will use them to refine the initial query. This can be done while using basic vector comparison in a full-dimensional term-document space, as well as in the reduced vector space used in LSI.

2 Project Organization

In this project two information retrieval models will be implemented and tested: basic vector comparison and LSI and we will apply relevance feedback in both models using a variation of

*Email: arnaud.fietzke@gmx.net
†Email: dividino@stud-cs.saarland.de

Rocchio's similarity-based relevance feedback algorithm. The main goal of the project is to evaluate the results of relevance feedback in these two models.

2.1 Improvements in the document collection

Increase the current document collection to comprise about 1000 to 1500 documents.
The documents are short news articles from cnn.com
cURL (curl.haxx.se) and *Perl* will be used for download and text extraction.
Estimated time: 1 day

2.2 Improvements in the Vector Space Model Algorithm

Implement vector space model algorithms in MatLab.
Estimated time: 1 day

2.3 Improvements in the LSI Algorithm

Implement LSI algorithms in MatLab.
Estimated time: 1 day

2.4 Implementation of Rocchio's Similarity-Based Relevance Feedback Algorithm

Implement relevance feedback algorithm in MatLab.
Estimated time: 1 day

2.5 Relevance Feedback with Basic Vector Comparison

Apply relevance feedback with basic vector comparison;
analyse effect of varying different parameters of the algorithm.
Estimated time: 1-2 days

2.6 Relevance Feedback with LSI

Apply relevance feedback with LSI;
analyse effect of varying different parameters of the algorithm.
Estimated time: 1-2 days

2.7 Results Analysis

Analyse results of basic vector comparison and LSI with and without relevance feedback.
Estimated time: 1-2 days

2.8 Report Results

Make some pretty diagrams and present results in a 1-page report.
Estimated time: 1 day

2.9 Deliverables

- 1-page report
- presentation
- source code of implemented algorithms

References

[1] Zhixiang Chen and Binhai Zhu. Some Formal Analysis of Rocchio's Similarity-Based Relevance Feedback Algorithm. In International Symposium on Algorithms and Computation, pages 108–119, 2000.

[2] J. J. Jr. Rocchio. Relevance feedback in information retrieval. In Prentice-Hall, In The Smart System — Experiments in Automatic Document Processing, pages 313–323, 1971.