

Lecture 7: Relevance Feedback and Query Expansion

Information Retrieval
Computer Science Tripos Part II

Ronan Cummins

Natural Language and Information Processing (NLIP) Group




**UNIVERSITY OF
CAMBRIDGE**

`ronan.cummins@cl.cam.ac.uk`

2017

- 1 Introduction
- 2 Relevance Feedback
 - Rocchio Algorithm
 - Relevance-Based Language Models
- 3 Query Expansion

- The same word can have different meanings (polysemy)
- Two different words can have the same meaning (synonymy)
- Vocabulary of searcher may not match that of the documents
- Consider the query = {*plane fuel*}
- While this is relatively unambiguous (wrt the meaning of each word in context), exact matching will miss documents containing *aircraft*, *airplane*, or *jet*
- Relevance feedback and query expansion aim to overcome the problem of *synonymy*

plane fuel 

[All](#) [Images](#) [News](#) [Shopping](#) [Videos](#) [More](#) [Settings](#) [Tools](#)

About 60,400,000 results (0.79 seconds)

Aviation fuel - Wikipedia

https://en.wikipedia.org/wiki/Aviation_fuel ▼

Jump to **Jet fuel** - Jet fuel is a clear to straw-colored fuel, based on either an unleaded kerosene (Jet A-1), or a naphtha-kerosene blend (Jet B). It is similar ...

[Types of aviation fuel](#) - [Production of aviation fuel](#) - [Energy content](#)

Jet fuel - Wikipedia

https://en.wikipedia.org/wiki/Jet_fuel ▼

Jet fuel, aviation turbine fuel (ATF), or avtur, is a type of aviation fuel designed for use in aircraft powered by gas-turbine engines. It is colorless to straw-colored ...

Density: 775.0-840.0 g/L

Melting point: -47 °C (-53 °F; 226 K)

Boiling point: 176 °C (349 °F; 449 K)

Flash point: 38 °C (100 °F; 311 K)

People also ask

Is jet fuel kerosene? ▼

Why kerosene is used as a jet fuel? ▼

Which fuel is used in airplanes? ▼

What is the octane level of jet fuel? ▼

[Feedback](#)

What type of fuel do airplanes use? | Reference.com

<https://www.reference.com> > [Vehicles](#) > [Airplanes](#) & [Helicopters](#) ▼

Full Answer. The two types of aviation fuel are jet fuel and aviation gasoline. The most common jet fuel is made from paraffin oil and kerosene and is called JET ...

Why do aeroplanes use kerosene (parafin) as fuel? - Quora

<https://www.quora.com/Why-do-aeroplanes-use-kerosene-parafin-as-fuel> ▼

Not all aircraft use kerosene, it really depends on the engine that the aircraft has. If it is a positive displacement engine (I.E. a piston engine), kerosene is not ...

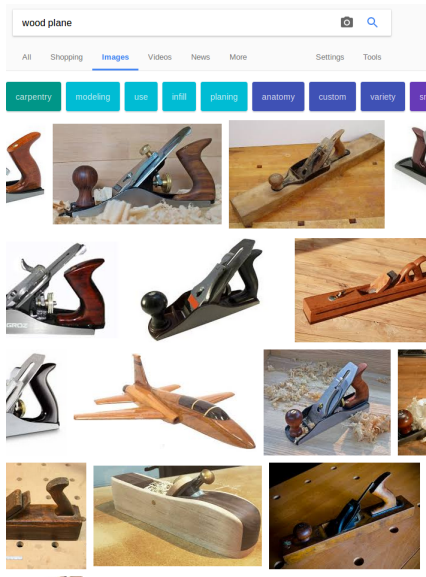
- Local analysis: query-time analysis on a portion of documents returned for a user query
 - Main local method: relevance feedback
- Global analysis: perform a global analysis once (e.g., of collection) to produce thesaurus
 - Use thesaurus for query expansion

- 1 Introduction
- 2 Relevance Feedback
 - Rocchio Algorithm
 - Relevance-Based Language Models
- 3 Query Expansion

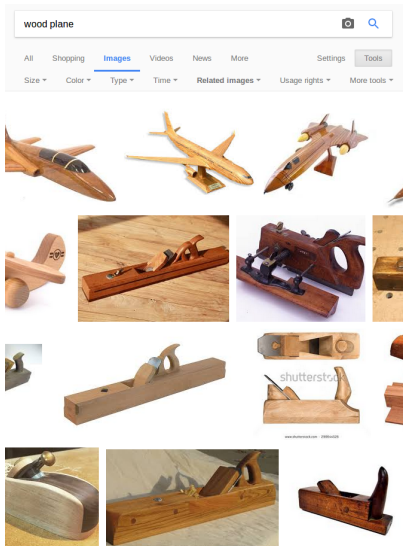
- The user issues a (short, simple) query.
- The search engine returns a set of documents.
- User marks some docs as relevant (possibly some as non-relevant).
- Search engine computes a new representation of the information need.
- Hope: better than the initial query.
- Search engine runs new query and returns new results.
- New results have (hopefully) better recall (and possibly also better precision).

A limited form of RF is often expressed as “more like this” or “find similar” .

Example



Example



- 1 Introduction
- 2 Relevance Feedback
 - Rocchio Algorithm
 - Relevance-Based Language Models
- 3 Query Expansion

- Developed in the late 60s or early 70s.
- It was developed using the VSM as its basis.
- Therefore, we represent documents as points in a high-dimensional term space.
- Uses centroids to calculate the center of a set of documents.

Rocchio aims to find the optimal query \vec{q}_{opt} that maximises:

$$\vec{q}_{opt} = \arg \max_{\vec{q}} [sim(\vec{q}, C_r) - sim(\vec{q}, C_{nr})] \quad (1)$$

where $sim(\vec{q}, C_r)$ is the similarity between a query q and the set of relevant documents C_r .

Rocchio aims to find the optimal query \vec{q}_{opt} that maximises:

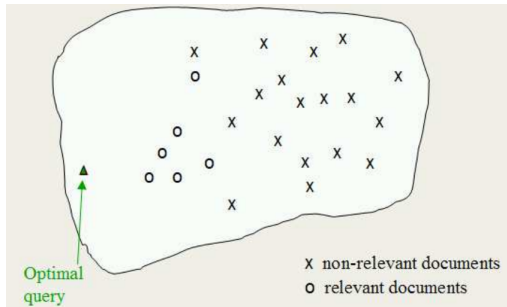
$$\vec{q}_{opt} = \arg \max_{\vec{q}} [sim(\vec{q}, C_r) - sim(\vec{q}, C_{nr})] \quad (1)$$

where $sim(\vec{q}, C_r)$ is the similarity between a query q and the set of relevant documents C_r . Using cosine similarity the optimal query becomes:

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \in C_{nr}} \vec{d}_j \quad (2)$$

which is the difference between the centroids of the relevant and non-relevant document vectors.

Rocchio Diagram



- However, we usually do not know the full relevant and non-relevant sets.
- For example, a user might only label a few documents as relevant.

Therefore, in practice Rocchio is often parameterised as follows:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|C_r|} \sum_{\vec{d}_j \in C_r} \vec{d}_j - \gamma \frac{1}{|C_{nr}|} \sum_{\vec{d}_j \in C_{nr}} \vec{d}_j \quad (3)$$

where α , β , and γ are weights that are attached to each component.

- Rocchio has been shown useful for increasing recall
- Contains aspects of positive and negative feedback
- Positive feedback is much more valuable (i.e. indications of what *is* relevant and $\gamma < \beta$)
- Reasonable values of the parameters are $\alpha = 1.0$, $\beta = 0.75$, $\gamma = 0.15$

- 1 Introduction
- 2 Relevance Feedback
 - Rocchio Algorithm
 - Relevance-Based Language Models
- 3 Query Expansion

Relevance-Based Language Models I

- The query-likelihood language model (earlier lecture) had no concept of relevance (if you remember)
- Relevance-Based language models take a probabilistic language modelling approach to modelling relevance
- The main assumption is that a document is generated from either one of two classes (i.e. relevant or non-relevant)
- Documents are then ranked according to their probability of being drawn from the relevance class

$$P(R|D) = \frac{P(D|R)P(R)}{P(D|R)P(R) + P(D|NR)P(NR)} \quad (4)$$

which is rank equivalent to ranking by log-odds

$$= \log \frac{P(D|R)}{P(D|NR)} \quad (5)$$

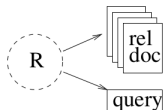


Figure 1: Queries and relevant documents are random samples from an underlying relevance model R . Note: the sampling process could be different for queries and documents.

- Lavrenko (2001) introduced the idea of relevance-based language models
- Outlined a number of different generative models
- One of the best performing models is one called RM3 (useful for both relevance and pseudo-relevance feedback)

Relevance-Based Language Models III

- Given a set of known relevant documents R one can estimate a relevance language model (e.g. multinomial θ_R)
- In practice, this can be smoothed with the original query model and a background model (not shown)

One could estimate the relevance model as:

$$(1 - \pi)\theta_R + \pi\theta_q \quad (6)$$

where π controls how much of the original query one wishes to retain.

Problems?

- Relevance feedback is expensive
- Relevance feedback creates long modified queries
- Long queries are expensive to process
- Users are reluctant to provide explicit feedback
- Its often hard to understand why a particular document was retrieved after applying relevance feedback

When does RF work?

- When users are willing to give feedback!
- When the user knows the terms in the collection well enough for an initial query.
- When relevant documents contain similar terms (similar to the cluster hypothesis)

The cluster hypothesis states that if there is a document from a cluster that is relevant to a search request, then it is likely that other documents from the same cluster are also relevant. - Jardine and van Rijsbergen

- How to evaluate if RF works?

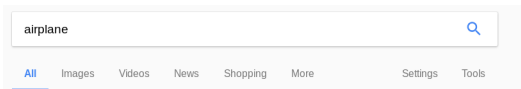
- How to evaluate if RF works?
- Have two collections, with relevance judgements for the same information needs (queries)
- User studies: time taken to find # of relevant documents (with and without feedback)

Other types of relevance feedback


- Implicit relevance feedback
- Pseudo relevance feedback - when does it work?

- 1 Introduction
- 2 Relevance Feedback
 - Rocchio Algorithm
 - Relevance-Based Language Models
- 3 Query Expansion

Query Expansion Motivation



Query Expansion Motivation

 
[All](#) [Images](#) [Videos](#) [News](#) [Shopping](#) [More](#) [Settings](#) [Tools](#)

Searches related to airplane

airplane **or** **aeroplane**

airplane **cartoon**

airplane **1980 cast**

airplane **2**

airplane **ticket booking**

airplane **definition**

airplane **lyrics**

airplane **full movie**

Query Expansion Introduction

- Query expansion is another method for increasing recall
- We use “global query expansion” to refer to “global methods for query reformulation”
- In global query expansion, the query is modified based on some global resource, i.e. a resource that is not query-dependent
- Often the problem aims to find (near-)synonyms
- Distributional Semantics (word embeddings)
- What’s the different between “local” and “global” methods?

- Use of a controlled vocabulary that is maintained by human editors (e.g. sets of keywords for publications - Medline)
- A manual thesaurus (e.g. wordnet)
- **An automatically derived thesaurus**
- Query reformulations based on query log mining (i.e. what the large search engines do)

- Let A be a term-document matrix
- Where each cell A_{td} is a weighted count of term t in document (or window) d
- Row normalise the matrix (e.g. L2 normalisation)
- Then $C = AA^T$ is a term-term similarity matrix
- The similarity between any two terms u and v is in C_{uv}
- Given any particular term q , the most similar terms can be easily retrieved

- Other approaches involve distributional semantics
- Where words with similar meanings appear in similar contexts
- Word embeddings - word2vec, glove, etc
- Can be useful but global expansion still suffers from problems of polysemy
- A naive approach to word-level expansion might lead to {apple computer} → {apple fruit computer}

- QE is transparent in that it allows the user to see (select) expansion terms
- Local approaches (PRF) to expanding queries tend to be more effective
- E.g. {apple computer} → {apple computer jobs iphone ipad macintosh}
- Local approaches tend to automatically disambiguate the individual query terms. Why?
- Query log mining approaches have also been shown to be useful

- Manning, Raghavan, Schütze: Introduction to Information Retrieval (MRS), chapter 9: Relevance feedback and query expansion, chapter 16.1: Clustering in information retrieval
- Victor Lavrenko and W. Bruce Croft: Relevance-Based Language Models