Lecture 7: Relevance Feedback and Query Expansion

Information Retrieval
Computer Science Tripos Part II

Helen Yannakoudakis[1]

Natural Language and Information Processing (NLIP) Group
UNIVERSITY OF CAMBRIDGE

helen.yannakoudakis@cl.cam.ac.uk

2018

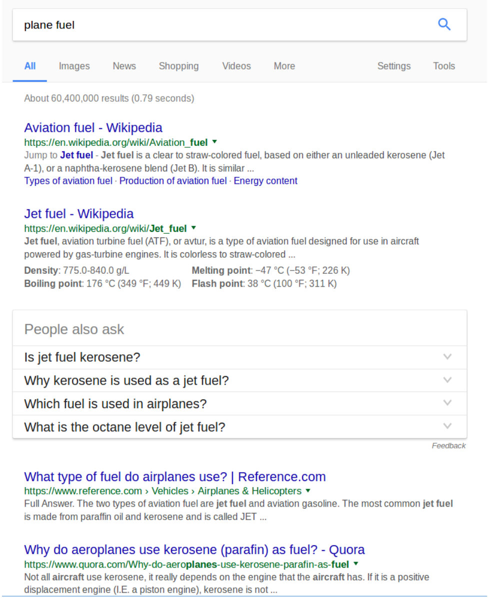
[1]Based on slides from Ronan Cummins

Overview

Motivation

- 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* $\rightarrow$ impacts recall.
- Relevance feedback and query expansion aim to overcome the problem of *synonymy*.

Example



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 ...

Improving Recall

Methods for tackling this problem split into two classes:

Improving Recall

Methods for tackling this problem split into two classes:

- Local methods: adjust a query relative to the documents returned (query-time analysis on a portion of documents)
  - Main local method: relevance feedback
- Global methods: adjust query based on some global resource / thesaurus (i.e., a resource that is not query dependent)
  - Use thesaurus for query expansion

Overview

1. Introduction
2. Relevance Feedback (RF)
   - Rocchio Algorithm
   - Relevance-based Language Models
3. Query Expansion

Relevance Feedback: The Basics

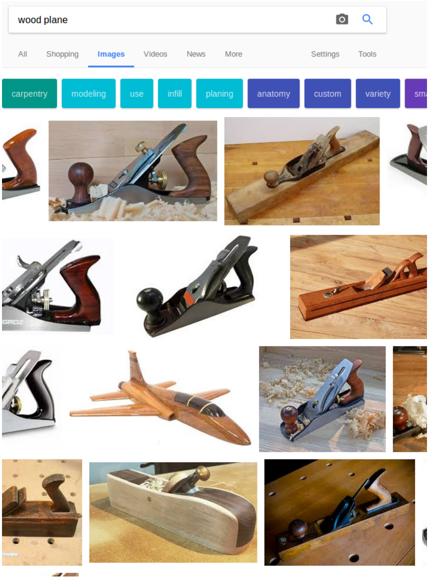
Main idea: involve the user in the retrieval process so as to improve the final result.

Relevance Feedback: The Basics

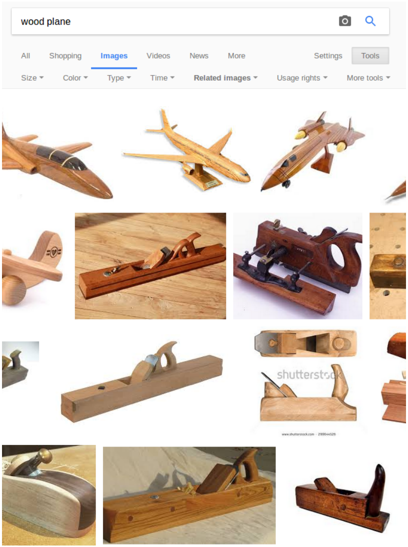
Main idea: involve the user in the retrieval process so as to improve the final result.

- 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).
  - Can have graded relevance feedback, e.g., "somewhat relevant", "relevant", "very relevant".
- Search engine computes a new representation of the information need based on feedback from the user.
  - 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).

Example


wood plane
All
Shopping
Images
Videos
News
More
Settings
Tools
carpentry
modeling
use
infill
planing
anatomy
custom
variety

Example


wood plane
All
Shopping
Images
Videos
News
More
Settings
Tools
Size
Color
Type
Time
Related images
Usage rights
More tools
shutterstock

Outline

1. Introduction
2. Relevance Feedback (RF)
   - Rocchio Algorithm
   - Relevance-based Language Models
3. Query Expansion

Rocchio algorithm: Basics

- Classic algorithm for implementing relevance feedback.
- It was developed using the Vector Space Model as its basis.
- Incorporates relevance feedback information into the VSM.
- Therefore, we represent documents as points in a high-dimensional term space.
- Uses centroids to calculate the center of a set of documents C: $\frac{1}{|C|} \sum_{\vec{d} \in C} \vec{d}$

Rocchio

Aims to find the query $\vec{q}$ that maximises similarity with the set of relevant documents C_r while minimising similarity with the set of non relevant documents C_{nr}:

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

Rocchio

Aims to find the query $\vec{q}$ that maximises similarity with the set of relevant documents C_r while minimising similarity with the set of non relevant documents C_{nr}:

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

Under cosine similarity, the optimal query for separating relevant and non relevant documents is:

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

which is the vector difference between the centroids of the relevant and non relevant documents.

Rocchio in practice

- In practice, however, we usually do not know the full set of relevant and non relevant sets.
- For example, a user might only label a few documents as relevant / non relevant.

Rocchio in practice

- In practice, however, we usually do not know the full set of relevant and non relevant sets.
- For example, a user might only label a few documents as relevant / non relevant.

Therefore, in practice Rocchio is often parameterised as follows:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

where $\vec{q}_0$ is the original query vector; D_r and D_{nr} are the sets of known relevant and non relevant documents.

- α, β, and γ are weight parameters attached to each component.
- Reasonable values are $\alpha = 1.0$, $\beta = 0.75$, $\gamma = 0.15$

Rocchio in practice

- In practice, however, we usually do not know the full set of relevant and non relevant sets.
- For example, a user might only label a few documents as relevant / non relevant.

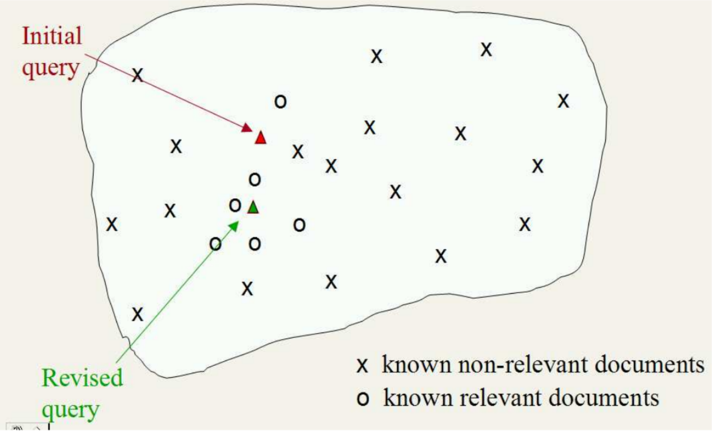
Therefore, in practice Rocchio is often parameterised as follows:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

where $\vec{q}_0$ is the original query vector; D_r and D_{nr} are the sets of known relevant and non relevant documents.

- α, β, and γ are weight parameters attached to each component.
- Reasonable values are $\alpha = 1.0$, $\beta = 0.75$, $\gamma = 0.15$
- Note: if final $\vec{q}_m$ has negative term weights, set to 0.

Example application of Rocchio



Rocchio in practice

- Represent query and documents as weighted vectors (e.g., tf–idf).
- Use Rocchio formula to compute new query vector (given some known relevant / non-relevant documents).
- Calculate cosine similarity between new query vector and documents.
- (E.g., supervision exercises 9.5 and 9.6 from the book).

Rocchio in practice

- Represent query and documents as weighted vectors (e.g., tf–idf).
- Use Rocchio formula to compute new query vector (given some known relevant / non-relevant documents).
- Calculate cosine similarity between new query vector and documents.
- (E.g., supervision exercises 9.5 and 9.6 from the book).

- Rocchio has been shown useful for increasing recall.
- Contains aspects of positive and negative feedback.
- Positive feedback is much more valuable than negative (i.e., indications of what *is* relevant)
- Most systems set $\gamma < \beta$ or even $\gamma = 0$.

Outline

1. Introduction
2. Relevance Feedback (RF)
   - 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.
- Relevance-based language models take a probabilistic language modelling approach to modelling relevance.

Relevance-based Language Models I

- The query-likelihood language model (earlier lecture) had no concept of relevance.
- 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:

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

Relevance-based Language Models I

- The query-likelihood language model (earlier lecture) had no concept of relevance.
- 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:

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

which is equivalent to ranking the documents by the (log) odds of their being observed in the relevant class:

$$= \frac{P(D|R)}{P(D|NR)} \sim \prod_{t \in D} \frac{P(t|R)}{P(t|NR)}$$

Relevance-Based Language Models II

$$\frac{P(D|R)}{P(D|NR)} \sim \prod_{t \in D} \frac{P(t|R)}{P(t|NR)}$$

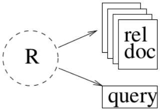


Figure 1: Queries and relevant documents are random samples from an underlying relevance model R. Note: the sampling process could be different for queries and documents.

- Lavrenko (2001) introduced the idea of relevance-based language models.
- Outlined a number of different generative models.

Relevance-Based Language Models II

$$\frac{P(D|R)}{P(D|NR)} \sim \prod_{t \in D} \frac{P(t|R)}{P(t|NR)}$$

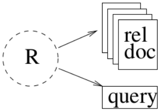


Figure 1: Queries and relevant documents are random samples from an underlying relevance model R. Note: the sampling process could be different for queries and documents.

- Lavrenko (2001) introduced the idea of relevance-based language models.
- Outlined a number of different generative models.
- $P(t|NR)$ estimated using document collection as most documents are non relevant.

Relevance-Based Language Models II

$$\frac{P(D|R)}{P(D|NR)} \sim \prod_{t \in D} \frac{P(t|R)}{P(t|NR)}$$

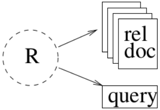


Figure 1: Queries and relevant documents are random samples from an underlying relevance model R. Note: the sampling process could be different for queries and documents.

- Lavrenko (2001) introduced the idea of relevance-based language models.
- Outlined a number of different generative models.
- $P(t|NR)$ estimated using document collection as most documents are non relevant.
- Assume that both the query and the documents are samples from an unknown relevance model R which gives $P(t|R)$.

Relevance-Based Language Models II

$$\frac{P(D|R)}{P(D|NR)} \sim \prod_{t \in D} \frac{P(t|R)}{P(t|NR)}$$

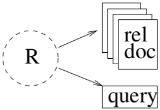


Figure 1: Queries and relevant documents are random samples from an underlying relevance model R. Note: the sampling process could be different for queries and documents.

- Lavrenko (2001) introduced the idea of relevance-based language models.
- Outlined a number of different generative models.
- $P(t|NR)$ estimated using document collection as most documents are non relevant.
- Assume that both the query and the documents are samples from an unknown relevance model R which gives $P(t|R)$.
- The query is the only sample we have from this unknown distribution.

Relevance-Based Language Models II

$$\frac{P(D|R)}{P(D|NR)} \sim \prod_{t \in D} \frac{P(t|R)}{P(t|NR)}$$

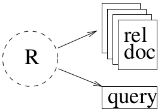


Figure 1: Queries and relevant documents are random samples from an underlying relevance model R. Note: the sampling process could be different for queries and documents.

- Lavrenko (2001) introduced the idea of relevance-based language models.
- Outlined a number of different generative models.
- $P(t|NR)$ estimated using document collection as most documents are non relevant.
- Assume that both the query and the documents are samples from an unknown relevance model R which gives $P(t|R)$.
- The query is the only sample we have from this unknown distribution.
- One of the best performing models is one called RM3 (useful for both relevance and pseudo-relevance feedback).

- 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):

$$(1-\lambda)P(t|\theta_R) + \lambda P(t|\theta_q)$$

Problems?

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

When does Relevance Feedback work?

- When users are willing to give feedback!
- When the user has sufficient initial knowledge, and 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

Relevance Feedback: Evaluation

How to evaluate if Relevance Feedback 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

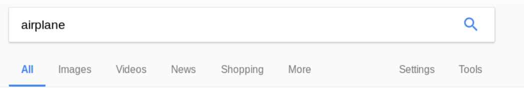
- Implicit relevance feedback
- Pseudo relevance feedback

Overview

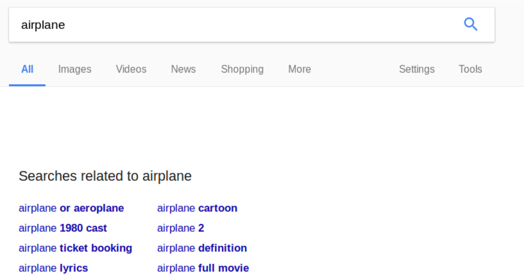
Query Expansion Introduction

- 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.
- What's the different between "local" and "global" methods?

Query Expansion: Example 1



Query Expansion: Example 1



Query Expansion: Example 2

Yahoo! My Yahoo! Mail Welcome, **Guest** [Sign In] Help

Web | Images | Video | Local | Shopping | more

palm **Search** Options

1 - 10 of about 534,000,000 for **palm** (About this page) - 0.11 sec.

Also try: palm trees, **palm** springs, **palm** centro, **palm** treo, More...

SPONSOR RESULTS

Palm - AT&T
att.com/wireless - Go mobile effortlessly with the **PALM** Treo from AT&T (Cingular).

Palm Handhelds
Palm.com - Organizer, Planner, WiFi, Music Bluetooth, Games, Photos & Video.

Palm, Inc.
Maker of handheld PDA devices that allow mobile users to manage schedules, contacts, and other personal and business information.
www.**palm.com** - Cached

Palm, Inc. - Treo and Centro smartphones, handhelds, and accessories
Palm, Inc., innovator of easy-to-use mobile products including **Palm**® Treo and Centro smartphones, **Palm** handhelds, services,

SPONSOR RESULTS

Handhelds at Dell
Stay Connected with Handheld PCs & PDAs. Shop at Dell™ Official Site.
www.**Dell.com**

Buy **Palm** Centro Cases
Ultimate selection of cases and accessories for business devices.
www.**Cases.com**

Free Plam Treo
Get A Free **Palm** Treo 700W Phone. Participate Today.

Query Expansion

- In relevance feedback, users give input on documents (are they relevant or not), which is used to refine the query.
- In query expansion, users give input on query terms or phrases.

Query Expansion 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).

Automatic thesaurus generation I

Hypothesis: words co-occurring in a document (or paragraph) are likely to be in some sense similar or related in meaning.

Automatic thesaurus generation I

Hypothesis: words co-occurring in a document (or paragraph) are likely to be in some sense similar or related in meaning.

- Let A be a term–document matrix.
- Where each cell $A_{t,d}$ is a weighted count of term t in document (or context window) d.
- Row normalise the matrix (e.g., L2 normalisation).
- Then $C = AA^T$ is a term–term similarity matrix.
  - Typically combined with an extra step of dimensionality reduction (e.g., Latent Semantic Indexing).
- The similarity between any two terms u and v is in $C_{u,v}$.
- Given any particular query term q, the most similar terms can be easily retrieved.

Automatic thesaurus generation II

- Distributional hypothesis: words with similar meanings appear in similar contexts (e.g., *car* and *motorbike*).
- Word embeddings – word2vec, glove, etc.

Summary

- Query Expansion is transparent in that it allows the user to see (select) expansion terms.
- Can be useful but global expansion still suffers from problems of polysemy.
- A naive approach to word-level expansion might lead to {apple computer} $\rightarrow$ {apple fruit computer}
- Local approaches to expanding queries tend to be more effective.
- E.g., {apple computer} $\rightarrow$ {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.

Reading

- 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