Information Retrieval (5LN712)

Relevance Feedback and Query Expansion

Ali Basirat

Department of Linguistics and Philology
Uppsala University

April 22, 2020

Table of contents

Introduction

- The users often interact with an IR system via queries
- Queries might not be good enough
- Users may not know enough how their information need is expressed in the document collection
- How an IR system can address these issues

Introduction

Synonymy Issue

- Synonymy: the same concepts are referred to by different terms (e.g., aircraft and airplane)
- The synonymy issue can affect the recall of IR systems
- Many documents that address the same concept but with different terms might be skipped
- Users themselves address this problem

Introduction

Synonymy Issue

The major methods to tackle the synonymy issue:

- Global methods: to expand and reformulate query terms independent of the query and results returned from it
- Local methods: to adjust a query with respect to the initial documents that match the query

Table of contents

Global Methods

- To aid the user in expanding the query
- To use a manually created thesaurus
- To use an automatically created thesaurus

Global Methods

Query Reformulation

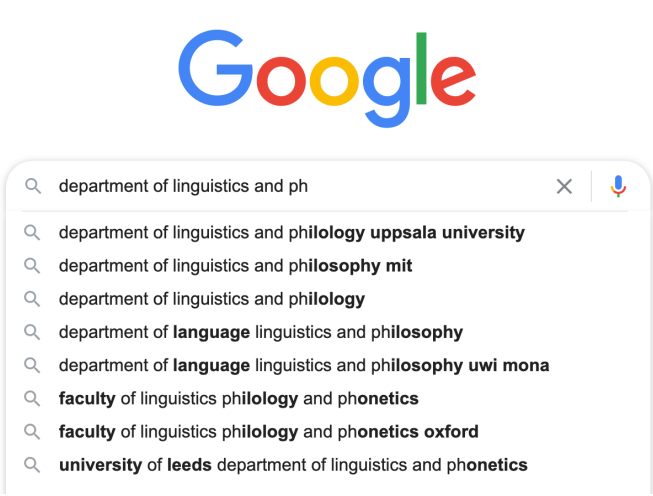
User supports to reformulate queries based on:

- information about omitted words
- suggestion about query terms
- list of the terms that are in the inverted index

Table of contents

Query Expansion

- Query expansion leads to a higher recall
- Queries can be reformulated based on the similar queries made by other users
- In web search, a search engine suggests related queries in response to a query
- The query expansion is done for each term of a query
- Each term can be expanded with synonyms and related words collected from a thesaurus



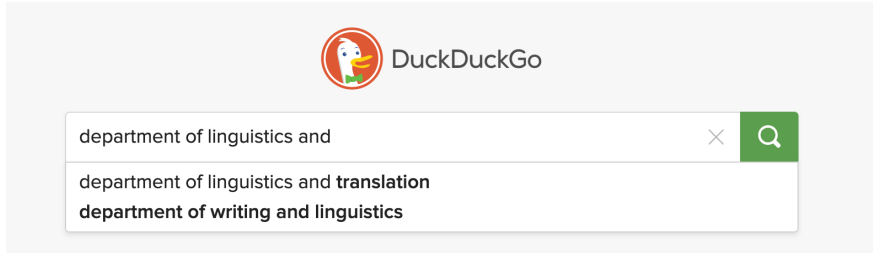


Table of contents

Thesaurus

Creation Approaches

Thesaurus is a list of words with their synonyms

- a controlled vocabulary in which concepts are associated with canonical terms are manually created
- a list of synonymous names for concepts are manually created
- a list of synonymous names ares automatically created based on the word co-occurrence statistics

Thesaurus

Automatic Generation

Two main approaches:

- to exploit word co-occurrence in a document or paragraph
- to exploit grammatical relations or dependencies between words

Thesaurus

Automatic Generation

To exploit word co-occurrences

- Start from a term-document matrix A
- Compute the similarity matrix $C = AA^T$
- Use C to find similar terms (words)

Global Methods

Advantages and Disadvantages

- Effective at increasing recall
- Costly to manually create thesauri and dictionaries
- General thesauri and dictionaries are not effective
- Domain specific thesauri are needed for scientific searches

Table of contents

Local Methods

- Relevance feedback
- Rocchio Algorithm
- Probabilistic Methods
- Indirect relevance feedback
- Pseudo relevance feedback

Relevance Feedback

- Relevance feedback (RF): to involve the user in the IR process
- Users give feedback on the relevance of initially returned documents

Relevance Feedback

Idea

- Users might not be able to formulate a good query when they do not know the collection well
- It is easy to judge particular documents
- It makes sense to iterativelly refine the query

Relevance Feedback

Procedure

1. The user issue a short and simple query
2. The system returns an initial set of retrieval results
3. The user marks some returned documents as relevant or non-relevant
4. The system computes a better representation of the information need based on the user feedback
5. The system displays a revised set of retrieval results
6. Go to the step 3 for a limited number of iterations

Relevance Feedback

Advantage

- RF is effective in tracking a user's evolving information need
- Seeing some documents may lead users to refine their understanding of the information they are seeking
- Interactive IR substantially improves the performance of an IR system

Example

Image search

- It is not easy to describe an image by words
- The user starts with a easy and simple description of the image
- The system returns some images
- It is easy for the user to find the relevant or non-relevant images

Table of contents

The underlying theory:

- To incorporate RF information into the vector space model
- To find a query vector $\vec{q}$ that maximizes the similarity with relevant documents while minimizing similarity with non-relevant documents
- Let C_r and C_{nr} be the sets of relevant and non-relevant documents to an information need, we want to find:

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

Relevance Feedback

Rocchio Algorithm - Theory

- What does the similarity between a vector and a set of vectors mean?
- One interpretation is the similarity between the vector and the centroid of the set of vector

$$\text{sim}(\vec{q}, C) = \frac{1}{|C|} \sum_{d_j \in C} \text{sim}(\vec{q}, \vec{d_j})$$

- If the cosine similarity is used

$$\text{sim}(\vec{q}, C) = \frac{\vec{q}}{|C|} \sum_{d_j \in C} \vec{d_j}$$

Relevance Feedback

Rocchio Algorithm - Theory

- Using the centroid vectors and the cosine similarity, the optimal query vector for separating the relevant and non-relevant documents is

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

- The optimal vector is the vector difference between the centroids of the relevant and non-relevant documents
- This is not a practical solution since we do not know the full set of relevant documents

Relevance Feedback

Rocchio Algorithm

- We have a user query and only partial knowledge of relevant and non-relevant documents
- Let D_r and D_{nr} be the sets of known relevant and non-relevant documents
- Let q_0 be the original query vector
- The modified query q_m is:

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

- α, β, and γ are real-valued weights that control the balance between trusting the judged document sets versus the query

Relevance Feedback

Rocchio Algorithm

- RF can improve both precision and recall
- The Rocchio algorithm moves the query vector toward the centroid of the relevant documents and far from the non-relevant documents
- Thus, it is most useful for increasing recall
- Reasonable values for the weights are: $\alpha = 1, \beta = 0.75$, and $\gamma = 0.15$
- Modern IR system account for only relevant documents - they set $\gamma = 0$

Table of contents

Relevance Feedback

Probabilistic Method

- To build a classifier to identify the relevant and non-relevant documents
- The classification is based on the judged documents

Relevance Feedback

Probabilistic Method

- A näive Bayesian probabilistic model
- Let R be an Bernoulli variable indicating the relevance of a document
- Let VR be the set of known relevant documents, and $\text{VR}_t \subset \text{VR}$ be the set of relevant documents containing the term t
- We can estimate the probability of a term appearing in a relevant or non-relevant document

$$\hat{P}(x_t = 1|R = 1) = \frac{|\text{VR}_t|}{|\text{VR}|}$$

$$\hat{P}(x_t = 1|R = 0) = \frac{\text{df}_t - |\text{VR}_t|}{N - |\text{VR}|}$$

Relevance Feedback

Probabilistic Methods

- The term probabilities can be used as the weights of query terms
- But, they are not sufficient because the probabilities are estimated only based on the documents that are judged relevant
- we need a way to make use of the original query as well
- This method will be discussed further in the probabilistic methods of IR

Relevance Feedback

on the Web

- Some search engines offer a similar/related pages
- Web users concern more about precision of the system
- RF is more useful to increase the recall

Relevance Feedback

on the Web

- Web search engines use click stream data - statistics about clicks on links
- It is a way of indirect RF

Relevance Feedback

Keys of Success

Two key assumptions on which a successful RF depends

- The initial query made by the user should be informative - the query vector should be close to the desired document vector
- The relevant documents should be clustered together

Relevance Feedback

Insufficiency

RF is not sufficient in the following cases

- Misspellings: if query terms are spelled differently than the spellings of the collection documents
- Cross-language IR: documents in different languages are not clustered together even if they have similar contents

Table of contents

Relevance Feedback

Evaluation

The general approach:

- Start with q_0 and compute a precision-recall curve.
- Re-compute the curve based on q_m after receiving feed-back from user

Relevance Feedback

Evaluation Strategies

1. Use all documents in the collection for both q_0 and q_m
2. Use all documents for evaluating q_0, and all minus those assessed by the user as relevant for q_m
3. Use two distinct collections for evaluating each of q_0 and q_m
4. Perform a user study

Relevance Feedback

Evaluation Strategies

If we use all collections for evaluating both q_0 and q_m

- We gain a lot from RF
- The known relevant documents (judged by the user) are ranked higher in the second round
- We see a significant performance boost from q_0 to q_m
- This is cheating

Relevance Feedback

Evaluation Strategies

If use documents in the residual collection for evaluating q_m

- We eliminate the relevant documents judged by the user from the second round evaluation
- Hence, the result of q_0 is expected to be better than the result of q_m specially if the set of relevant documents is small

Relevance Feedback

Evaluation Strategies

If we use two distinct sets for evaluating q_0 and q_m

- Mitigates the shortcomings of the two other approaches
- Provide for measuring the effectiveness of RF

Relevance Feedback

Evaluation Strategy

To perform a user study

- Seems to be the best strategy to evaluate the utility of RF
- Measure the effectiveness of RF using time-based comparisons
- How fast the user find relevant documents with RF
- How many relevant documents does a user find in a certain amount of time

Table of contents

Indirect Relevance Feedback

Implicit Feedback

- To use indirect sources of evidence rather than explicit feedback on relevance
- For example, the number of clicks on links is considered as a relevance indicator
- If a documents is seen more others then the documents is more likely to be relevant

Table of contents

Pseudo Relevance Feedback

Blind Relevance Feedback

- To automate the manual part of RF
- It contains no evidence of user judgments
- The user does not do any judgement interaction with the system
- We assume that the top k ranked documents returned for the user query are relevant

Pseudo Relevance Feedback

Outline

- The user submits a query q_0
- The system returns a ranked list of documents
- The top k documents of the returned list are used to create the modified query q_m

Table of contents

Summary

- Relevance feedback effectively improve the recall of an IR system
- It is a way to mitigate the lack of knowledge of users about the collection
- RF takes the user into the game of IR to bridge between the user's mind and the content of the collection
- The global methods of RF rely on the statistics about the collection, but the local methods takes the need of users into account
- Rocchio algorithm is a way to modify the user query to make it more compatible with the content of the collection
- How to evaluate an IR system with RF
- RF can be very costly and sometimes boring from the user's point of view
- Two ways toward automating RF: indirect and pseudo RF

References I

Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich.
Introduction to Information Retrieval.
Cambridge University Press, 2008.