

Query Expansion

Dipasree Pal

Indian Statistical Institute, Kolkata

December 22, 2010

Overview

- ▶ Why is it necessary
- ▶ Some Query Expansion techniques

Query Expansion

Science of searching for information in response to a query.

- ▶ How is it done?
Keyword matching between a query and a document
 - ▶ Information need is expressed using some keywords
 - ▶ Documents containing those keywords are retrieved

Example: President Obama visits India

Query Expansion: Motivation

- ▶ Vocabulary mismatch problem
 - ▶ Useful documents may not contain keywords given by the user
Example: *Economic impact* of recycling tyres.
- ▶ Solution: Query Expansion
Add terms to query in order to
 - ▶ Match more documents
 - ▶ Improve match with relevant documents
Example: *Economic impact* of recycling tyres.
+ (rubber, sale, revenue, savings, cost, etc.)

Query Expansion: Different types

Sources of expansion terms:

- ▶ Retrieved documents
 - ▶ Relevance feedback
 - ▶ Blind feedback
- ▶ Target corpus
 - ▶ Phrase finder
- ▶ External resources
 - ▶ Other document repositories
 - ▶ Wikipedia etc.
 - ▶ Web
 - ▶ Query logs
 - ▶ Linguistic resources
 - ▶ Ontology, WordNet etc.

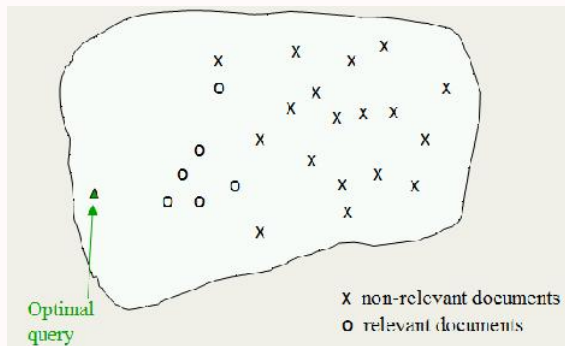
Sources of expansion terms: Retrieved documents

Users are not good at forming detailed/long queries.

- ▶ Relevance feedback
 - ▶ Original short query is used to retrieve a few docs.
 - ▶ User examines some of the retrieved documents and provides feedback about which documents are relevant and which are non-relevant
 - ▶ System uses feedback to “learn” better query:
 - ▶ select/emphasize words that occur more frequently in relevant documents than non-relevant documents
 - ▶ eliminate/de-emphasize words that occur more frequently in non-relevant than in relevant documents
 - ▶ Resulting query should bring in more relevant documents and fewer non-relevant documents

Rocchio's algorithm

- ▶ Maximize the similarity with relevant documents
- ▶ Minimize the similarity with non relevant documents



Rocchio's algorithm

- ▶ Maximize the difference between average similarities for relevant and non-relevant documents

$$\begin{aligned} C &= \frac{1}{N_{rel}} \sum_{D_i \in Rel} Sim(Q, D_i) - \frac{1}{N_{nonrel}} \sum_{D_i \in NRel} Sim(Q, D_i) \\ &= \vec{Q} \cdot \left[\frac{1}{N_{rel}} \sum_{D_i \in Rel} \vec{D}_i - \frac{1}{N_{nonrel}} \sum_{D_i \in NRel} \vec{D}_i \right] \end{aligned}$$

- ▶ In practice:

$$\vec{Q}_{new} = \alpha \vec{Q}_{old} + \frac{\beta}{N'_{rel}} \sum_{D_i \in Rel} \vec{D}_i - \frac{\gamma}{N'_{nonrel}} \sum_{D_i \in NRel} \vec{D}_i$$

Sources of expansion terms: Retrieved documents

Users want good results straightaway (without interaction)

- ▶ Blind feedback: In the absence of feedback, *assume* top-ranked documents are relevant

TREC7 (Carpineto et al, TOIS, 2002)

	No Fdbk.	Rocchio
Rel-Ret	2751	3009 +9.38%
MAP	0.2291	0.2625 +14.54%

Blind feedback

- ▶ Obvious danger: query drift
If initial retrieval is poor, adhoc feedback can aggravate the problem
Example: terrier
- ▶ Suggestion: find ways to improve the initial retrieval


Query drift

Web [Images](#) [Videos](#) [Maps](#) [News](#) [Shopping](#) [Gmail](#) [more](#) ▼

Google

terrier

About 19,600,000 results (0.24 seconds)

 Everything

 Images

 Videos

 News

 More

Any time

Latest

Past 2 days

All results

Sites with images

More search tools

Something different

[schnauzer](#)

[dachshund](#)

[jack russell](#)

[Images for terrier](#) - [Report images](#)



[Terrier](#) - [Wikipedia, the free encyclopedia](#)

A **terrier** is a dog of any one of many breeds or landraces of **terrier** type, which are typically small, wiry, very active and fearless dogs. ...

[Terriers \(TV series\)](#) - [List of dog types](#) - [Category](#) - [Yorkshire Terrier](#)
en.wikipedia.org/wiki/Terrier - [Cached](#) - [Similar](#)

[AKC Breeds by Group - Terrier Group](#)

People familiar with this Group invariably comment on the distinctive **terrier** personality.

Query drift

Web Images Videos Maps News Shopping Gmail more ▾



terrier ir

About 872,000 results (0.19 seconds)

 Everything

 Images

 Videos

 More

Show search tools

[Terrier IR Platform - Homepage](#)

Welcome to the **Terrier IR** Platform. Terrier is a highly flexible, efficient, and effective open source search engine, readily deployable on large-scale ...

[terrier.org/](#) - Cached

Terrier 3.0 Forum
Download Terrier 2.2.1

[More results from terrier.org »](#)

[Terrier IR Platform - Download](#)

The latest stable version of **Terrier** is 3.0, released 10/03/2010. Released ...

[terrier.org/download/](#) - Cached

[Terrier IR Platform - People](#)

The following people have been involved in researching, developing or ...

[terrier.org/people.html](#) - Cached

[Documentation - The Terrier Project](#)

Overview, An overview of what the **Terrier** platform is, and what it can be ...

[terrier.org/docs/v2.2.1/](#) - Cached

 [Show more results from terrier.org](#)

[\[Wing\] Terrier IR Platform](#)

18 Jun 2010 ... On Fri, Jun 18, 2010 at 6:00 PM, Markus HAENSE <mhaense at gmail.com>

Sources of expansion terms: Target corpus

- ▶ Automatic thesaurus built from the target corpus

Sources of expansion terms: Target corpus

- ▶ Automatic thesaurus built from the target corpus
- ▶ Term-term relationship

Sources of expansion terms: Target corpus

- ▶ Automatic thesaurus built from the target corpus
- ▶ Term-term relationship
- ▶ Term co-occurrence information

Sources of expansion terms: Target corpus

- ▶ Automatic thesaurus built from the target corpus
- ▶ Term-term relationship
- ▶ Term co-occurrence information
 - ▶ Association: if two terms co-occur within the same document, they constitute an association

Sources of expansion terms: Target corpus

- ▶ Automatic thesaurus built from the target corpus
- ▶ Term-term relationship
- ▶ Term co-occurrence information
 - ▶ Association: if two terms co-occur within the same document, they constitute an association
 - ▶ Gather data about term associations over the corpus: (term1, term2, assoc. freq.)

Sources of expansion terms: Target corpus

Example: **Phrasefinder**(Jing and Croft 1994)

- ▶ Each term is represented by a vector of associated terms

$$T = (\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \dots, \langle t_n, w_n \rangle)$$

⇒ term = pseudo document

Sources of expansion terms: Target corpus

Example: **Phrasefinder**(Jing and Croft 1994)

- ▶ Each term is represented by a vector of associated terms

$$T = (\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \dots, \langle t_n, w_n \rangle)$$

⇒ term = pseudo document

- ▶ Compare query to the term vectors
If $\text{Score}(Q, T)$ is high, T co-occurs with many query terms.

Sources of expansion terms: Target corpus

Example: **Phrasefinder**(Jing and Croft 1994)

- ▶ Each term is represented by a vector of associated terms

$$T = (\langle t_1, w_1 \rangle, \langle t_2, w_2 \rangle, \dots, \langle t_n, w_n \rangle)$$

⇒ term = pseudo document

- ▶ Compare query to the term vectors
If $\text{Score}(Q, T)$ is high, T co-occurs with many query terms.
- ▶ Terms co-occurring with more query terms are good expansion terms

Sources of expansion terms: External resources

- ▶ Other document repositories
 - ▶ Wikipedia etc. (Xu et al., SIGIR 2009)
- ▶ Web
 - ▶ Query logs (Cui et al, WWW2002)

Query → terms → related queries → clicked documents → related terms

- ▶ Linguistic resources
 - ▶ Ontology, WordNet etc.

Sources of expansion terms: Wordnet

Defines semantic relatedness

- ▶ Synset
Set of synonyms
- ▶ Hypernym and Hyponym
Kind of relation
Lion (Hyponym) is a kind of animal
(Hypernym)
- ▶ Meronymy and Holonymy
Part-whole relation
Branch (Meronymy) is a part of tree (Holonymy)
- ▶ Antonym
Opposite meaning

Sources of expansion terms: Wordnet

- ▶ Definition provides valuable information about the semantic meaning of a term
- ▶ The more common words the definitions of two terms have, the more similar these terms are (Banerjee and Pedersen, 2005)
- ▶ Semantic similarity based on synset definitions
- ▶ Terms which have semantic similarity with query terms are added

Summary

- ▶ There are different ways of choosing terms
 - ▶ Retrieved documents
 - ▶ Target corpus
 - ▶ External resources
- ▶ Most groups found query expansion useful on average
- ▶ Many issues still to be investigated

THANK YOU!