Data Mining and Machine Learning

Lecture 5
Query Expansion

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

- To understand how the use of semantic relationships between words can improve the performance of a text IR system
  - Query expansion
  - Generalisation
  - Synonyms, hypernyms & hyponyms
  - WordNet

Query Processing

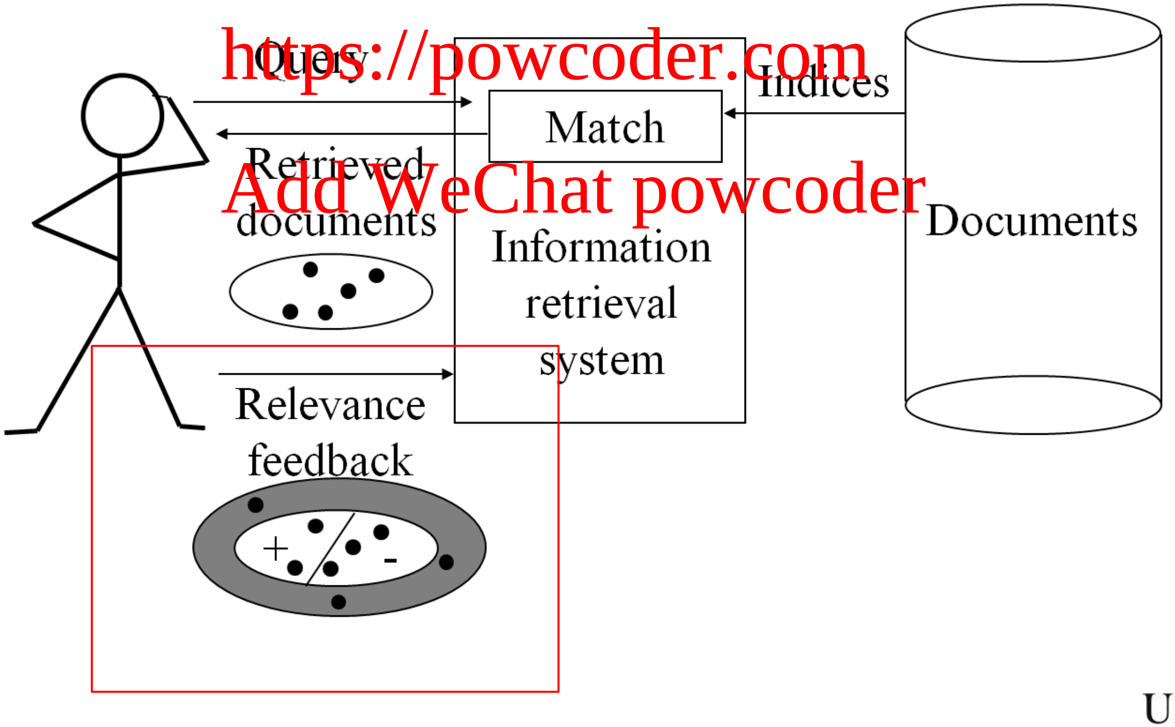
- Remember how we previously processed a query:
- Example:
  - “I need information on distance running”
- Stop word removal
  - information, distance, running
- Stemming
  - information, distance, run
- But what about:
  - “The London marathon will take place…”

Query Expansion

- Add terms to the query to increase the overlap between it and potentially relevant documents...
- ...but not irrelevant documents
- Two approaches:
  - User feedback
  - Linguistic knowledge

Feedback-based Query Expansion

- User provides feedback on the results of retrieval
  - Which of the returned documents are particularly relevant
  - Which are irrelevant



Query reformulation

- Revise the query in response to the user feedback
  - Query expansion: Find terms in the ‘relevant’ documents that are not in the query. Add them to the query (of maybe just those with large TF-IDF weights)
  - Term reweighting: Increase the weight of query terms in relevant documents and decrease the weight of query terms in irrelevant documents. For example

$$w_{td} = \boxed{\lambda} \times f_{td} \times IDF(t)$$

  - Various methods for determining λ have been proposed

Knowledge-Based Query Expansion

- Recall:
  - *q* = “I need information on distance running”
  - *d* = “The London marathon will take place…”

- We know there is a relationship between
  – run, distance, and marathon

- Words with the same meaning are <u>synonyms</u>
- If a $\boldsymbol{q}$ contains $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ is a synonym of $\boldsymbol{w_1}$, then add $\boldsymbol{w_2}$ to $\boldsymbol{q}$

Thesaurus

- A thesaurus is a 'dictionary' of synonyms and semantically related words and phrases
- E.G: Roget's Thesaurus
- Example:

physician

```
syn: || croaker, doc, doctor, MD, medical, mediciner, medico ||
rel: medic, general practitioner, surgeon
```

Peter Mark Roget 1779 –1869

- Born London 1779
- Founder of the Royal Society of Medicine
- Invented the log-log slide rule
- Professor of Physiology at the Royal Institution, 1834
- Retired 1840
- Roget’s *Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition* appeared in 1852.
- Died 1869. Buried St James’ Church, West Malvern, Worcestershire.

Hyponyms

- Not only synonyms are useful for query expansion
- Query $\boldsymbol{q}$ = “Tell me about England”
- Document $\boldsymbol{d}$ = “A visit to London should be on everyone’s itinerary”
- ‘London’ is a hyponym of ‘England’
- Hyponym ~ subordinate ~ subset
- If a query $\boldsymbol{q}$ contains a word $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ is a hyponym of $\boldsymbol{w_1}$, then $\boldsymbol{w_2}$ should be added to $\boldsymbol{q}$

Hypernyms

- Hypernyms are also useful for query expansion
- Query ***q*** = “Tell me about England”
- Document ***d*** = “Places to visit in the British Isles”
- ‘British Isles’ is a hypernym of ‘England’
- Hypernym ~ generalisation ~ superset
- If a query ***q*** contains a word $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$ is a hypernym of $\boldsymbol{w_1}$, then $\boldsymbol{w_2}$ should be added to ***q***

WordNet

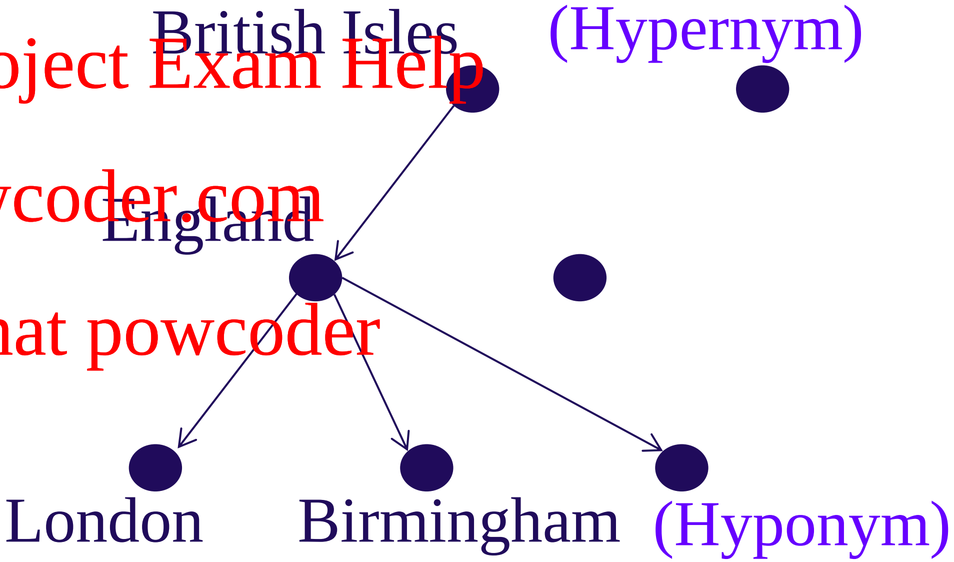
- Online lexical database for the English Language
- http://www.cogsci.princeton.edu/~wn

Category	***Forms***	***Meanings (syn sets)***
Nouns	57,000	48,800
Adjectives	19,500	10,000
Verbs	21,000	8,400

See Belew, chapter 6

WordNet

- Organised as a set of hierarchical trees
- For example, 25 trees for nouns
- ‘Children’ of a node are hyponyms
- Words become more specific as you move deeper into the tree



Noun Categories	
act, action, activity	natural object
animal, fauna	natural phenomenon
artefact	person, human being
attribute, property	plant, flora
body, corpus	possession
cognition, knowledge	process
communication	quantity, amount
event, happening	relation
feeling, emotion	shape
food	state, condition
group, collection	substance
location, place	time
motive	

Query-document scoring

- A query ***q*** is expanded to include hyponyms and synonyms
- Recall that for a document ***d***

$$w_{td} = f_{td} \cdot IDF(t)$$

$$Sim(q,d) = \frac{\sum\limits_{t \in q \cap d} w_{td} \cdot w_{tq}}{\|d\| \cdot \|q\|}$$

Query expansion

- Suppose:
  - ***t*** is the original term in the query,
  - ***t’*** is a synonym or hyponym of ***t*** which occurs in ***d***

- Then we could define:

$$w_{t'd} = \lambda_{tt'} \times f_{t'd} \times IDF(t) \qquad 0 \le \lambda_{tt'}$$

- Where $\lambda_{tt'}$ is a weighting depending on how ‘far’ ***t*** and ***t’*** are apart according to WordNet ($\lambda_{tt}=1$)

Example

- Query *q* is:
  - *Is the Dark Knight on at the town cinema?*
  - *q* becomes: *dark knight town cinema*
- Document *d* is:
  - *The latest Batman movie places the caped crusader in a dark urban environment*
  - *d* becomes: *late batman move cape crusade dark urban environment*

Example (continued)

- In the similarity calculation, $q \cap d = \{dark\}$
- But:
  - *move* and *cinema* are synonyms (compare "go to the cinema" with "go to the movies")
  - *crusader* is a hyponym of *knight*
  - *urban* is a hypernym of *town*
- Therefore, after query expansion,

$$q \cap d = \{dark, move\ (syn(cinema)), crusade(hypo(knight)), urban(hyper(town))\}$$

Example (continued)

- So, if $\lambda = 1$, $\lambda_{syn} = 0.8$, $\lambda_{hypo} = 0.5$ and $\lambda_{hyper} = 0.3$, then the numerator in the calculation of *sim*(*q*,*d*) becomes

$$w_{dark,d} * w_{dark,q}$$
$$+ \ 0.8*w_{movie,d} * w_{cinema,q}$$
$$+ \ 0.5*w_{crusader,d} * w_{knight,q}$$
$$+ \ 0.3*w_{urban,d} * w_{town,q}$$

Note: this is just a ‘made up’ example. I haven’t consulted WordNet for synonym, hyponym or hypernym information and the weights λ are just for illustration

Example (continued)

- The drawback of query expansion is that as well as increasing the overlap between a query *q* and a *relevant* document *d*, it may also increase the overlap with an *irrelevant* document
- Consider:
- *The crusades were a dark period in our history when knights moved from across Europe to join crusades to the holy land*
- This becomes: *crusade dark period history knight move europe crusade holy land*

Example (continued)

- In this case

  $q \cap d$ = {*dark*, *knight*, *move (syn(cinema))*,
  2 x *crusade(hypo(knight))*,
  urban(hyper(town)), *land(hyper(town))*}

- This document is likely to score higher similarity than the previous one

- So, the challenge is:
  - Expand queries *enough* to promote overlap with relevant documents...
  - ...but not so much that they overlap with irrelevant documents

Summary

- Query expansion
  - Feedback-based
  - Knowledge-based: Synonyms, hyponyms and hypernyms
- Goal is to increase overlap with query and relevant documents
- WordNet
- Generalization
- Example “toy” calculation