Query Operations

Relevance Feedback &
Query Expansion
Relevance Feedback

- After initial retrieval results are presented, allow the user to provide feedback on the relevance of one or more of the retrieved documents.
- Use this feedback information to reformulate the query.
- Produce new results based on reformulated query.
- Allows more interactive, multi-pass process.
Relevance Feedback Architecture

Query String

Revise Query

Query Reformulation

Feedback

Document corpus

IR System

Ranked Documents

ReRanked Documents

1. Doc1
2. Doc2
3. Doc3

1. Doc1
2. Doc2
3. Doc3

1. Doc2
2. Doc4
3. Doc5

.
Query Reformulation

• Revise query to account for feedback:
  – **Query Expansion**: Add new terms to query from relevant documents.
  – **Term Reweighting**: Increase weight of terms in relevant documents and decrease weight of terms in irrelevant documents.

• Several algorithms for query reformulation.
Query Reformulation for VSR

- Change query vector using vector algebra.
- **Add** the vectors for the *relevant* documents to the query vector.
- **Subtract** the vectors for the *irrelevant* docs from the query vector.
- This adds both positively and negatively weighted terms to the query as well as reweighting the initial terms.
Optimal Query

• Assume that the relevant set of documents $C_r$ are known.
• Then the best query that ranks all and only the relevant queries at the top is:

$$
\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{d_j \in C_r} d_j - \frac{1}{N - |C_r|} \sum_{d_j \notin C_r} d_j
$$

Where $N$ is the total number of documents.
Example

- Vocabulary: *(information, method, performance, retrieval, system)*
- **D1**: *(1,0,1,1,0)*  *information retrieval performances*
- **D2**: *(1,0,1,1,1)*  *performance of information retrieval systems*
- **D3**: *(0,1,0,0,1)*  *system’s method*
- **C_r**: **D1, D2**; **N-C_r** = **D3

\[
q_{opt} = \frac{1}{2} \left\{ (1,0,1,1,0) + (1,0,1,1,1) \right\} - \frac{1}{3-2} (0,1,0,0,1) = \\
\frac{1}{2} (2,0,2,2,1) - (0,1,0,0,1) = (1,0,1,1,0.5) - (0,1,0,0,1) = \\
(1,-1,1,1,-0.5)
\]
Standard Rocchio Method

- Since all relevant documents are unknown, just use the known relevant \((D_r)\) and irrelevant \((D_n)\) sets (among the first \(k\) ranked) of documents and include them in initial query \(q\).

\[
\tilde{q}_m = \alpha \tilde{q} + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} \tilde{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall d_j \in D_n} \tilde{d}_j
\]

\(\alpha\): Tunable weight for initial query.

\(\beta\): Tunable weight for relevant documents.

\(\gamma\): Tunable weight for irrelevant documents.
Dr: saga, movie, director, david slade, licantropus, melissa rosenberg..

Dn: foundation, software, development, tool, environment....
Ide Regular Method

• Since more feedback should perhaps increase the degree of reformulation, do not normalize:

\[
\tilde{q}_m = \alpha \tilde{q} + \beta \sum_{\forall d_j \in D_r} \tilde{d}_j - \gamma \sum_{\forall d_j \in D_n} \tilde{d}_j
\]

\(\alpha\): Tunable weight for initial query.
\(\beta\): Tunable weight for relevant documents.
\(\gamma\): Tunable weight for irrelevant documents.
Ide “Dec Hi” Method

• Bias towards rejecting just the highest ranked of the irrelevant documents:

\[
\tilde{q}_m = \alpha \tilde{q} + \beta \sum_{\forall d_j \in D_r} \tilde{d}_j - \gamma \max_{\text{non-relevant}} (\tilde{d}_j)
\]

\(\alpha\): Tunable weight for initial query.
\(\beta\): Tunable weight for relevant documents.
\(\gamma\): Tunable weight for irrelevant document.
Comparison of Methods

• Overall, experimental results indicate no clear preference for any one of the specific methods.
• All methods generally improve retrieval performance (recall & precision) with feedback.
• Generally just let tunable constants equal 1.
Why is Feedback Not Widely Used

- Users sometimes reluctant to provide explicit feedback.
- Results in long queries that require more computation to retrieve, and search engines process lots of queries and allow little time for each one.
- Makes it harder to understand why a particular document was retrieved.
Pseudo Feedback

- Use relevance feedback methods **without explicit user input**.
- Just **assume** the top $m$ retrieved documents are relevant, and use them to reformulate the query.
- Allows for query expansion that includes terms that are correlated with the query terms.
- Would not work well for previous Eclypse example but common queries are less ambiguous,
- E.g. Eclypse licantropous, Eclypse moon
Pseudo Feedback Architecture

Query String

Document corpus

IR System

ReRanked Documents

Revised Query

Ranked Documents

Query Reformulation

Pseudo Feedback

1. Doc1
2. Doc2
3. Doc3

1. Doc1
2. Doc2
3. Doc3

1. Doc2
2. Doc4
3. Doc5

PseudoFeedback Results

- Found to improve performance on TREC competition ad-hoc retrieval task.
- Works even better if top documents must also satisfy additional boolean constraints in order to be used in feedback (especially negative constraints like eclipse AND (licantropus OR not moon)).
OTHER METHODS FOR QUERY EXPANSION
Thesaurus

• A thesaurus provides information on synonyms and semantically related words and phrases.

• Example:

physician

    syn: ||croaker, doc, doctor, MD, medical, mediciner, medico, ||sawbones

    rel: medic, general practitioner, surgeon,
Thesaurus-based Query Expansion

- For each term, $t$, in a query, expand the query with synonyms and related words of $t$ from the thesaurus.
- Can weight added terms less than original query terms (= discount factor for terms not in original query).
- Generally increases recall.
- May significantly decrease precision, particularly with ambiguous terms.
  - “interest rate” $\rightarrow$ “interest rate fascinate evaluate”
WordNet

- A more detailed database of semantic relationships between English words.
- Developed by famous cognitive psychologist George Miller and a team at Princeton University.
- About 144,000 English words.
- Nouns, adjectives, verbs, and adverbs grouped into about 109,000 synonym sets called *synsets*. 
WordNet Synset Relationships

- **Antonym**: front → back
- **Attribute**: benevolence → good (noun to adjective)
- **Pertainym**: alphabetical → alphabet (adjective to noun)
- **Similar**: unquestioning → absolute
- **Cause**: kill → die
- **Entailment**: breathe → inhale
- **Holonym**: chapter → text (part-of)
- **Meronym**: computer → cpu (whole-of)
- **Hyponym**: plant → tree (specialization)
- **Hypernym**: apple → fruit (generalization)
WordNet Query Expansion

- Add synonyms in the same synset.
- Add hyponyms to add specialized terms.
- Add hypernyms to generalize a query.
- Add other related terms to expand query.
- In case of ambiguity, which synset?
Not all senses available

<table>
<thead>
<tr>
<th>Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <em>S:</em> (n) <em>apple</em> (fruit with red or yellow or green skin and sweet to tart crisp whitish flesh)</td>
</tr>
<tr>
<td>○ <em>direct hyponym</em> / <em>full hyponym</em></td>
</tr>
<tr>
<td>• <em>S:</em> (n) <em>crab apple, crabapple</em> (small sour apple; suitable for preserving) &quot;crabapples make a tangy jelly&quot;</td>
</tr>
<tr>
<td>• <em>S:</em> (n) <em>eating apple, dessert apple</em> (an apple used primarily for eating raw without cooking)</td>
</tr>
<tr>
<td>• <em>S:</em> (n) <em>cooking apple</em> (an apple used primarily in cooking for pies and applesauce etc)</td>
</tr>
<tr>
<td>○ <em>direct hypernym</em> / <em>inherited hypernym</em> / <em>sister term</em></td>
</tr>
<tr>
<td>○ <em>part holonym</em></td>
</tr>
<tr>
<td>• <em>S:</em> (n) <em>apple, orchard apple tree, Malus pumila</em> (native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits)</td>
</tr>
</tbody>
</table>

Computer sense of apple is missing
A better source: Wikipedia (disambiguation page)

Apple (disambiguation)

From Wikipedia, the free encyclopedia

The **apple** is the pomaceous edible fruit of a temperate-zone deciduous tree.

**Apple** or **apples** may also refer to:

**Plants and plant parts**

- *Malus*, the genus of all apples and crabapples
- *Cashew apple*, the fruit that grows with the cashew nut
- Several fruits called **Custard apple**
- Love apple
  - *Tomato*
  - *Syzygium samarangense*
- Plants called *Mammea apple*
- May apple, *Podophyllum peltatum*
- Oak apple, a type of gall that grows on oak trees
- Several fruits called **rose apple**
- Thorn apple:
  - *Crataegus* species
  - *Datura* species
- Wax apple, *Syzygium samarangense*

**Companies**

- **Apple Corps**, a multimedia corporation founded in the 1960s by The Beatles
- **Apple Inc.**, a consumer electronics and software company founded in the 1970s
- **Apple Bank**, an American bank in the New York City area

**Films**

- **The Apple** (1980 film), a 1980 musical science fiction film
- **The Apple** (1998 film), by Samira Makhmalbaf

**Television**

- "The Apple" (Star Trek: The Original Series), a 1967 second season episode

**Music**

- **Apple (album)**, a 1990 album by Mother Love Bone
- **Apple** (band), an American band
An even better source: BabelNet

BabelNet 2.0
A very large multilingual encyclopedic dictionary and semantic network

Type a term: apple

Noun
fruit with red or yellow or green skin and sweet to tart crisp whitish flesh
native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits
Apples appear in many religious traditions, often as a mystical or forbidden fruit.
Apple is the only full-length studio album by the American rock band Mother Love Bone.
The Apple was a short-lived American automobile manufactured by Apple Automobile Company in Dayton, Ohio from 1915 to 1917.
Apple were a British psychedelic rock band.
The Ariane Passenger PayLoad Experiment, was an experimental communication satellite with a C-Band transponder launched by Indian Space Research Organisation satellite on June 19, 1981 by Ariane, a launch vehicle of the European Space Agency from Centre Spatial Guyanais near Kourou in French Guiana.
Apple Inc. is an American multinational corporation that designs and sells consumer electronics, computer software, and personal computers.
The Russian United Democratic Party "Yabloko" is a Russian social liberal party founded by Grigory Yavlinsky and currently led by Sergey Mitrokhin, an opposition member of the Moscow City Duma. The party logo consists of a red circle and a green isosceles triangle, suggesting an apple in a constructivist style.
The Apple Retail Store is a chain of retail stores owned and operated by Apple Inc., dealing in computers and consumer electronics.
The Party of Free Democrats is a political party in Ukraine led by Mykhaylo Brodskyi.
Use context in query to disambiguate (e.g. “apple computer”) or click on right sense

Apple Inc. is an American multinational corporation that designs and sells consumer electronics, computer software, and personal computers.
STATISTICAL QUERY EXPANSION
Statistical Expansion

- Existing human-developed thesauri are not easily available in all languages.
- Human thesauri are limited in the type and range of synonymy and semantic relations they represent.
- Semantically related terms can be discovered from statistical analysis of corpora.
- E.g. “licantropie” and “eclipse” do not co-occur in thesauri: “The Twilight Saga: Eclipse, commonly referred to as Eclipse, is a 2010 American romantic fantasy film based on Stephenie Meyer's 2007 novel, Eclipse” but they do co-occur in texts
Automatic Global Analysis

- Determine term similarity through a **pre-computed statistical analysis** of the complete corpus.
- Compute association matrices which quantify term correlations in terms of how frequently they co-occur.
- Expand queries with statistically most similar terms.
# Association Matrix

\[
\begin{array}{cccccc}
  & w_1 & w_2 & w_3 & \cdots & w_n \\
 w_1 & c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\
 w_2 & c_{21} & & & & \\
 w_3 & c_{31} & & & & \\
 \vdots & \vdots & & & & \\
 w_n & c_{n1} & & & & \\
\end{array}
\]

- \( c_{ij} \): Correlation factor between term \( i \) and term \( j \)

\[
c_{ij} = \sum_{d_k \in D} f_{ik} \times f_{jk}
\]

- \( f_{ik} \): Frequency of term \( i \) in document \( k \)

- \( c_{ij} = 0 \) if either \( i \) or \( j \) do not occur in \( d_k \)

- \( c_{ii} = \) sum of quadratic frequencies
Normalized Association Matrix

- Frequency based correlation factor favors more frequent terms.
- Normalize association scores:
  \[ s_{ij} = \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}} \]
- Normalized score is 1 if two terms have the same frequency in all documents.

**Numerator:** SUM(product of \(i\)-\(j\) frequencies)

**Denominator:** SUM(frequency of \(i\))^2 + SUM(frequency of \(j\))^2 – numerator
Example (assuming freq=1 or 0 in all docs)

- Documents with “information”: 5500
- Documents with “retrieval”: 2600
- Documents with both: 2500

\[
\frac{2500}{5500 + 2600 - 2500} = \frac{2500}{5600} = 0.45
\]
Metric Correlation Matrix

- Association correlation does not account for the proximity of terms in documents, just co-occurrence frequencies within documents.
- Metric correlations account for term proximity.

\[ c_{ij} = \sum_{k_u \in V_i} \sum_{k_v \in V_j} \frac{1}{r(k_u, k_v)} \]

- \( V_i \): Set of all occurrences of term \( i \) in any document.
- \( r(k_u, k_v) \): Distance in words between word occurrences \( k_u \) and \( k_v \) (\( \infty \) if \( k_u \) and \( k_v \) are occurrences in different documents).
Normalized Metric Correlation Matrix

• Normalize scores to account for term frequencies:

\[ s_{ij} = \frac{c_{ij}}{|V_i| \times |V_j|} \]
Query Expansion with Correlation Matrix

- For each term $i$ in query, expand query with the $n$ terms, $j$, with the highest value of $c_{ij}$ ($s_{ij}$).
- This adds semantically related terms in the “neighborhood” of the query terms.
Problems with Global Analysis

• Term ambiguity may introduce irrelevant statistically correlated terms.
  – “Apple computer” → “Apple red fruit computer”

• Since terms are highly correlated anyway, expansion may not retrieve many additional documents.
Automatic Local Analysis

• At query time, dynamically determine similar terms based on analysis of top-ranked retrieved documents.
• Base correlation analysis on only the “local” set of retrieved documents for a specific query.
• Avoids ambiguity by determining similar (correlated) terms only within relevant documents.
  – “Apple computer” →
  “Apple computer Powerbook laptop”
Example (apple computer)

Apple Computer - Get great deals for Apple Computer on eBay!
popular.ebay.com/computers../apple... - Stati Uniti - Traduci questa pagina
The Apple Computer Co. began in the 1970s with the production of the behemoth Apple II microcomputer. Based in Cupertino, CA, in the heart of Silicon Valley, ...

AAPL Stock Price Today - Apple Inc. Stock Quote - WSJ.com
quotes.wsj.com/AAPL - Traduci questa pagina
Apple Inc. AAPL (U.S.: Nasdaq). Help. Real-time prices for U.S.-listed stocks, including premarket and after hours, reflect trading through Nasdaq only.

AAPL: Summary for Apple Inc.- Yahoo! Finance
finance.yahoo.com/q?s=AAPL - Stati Uniti - Traduci questa pagina
View the basic AAPL stock chart on Yahoo! Finance. Change the date range, chart type and compare Apple Inc. against other companies.

Apple Inc.: NASDAQ:AAPL quotes & news - Google Finance
www.google.com/finance?cid=22144 - Traduci questa pagina
Get detailed financial information on Apple Inc. (NASDAQ:AAPL) including real-time stock quotes, historical charts & financial news, all for free!
Global vs. Local Analysis

- Global analysis requires intensive term correlation computation only once at system development time.
- Local analysis requires intensive term correlation computation for every query at run time (although number of terms and documents is less than in global analysis).
- But local analysis gives better results.
Global Analysis Refinements

• Only expand query with terms that are similar to all terms in the query.

\[ \text{sim}(k_i, Q) = \sum_{k_j \in Q} c_{ij} \]

– “fruit” not added to “Apple computer” since it is far from “computer.”
– “fruit” added to “apple pie” since “fruit” close to both “apple” and “pie.”

• Use more sophisticated term weights (instead of just frequency) when computing term correlations.
Query Expansion Conclusions

- Expansion of queries with related terms can improve performance, particularly recall (more terms = more documents with same rank threshold).

- However, must select similar terms very carefully to avoid problems, such as loss of precision (e.g. if unrelated terms are added, precision might considerably decrease).
Google query expansion

- Use query logs:

Query is expanded with “hints” as you type words into the query window
Query expansion with query logs

Query logs: given a query, which documents have been accessed?

Prob of term j in query k

Relevance of term i in doc j
The Knowledge Graph
The Knowledge Graph

- A very large knowledge base of interlinked objects
- Probably extracted from Freebase (Freebase has many things in common with Wikipedia)
- (summer 2013) > 500mil. Objects, 3.5 billion facts
- Probably the newest biggest revolution in Google Search
Freebase is an online collaborative knowledge base of structured data harvested from many sources, including Wikipedia and individual users.
Want to know more?

The Anatomy of a Large-Scale Human Computation Engine

Shailesh Kochhar  
Metaweb Technologies  
kochhar@metaweb.com

Stefano Mazzocchi  
Metaweb Technologies  
stefano@metaweb.com

Praveen Paritosh  
Metaweb Technologies  
paritosh@metaweb.com
Query expansion with Knowledge Graph

Eclipse - The Eclipse Foundation open source community
https://www.eclipse.org/
A project aiming to provide a universal toolset for development. Open Source IDE, mostly provided in Java, but the development language is independent and ...

Downloads
Downloaded 1,322,740 Times Other Downloads The Eclipse ...

Eclipse Marketplace
Eclipse Marketplace is the source for Eclipse-based solutions ...

Documentation
Current releases. Eclipse Kepler (4.3) Documentation (HTML Help ...

PHP Development Tools
Downloads - PDT Documents - PDT Incubator - 3.2 - PDT - ...

CDT
CDT Downloads - Documentation - C/C++ Development Tooling

The Eclipse IDE
The award winning Eclipse IDE, and all the tools needed to ...

Eclipse (informatica) - Wikipedia
it.wikipedia.org/wiki/Eclipse_(informatica)
Eclipse è un ambiente di sviluppo integrato multi-linguaggio e multiplattaforma. Ideato da un consorzio di grandi società quali Ericsson, HP, IBM, Intel, ...

The Twilight Saga: Eclipse - Wikipedia
it.wikipedia.org/wiki/The_Twilight_Saga:_Eclipse
The Twilight Saga: Eclipse è un film del 2010 diretto da David Slade. Sceneggiato da Melissa ...

Eclissi
L'eclissi, o eclipsa, è un evento astronomico che avviene quando un corpo celeste, come un pianeta ...