Query Expansion

COMP3009J: Information Retrieval

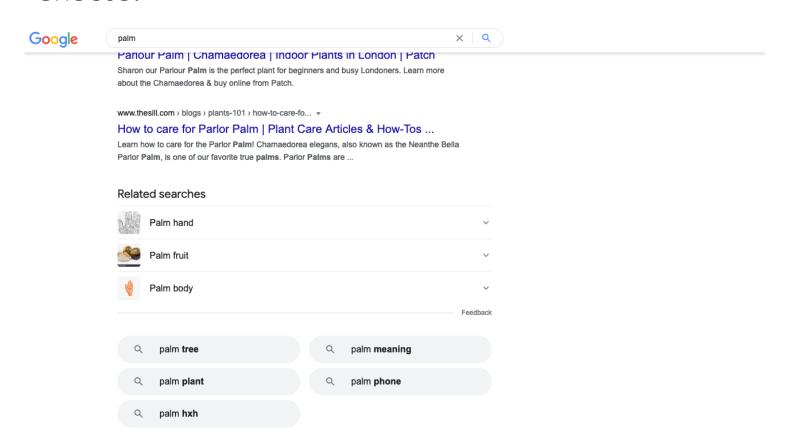Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

Query Expansion

- "**Query expansion** (QE) is a process in Information Retrieval which consists of selecting and adding terms to the user's query with the goal of minimizing query-document mismatch and thereby improving retrieval performance."*

  - Generally works by increasing **recall** by matching more documents than the original query would match.

  - Can be done **automatically** or **interactively** (with user involvement, similar to relevance feedback).

* Vechtomova, O. (2009). Query Expansion for Information Retrieval, https://doi.org/10.1007/978-0-387-39940-9_947
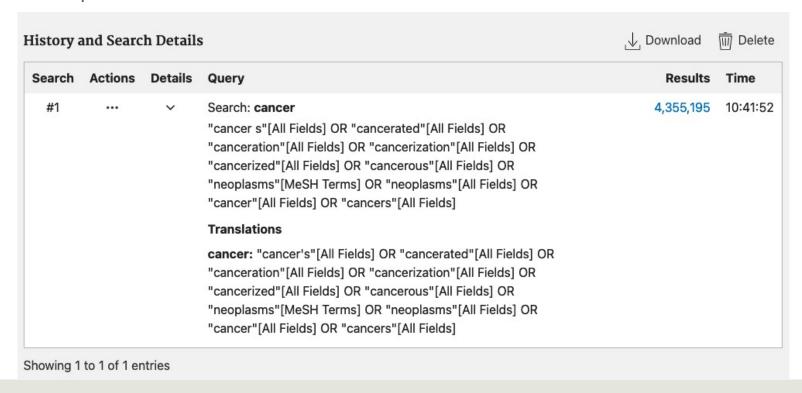
Interactive Query Expansion

■ In results for the search query "palm", Google has also suggested some query expansion options for the user to choose.

Automatic Query Expansion

- This query is shown from the medical search engine **PubMed**.

- The original query "cancer" has been automatically expanded with several related search terms.

History and Search Details   ↓ Download   🗑 Delete

Search	Actions	Details	Query	Results	Time
#1	⋯	⌄	Search: **cancer** "cancer s"[All Fields] OR "cancerated"[All Fields] OR "canceration"[All Fields] OR "cancerization"[All Fields] OR "cancerized"[All Fields] OR "cancerous"[All Fields] OR "neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "cancer"[All Fields] OR "cancers"[All Fields] **Translations** **cancer:** "cancer's"[All Fields] OR "cancerated"[All Fields] OR "canceration"[All Fields] OR "cancerization"[All Fields] OR "cancerized"[All Fields] OR "cancerous"[All Fields] OR "neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "cancer"[All Fields] OR "cancers"[All Fields]	4,355,195	10:41:52

Showing 1 to 1 of 1 entries

Resources for Query Expansion

- The main information used is **synonymy** (or near-synonymy), i.e. words that have the same or similar meanings.

- A **thesaurus** is a type of database that collects synonyms (or near-synonyms)
  - **Manual** thesaurus (maintained by human editors)
  - **Automatically derived** thesaurus
  - Using **word embeddings** as a similar effect, without using a thesaurus.

- **Query Log Mining**: based on queries previously submitted by other users.

Thesaurus-based Query Expansion

- For each term in the query, expand the query by adding words from the thesaurus that are similar to it.
  - e.g. "cancer" → "cancer, cancerous, neoplasms" etc.

- Generally increases recall, but may significantly reduce precision.
  - e.g. "interest rate" → "interest fascinate rate"

- Widely used in specialised search engines for science and engineering.

- Creating and maintaining a manual thesaurus is expensive, and is a lot of work!
  - Therefore, we seek automated methods to create a thesaurus.

Automatic Thesaurus Creation

- A thesaurus can be created automatically by analysing the **distribution of words in documents**.

- Attempt to find similarity between words.

- 2 possible definitions:
  1. Two words are similar if they co-occur with similar words (e.g. "car" and "truck" may be similar because they both occur with words like "road", "driver", "license", etc.)
  2. Two words are similar if they occur in particular grammatical relations with the same words (e.g. you can "harvest", "peel", "eat", "prepare" apples and pears, so apples and pears may be similar).

Automatically Generated Thesaurus (example)

Word	Nearest Neighbours
absolutely	absurd, whatsoever, totally, exactly, nothing
bottomed	dip, copper, drops, topped, slide, trimmed
captivating	shimmer, stunningly, superbly, plucky, witty
doghouse	dog, porch, crawling, beside, downstairs
makeup	repellent, lotion, glossy, sunscreen, skin, gel
mediating	reconciliation, negotiate, case, conciliation
keeping	hoping, bring, wiping, could, some, would
lithographs	drawings, Picasso, Dali, sculptures, Gauguin
pathogens	toxins, bacteria, organisms, bacterial, parasite
senses	grasp, psyche, truly, clumsy, naïve, innate

Word Embeddings for Query Expansion

- In recent times, researchers have used neural network approaches to create **word embeddings**: a representation of word as vectors in an n-dimensional space.
  - `word2vec` is probably the most famous: 300-dimension vectors.
    - Other, newer ones include `GloVe`, `ELMo`, or embeddings from language models like `BERT`.
  - Because words are represented as vectors, a **similarity** between words can easily be calculated.
    - It is assumed that words that are close to one another in the vector space have similar meaning.

* Diaz, F., Mitra, B., and Craswell, N. (2016) Query Expansion with Locally-Trained Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Search Engine Query Expansion (Query Log Mining)

- ☐ The **query log** stores details about previous queries and user behaviour.

- ☐ **Example:** After searching for "palm", many users next search for "palm oil".

- ☐ **Example:** Users searching for "football" and "soccer ball" often click on the same URL in the search results. Therefore, these are potential expansions of each other.

Summary

- Query Expansion is the process of adding terms to a query in order to **increase recall**.

- However, it is not a perfect process, and it can have negative effects also.

- A **manual thesaurus** is the most reliable method, but is expensive and difficult to maintain.

- Therefore, several **automatic** methods have been proposed.