Query Modification

41

Query Modification

There is a gap between the information needed and the query that is/was actually made:

- **Information needed**: the information that a user needs to achieve a goal
- **Query**: a representation of an information need for the IR system, tipically using a list of words, natural language, or other formats.
- **The effectiveness** of the IR system also depends on the user's ability to formulate a query. For the same information need, better or worse queries can be introduced
- Thefore, the expressiveness of the query language is important

42

Query Modification (cont.)

- **Goal:** It is possible to **increase the recall**, i.e. the number of relevant documents returned to user.
- To improve the results different query modification techniques can be used:
  - Global methods:
    - Query expansion;
  - Local methods:
    - Relevance feedback
    - Pseudo feedback.

43

Query Modification: Global Query Expansion

Idea: Add terms based on "global information" that are not query-specific.

- Each word of the query is expanded with synonyms or related words from a thesaurus.
- **Added terms should have less weight than the original ones**.
- Generally, it increases recall. Example: hospital → medical.
- It is widely used in specialized search engines for science and engineering.
- The method can also determine word similarity based on their co-occurrence with other similar words. Example: car → motorcycle.
- Disadvantage:
  - The construction of the thesaurus is expensive.

44

Query Modification: Relevance Feedback

Idea: Add terms that are based on "local" information of the result list.

- **Process**:
  1. The user issues a (short, simple) query.
  2. The search engine returns a set of documents.
  3. User marks some docs as relevant and some as non-relevant.
  4. The search engine computes a new representation of the information need.
  5. Search engine runs new query and returns new results.
  6. New results have (hopefully) a better recall.
  7. We can iterate this: several rounds of relevance feedback

45

Query Modification: Pseudo-Relevance Feedback

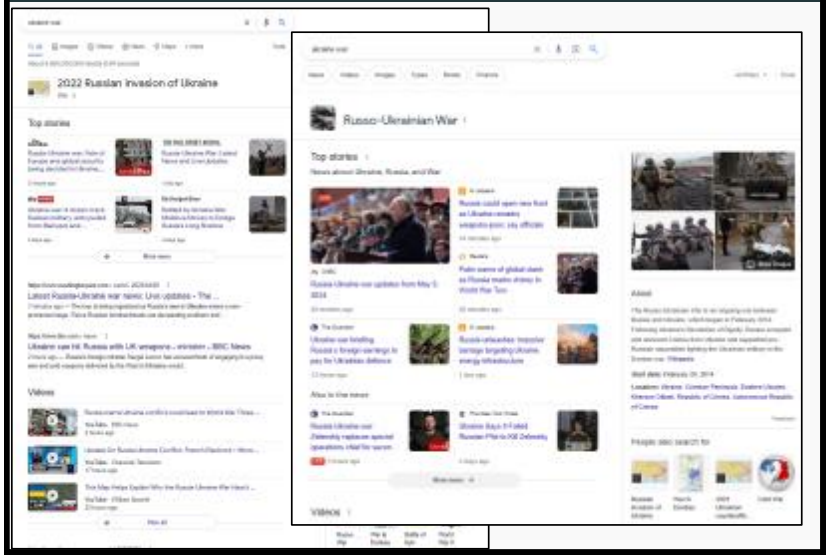
- Also known as "blind relevance feedback"
- Pseudo-relevance feedback **automates** the "manual" part of true relevance feedback.
- Pseudo-relevance algorithm:
  - retrieve a ranked list of hits for the user's query;
  - assume that the **top k** documents are relevant;
  - do query expansion with the k assumed relevant documents.
- Works very well on average!
- Disadvantages:
  - Can go horribly wrong for some queries.
  - Several iterations can cause **query drift**.

46

Web Retrieval

48

What makes the Web special for Information Retrieval?



49

What makes the Web special for Information Retrieval? (cont.)

The Web is a "special collection"

- Distributed Data: documents spread over millions of different web servers.
- Volatile Data: many documents change or disappear rapidly (e.g. dead links).
- Large Volume: billions of separate documents.
- Unstructured and Redundant Data: no uniform structure, HTML errors, up to 30% (near) duplicate documents.
- Quality of Data: no editorial control, false information, poor quality writing, typos, etc.
- Heterogeneous Data: multiple media types (images, video, VRML), languages, character sets, etc.
- Users: different backgrounds and knowledge.
- Hyperlinks: relations among web pages.

50

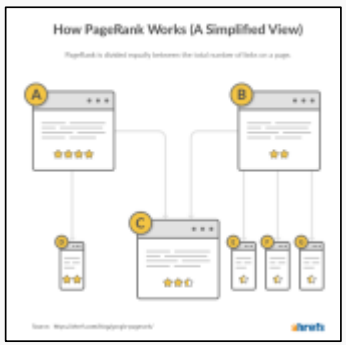
What makes the Web special for Information Retrieval? (cont.)

- Weighting the terms is not so simple in Web retrieval.
- It depends on where the html-tag occurs. For example, in the title of the web page, emphasized, in the text of a link.
- The terms occuring in these places will be configured as more relevant terms than those occurring in the body.
- Web page pre-processing involves:
  1. identifying different text fields (e.g., title, metadata, body, h1, h2, etc.);
  2. identifying anchor text;
  3. removing html tags;
  4. identifying main content blocks (e.g., disregard advertising information).

51

What makes the Web special for Information Retrieval? (cont.)

- **The relevance value of a web page depends on:**
- the score computed by the retrieval model given a query
- the quality of the document, independently on the query (e.g., using link analysis)
- Combination of both values gives a unique score for each document



52

References

53

References

- Aggarwal, C. C. (2015).
  ***Data Mining, The Texbook*.**
  Springer.
- Gandomi, A. and Haider, M. (2015).
  Beyond the hype: Big data concepts, methods, and analytics.
  International Journal of Information Management, 35(2):137 - 144.
- Han, J., Kamber, M., and Pei, J. (2011).
  ***Data Mining: Concepts and Techniques*.**
  Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.
- Liu, B. (2011).
  ***Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*.**
  Springer, 2nd edition.
- Ravi, K. and Ravi, V. (2015).
  A survey on opinion mining and sentiment analysis: Tasks, approaches and applications.
  Knowledge-Based Systems, 89:14 - 46.
- Zhang, L. and Liu, B. (2017).
  ***Sentiment Analysis and Opinion Mining*, pages 1152–1161.**
  Springer US, Boston, MA.

54