Information Retrieval Methods

Helena Ahonen-Myka

Spring 2007, part 11

Retrieval strategies

User interfaces and visualisation

Translation from Finnish: Greger Lindén

In this part

- Retrieval strategies
  - querying, browsing, navigation, scanning
  - filtering and routing
- User interfaces and visualisation

Retrieval process

1. The user has an information need
2. The user forms a query
3. The user sends the query to a system
4. The system returns an answer set
5. The user eyes and evaluates the results
6. If the user is satisfied, s/he stops
7. If the user is not satisfied, s/he modifies the query and returns to step 3

Retrieval process

- Background hypothesis:
  - the information need of the information seeker does not change during the retrieval process
  - the process is successful if, by modifying the query iteratively, the end result is a set of all relevant documents and no non-relevant ones

Retrieval process

- In practice the user learns new things during the process
  - the user eyes the titles of the result list, search terms in context, result documents and navigates following hyperlinks
- "the berry picking model"
  - the user's information need changes during the process
  - the information need is satisfied during the retrieval process by eyeing or reading information fragments
- in addition to querying, other retrieval strategies are scanning, browsing and navigation

Querying, browsing, navigation and eyeing

- querying
  - documents are described explicitly with query words
  - the result is ad hoc document clusters
- browsing
  - the user starts from some possibly interesting topic/idea/document and browses documents to find relevant ones
  - if no relevant documents are found, the user will move to somewhere else
  - the starting point can be found by querying
  - assumption: documents on the same topic are organised together

Querying, browsing, navigation and eyeing

- navigating
  - the user follows hyperlinks towards a known goal (e.g. the phone number of N.N. at the Department of Computer Science)
  - the route is assumed to be known, or it is easily found out while navigating
- scanning
  - the user scans the titles of the answer list, documents, hyperlinks, meta data, etc.
- selection
  - auxiliary operation: e.g. when scanning, the seeker selects a hyperlink to follow

7

Content-based information filtering and routing

- filtering
  - the goal is to select for a person or an organisation from a document flow (e.g. today's news, emails) interesting documents or remove unwanted ones
- routing
  - a document from a document flow is routed to a person who is interested in the document or to whose field of activities it belongs (e.g. questions by customers are routed to different experts)

8

Content-based information filtering and routing

- filtering and routing are based on filters (profiles)
- the document collection in a retrieval system is usually quite static, but queries vary
- in filtering and routing, the document collection changes continuously, but the filters are used for a long time and change only rarely (filters are like static queries)

9

Content-based information filtering and routing

- filters can be based on meta data of documents (e.g. the sender of emails), but also on the contents of documents
- exact matching: the filters are applied as Boolean queries on each incoming document

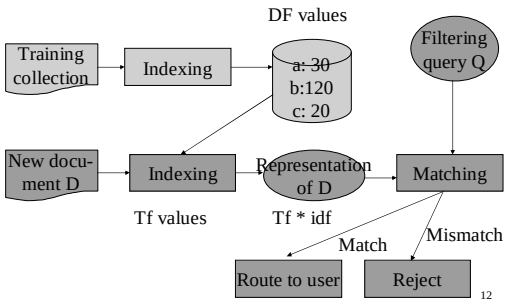
10

Content-based information filtering and routing

- Partial matching: the relevance of each document to the filter → accept/reject or the best receiver is selected
  - Problem: the collection does not actually exist → how can we compute df values for term weights
  - Solution: the df values of terms can be learnt from similar training materials (collections)

11

Content-based information filtering and routing



12

User interfaces and visualisation

- Overview of the document collection(s)
- Interfaces for specifying queries
- Visualisation of search results and their context

- This part based on
  - Chapter 10 “User Interfaces and Visualization” (by Marti A. Hearst) in Baeza-Yates&Ribeiro-Neto’s book Modern Information Retrieval
  - Chapter also available on the web (link from our course page)

13

Overview of the document collection(s)

- we can generate overfiews by clustering
  - with labels for clusters
  - e.g. scatter/gather method (see part 6)
- graphical visualizations
  - e.g. WEBSOM (websom.hut.fi)
- manually (semi-automatically) generated hierarchies
  - e.g. Yahoo!, medical concept hierarchies (MeSH)

14

Interfaces for specifying queries

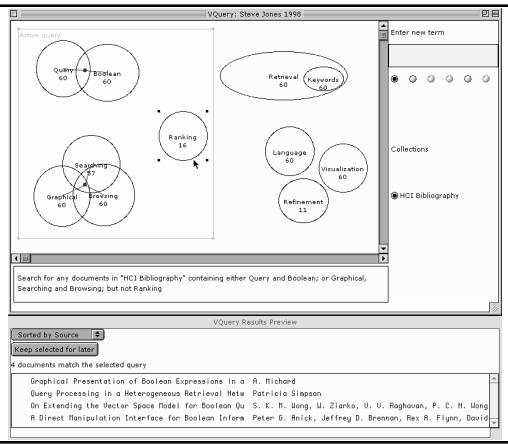
- forming Boolean queries can be difficult for many users
  - e.g. AND and OR do not correspond to their counterparts in standard language
  - “dogs and cats”, “tee or coffee”
- quorum search may help
  - automatic reformulation of the query from strict to loose
- also interfaces to define flexible forms of faceted queries can be offered
  - (osteoporosis OR ‘bone loss’)
  - (drugs OR pharmaceuticals)
  - (prevention OR cure)

15

Interfaces for specifying queries: graphical solutions

- Venn diagrams (Hearst: figure 10.10)
- the user can assign any number of query terms to ovals
  - if two or more ovals are placed such that they overlap with another, and if the user selects the area of their intersection → an AND operation is implied among the terms
  - if the user selects outside the area of intersection but within the ovals, an OR is implied among the corresponding terms
  - a NOT operation is associated with any term whose oval appears in the active area of the display but which remains unselected
- an active area indicates the current query: all groups of ovals within the active area are considered to in the query
  - ovals containing query terms can be moved out of the active area for later use

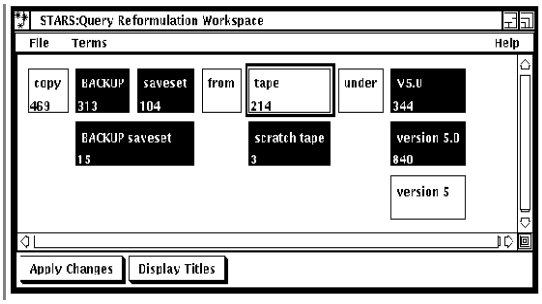
16



Interfaces for specifying queries: graphical solutions

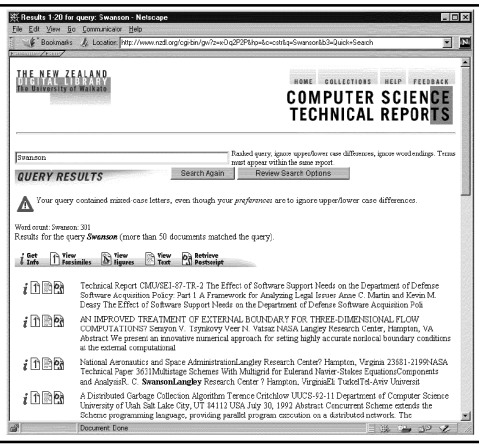
- block-oriented diagrams (restricted and parallel concepts) (Hearst: figure 10.12)
- the user types a natural language query which is automatically converted to a representation in which each query term is represented within a block
- the block are arranged into rows and columns
  - two or more blocks are in the same row → AND
  - two or more blocks are in the same column → OR
- the user can experiment with different combinations of terms by activating and deactivating blocks

18



Visualization of search results and their context

- a typical way: document surrogates
  - document titles, a fragment from the beginning, a link to an abstract, the class code, similarity value… (Hearst: figure 10.14)

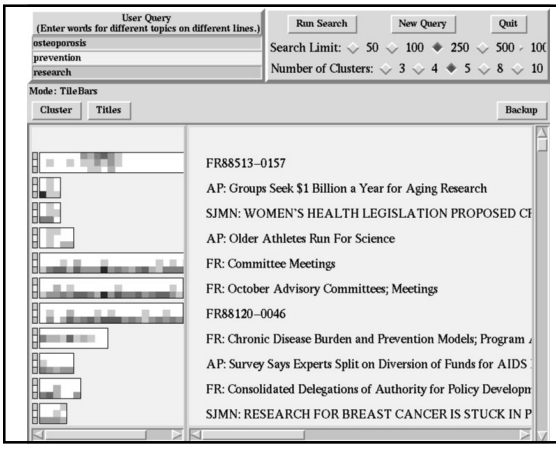


Visualization of search results and their context

- highlighting of query terms
  - the user can more easily perceive the answer set, if the occurrences of the search words are somehow highlighted in the documents
- KWIC (keyword-in-context)
  - sentences where the query terms occur: summarize the ways the terms are used within a document
  - decisions:
    - How many sentences?
    - Which sentences? E.g. sentences near the beginning with the largest subset of query terms.
    - Which order? Usually in order of occurrence, independent of how many query terms they contain.
  - the retrieval system must have a copy of the original document (web search engines may not have)

Visualization of search results and their context

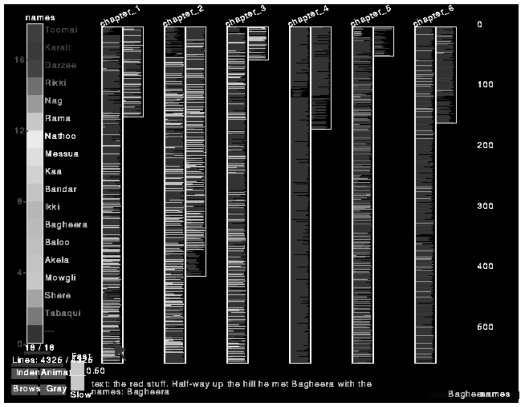
- TileBars
  - the user enters a query in faceted format
  - the system displays a graphical bar next to the title of each retrieved document, showing the degree of match for each facet
    - the user can see in which documents all the facets are present
    - (Hearst: figure 10.15, better picture in the PDF version)



Visualization of search results and their context

- SeeSoft
  - visualising occurrences of search terms
  - each column denotes one section (in a book)
  - each colour denotes the occurrence of some person (name) in the text
  - (Hearst: figure 10.16)

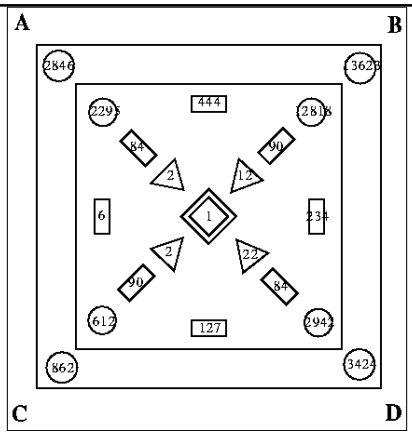
25



Visualization of search results and their context

- InfoCrystal
  - for each combination of search terms: in how many documents does the combination occur
  - the user does not have to specify Boolean operators in the query
  - allows visualization of all possible relations among N user-specified terms (although beyond 4 terms the interface becomes difficult to understand)
  - (Hearst: figure 10.17)

27

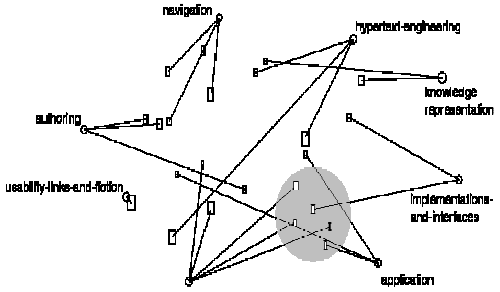


28

Visualization of search results and their context

- Vibe
  - query terms are placed in conceptual space
  - after the search, result documents are positioned in this space
    - a set of document that contain three query terms are shown at a point midway between the representations of the three terms
  - (Hearst: figure 10.18)

29

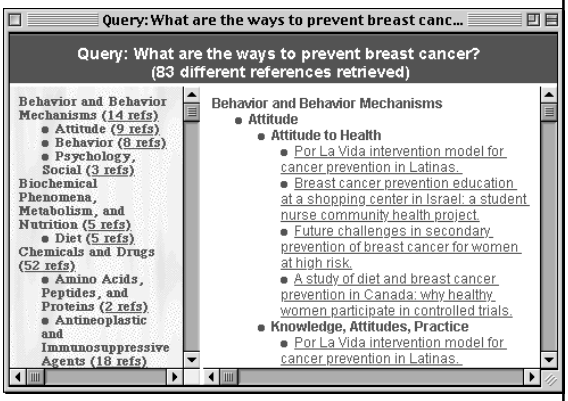


30

Visualization of search results and their context

- DynaCat
  - the answer set is ordered according to a classification system
  - all classes are not shown, only those that are relevant according to predefined query types
    - example of a type: ”Behaviour and behaviour mechanisms”
    - a query that belongs to the type: ”what are the ways to prevent breast cancer?”
  - (Hearst: figure 10.20)

31



User interfaces and visualisation

- there are naturally many other subfields in designing user interfaces for retrieval systems
  - relevance feedback: what is automated, what is left in control of the user
  - supporting the retrieval process : e.g. how is the retrieval history stored; using a result as input for the next phase (query)
  - supporting long-term retrieval processes e.g. continuous follow-up of competing enterprises

33

In this part

- Retrieval strategies
  - querying, browsing, navigation, scanning
  - filtering and routing
- User interfaces and visualisation
  - Overview of the document collection(s)
  - Interfaces for specifying queries
  - Visualisation of search results and their context

34

Presentation of project work (19 February)

- Each project group will give an informal presentation during the last exercise session on Monday February 19th (starting at 12.15 in C221)
- The length of the presentation should be about 15-20 minutes
- The project work does not have to be completed at the time of the presentation
  - the aim is to give an overview of the progress so far (what is your topic, what kind of queries and results you have studied, etc.)
- Remember that the project report deadline is on Friday, March 9th

35