How can library materials be ranked in the OPAC?

Prof. Dr. Dirk Lewandowski
University of Applied Sciences Hamburg
dirk.lewandowski@haw-hamburg.de

9th International Bielefeld Conference
Bielefeld, 4 February 2009

Agenda

The state of the OPAC and the importance of relevance ranking

Ranking factors

The composition of results lists

Conclusions

Agenda

The state of the OPAC and the importance of relevance ranking

Ranking factors

The composition of results lists

Conclusions

What’s wrong with library catalogueues?

- **catalogueues are incomplete**
  - Items from journal article collections, abstracting and indexing databases

- **“Electronic card catalogueue”?**

- **User behaviour changed**
  - Short queries, fast results, one set of results
  - Search engines strongly influence users’ demands

- **Known item vs. topic-based search**
  - OPACs should accomodate both.

Some ideas to improve the OPAC (“catalogue 2.0”)

- **Let users participate**
  - Write reviews
  - Rate titles
- **Enrich bibliographic data**
  - Add reviews
  - Add TOC
- **Improve navigation**
  - Drill-down menus on results pages to combine searching and browsing
- **Extend the database**
  - Federated search

Core of all search appliances: Relevance ranking

- **While Web 2.0 features add value to the catalogue, search is still the core.**
- **“Search must work”**
- **Users’ needs**
  - Users want results quickly.
  - Users are not willing to think too much about formulating their queries.
  - Users are not willing to search for the right database before conducting their search.
  - Users are only willing to view a few results on the first results page before deciding to continue.

Misconceptions about relevance ranking

- **A clear sorting criterion is better than relevance ranking.**
  - Ranking does not reduce the number of results, but puts them in a certain order.
  - Other ordering options can be given.
- **Library catalogues do not apply any form of ranking.**
  - Even conventional OPACs rank the results (according to publication date).
- **Relevance ranking is useless because it simply doesn't work.**
  - "Relevance" is hard to determine and depends on the context and on the individual user. However, a good relevance ranking can at least produce sufficient results lists.
- **Ranking is not that complicated. One must only apply standard measures such as TF/IDF.**
  - For a good ranking, text matching alone is insufficient.

Agenda

The state of the OPAC and the importance of relevance ranking

Ranking factors

The composition of results lists

Conclusions

Ranking factors in web search engines

- **Text matching**
  - Measures matching between query and document.
  - Term frequency, position of search terms within the documents, etc.
  - Text from document fulltexts, anchor texts.
- **Popularity**
  - Measures popularity of the document (overall popularity or topic-based)
  - Link popularity (PageRank etc.), click popularity.
- **Freshness**
  - Fresh documents can sometimes be very useful.
  - Derived from documents or from structural data (e.g., linkage)
- **Locality**
  - Mainly expressed in differing rankings for country-specific search interfaces.

Text matching

- **Factors**
  - Term frequency, inverted document frequency
  - Fields: Title, subject headings, author, etc.

- **Availability of text elements as a ranking factor**
  - Fulltext, TOC, reviews, user comments

- **Problems with text matching**
  - Not enough text in metadata.
  - Amount of text varies considerably (from mere bibliographic data to hundreds of pages of fulltext).

Popularity

- **Popularity of**
  - Item
  - Author/editor
  - Publisher
  - Book series
- **Measures**
  - Number of items (by author, publisher, etc.)
  - Usage (circulation rate, download requests)
  - Average user rating
  - Citations

Freshness

- **Freshness is the most-used ranking criterion in catalogues today.**

- **It is often difficult to determine whether fresh items will be relevant to a certain query.**

- **Need for fresh items can be derived from**
  - Circulation rate for the individual item
  - Circulation rates for items from a certain group (from broad disciplines to specific subject headings)

Locality

- **Availability of item**
  - from the local library; within a certain distance.
  - Item currently available.

- **Physical location of the user**
  - At home (electronic items strongly preferred)
  - At the library

Other ranking factors

- **Size of item (no. of pages)**
- **Document types**
  - – Monograph, edited book, proceedings, etc.
  - – Article vs. Book
  - – Physical vs. online materials
- **User group**
  - – Professor, undergraduate student, graduate student, etc.
- **Personalization**
  - – Individual usage data
  - – Click-stream data from navigation

Agenda

The state of the OPAC and the importance of relevance ranking

Ranking factors

The composition of results lists

Conclusions

Data needed

- **Data from the catalogue**
- **Circulation data**
  - Anonymous
- **Location data**
  - From IP ranges
- **User data**
- **Data from remote resources**
  - Abstracts (and fulltexts) from publishers.

Collections and databases

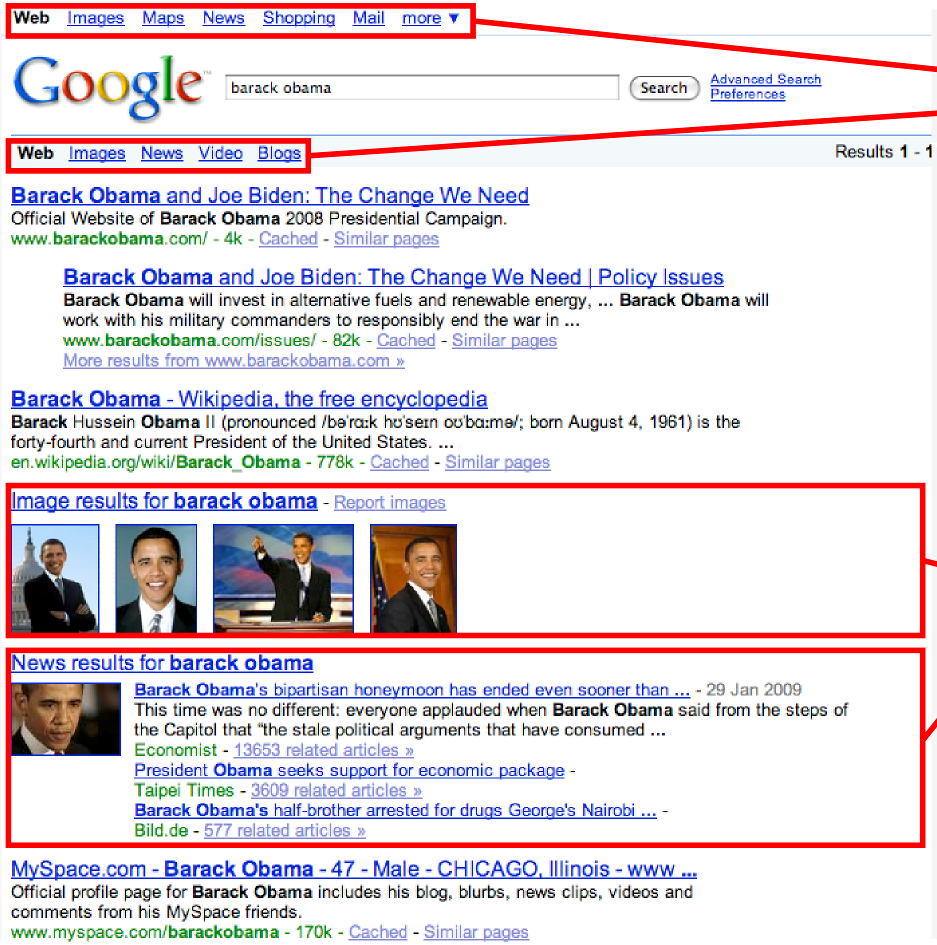
- **Library controlled**
  - catalogue
  - Local digital repositories
  - Course management systems
  - Institutional web sites

- **External collections**
  - A&I databases
  - E-journal collections

Mixed results lists

- **Ranking algorithms prefer "more of the same". This does not satisfy users' needs for a variety of results.**
- **Example for a broad query**
  - Reference works (from subject headings + items from reference collection)
  - Text books
  - Relevant databases
  - Some current items
  - Relevant journals

“Universal Search”



Web Images Maps News Shopping Mail more ▼

barack obama | Search | Advanced Search Preferences

Web Images News Video Blogs — Results 1 - 1

Barack Obama and Joe Biden: The Change We Need
Official Website of **Barack Obama** 2008 Presidential Campaign.
www.**barackobama**.com/ - 4k - Cached - Similar pages

Barack Obama and Joe Biden: The Change We Need | Policy Issues
Barack Obama will invest in alternative fuels and renewable energy, ... **Barack Obama** will work with his military commanders to responsibly end the war in ...
www.**barackobama**.com/issues/ - 82k - Cached - Similar pages
More results from www.barackobama.com »

Barack Obama - Wikipedia, the free encyclopedia
Barack Hussein **Obama** II (pronounced /bəˈrɑːk hʊˈseɪn oʊˈbɑːmə/; born August 4, 1961) is the forty-fourth and current President of the United States. ...
en.wikipedia.org/wiki/**Barack_Obama** - 778k - Cached - Similar pages

Image results for **barack obama** - Report images

News results for **barack obama**
Barack Obama's bipartisan honeymoon has ended even sooner than ... - 29 Jan 2009
This time was no different: everyone applauded when **Barack Obama** said from the steps of the Capitol that "the stale political arguments that have consumed ...
Economist - 13653 related articles »
President **Obama** seeks support for economic package -
Taipei Times - 3609 related articles »
Barack Obama's half-brother arrested for drugs George's Nairobi ... -
Bild.de - 577 related articles »

MySpace.com - **Barack Obama** - 47 - Male - CHICAGO, Illinois - www ...
Official profile page for **Barack Obama** includes his blog, blurbs, news clips, videos and comments from his MySpace friends.
www.myspace.com/**barackobama** - 170k - Cached - Similar pages

Additional databases

One box results (e.g., news or images)

Agenda

The state of the OPAC and the importance of relevance ranking

Ranking factors

The composition of results lists

Conclusions

Conclusions

- **Search is the core of the library catalogue.**
  - However, other elements must be considered, too:
    - Usability
    - User guidance
    - Spelling corrections
    - etc.
- **A good ranking is always a mixture of ranking factors**
- **In addition, results lists should be mixed.**
  - Items from different collections.
  - Mixture of direct results and pointers to other collections.
- **Future: catalogue will become more like a search engines.**

Thank you for your attention.

Prof. Dr.
Dirk Lewandowski

Hamburg University of Applied Sciences
Department Information
Berliner Tor 5
D - 20099 Hamburg
Germany

www.bui.haw-hamburg.de/lewandowski.html
E-Mail: dirk.lewandowski@haw-hamburg.de