The miracle of full-text searching: myth or reality?

Tim Hutchinson
University of Saskatchewan Archives
ACA/AAO – June 3, 1999

Outline

- How do we measure retrieval performance?

- Library/information science literature

- Archival literature

- Conclusions
  - ▶ Search interfaces/system design
  - ▶ Further research
  - ▶ Options for controlled vocabulary

How do we measure retrieval performance?

- Most established – recall and precision
- Comparing systems/strategies – overlap rates
- Other suggested techniques, e.g. failure analysis (qualitative), user-centred measures

How do we measure retrieval performance?

Problems/debates

- "Real" users vs. subject specialists or information professionals
- What is "relevance"?
- Interactive searching vs. batch searching
- Controlled conditions vs. "real world"

Recall and precision

Recall

(# of relevant documents retrieved)

(# total of relevant documents)

Recall and precision

Precision

$$\frac{(\text{\# of relevant documents retrieved})}{(\text{\# of documents retrieved})}$$
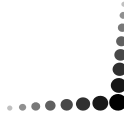
Recall and precision

Relationship?

- Controlled vocabulary increases precision
- Free text/full text: increases recall but decreases precision (false drops)
- Adding access points: decreases precision
- Controlled vocabulary: groups like items but not comprehensive
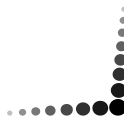- Inverse relationship often found

Library/information science literature

- Finding aids = full text; Fonds-level descriptoins = abstracts ???

- Vast literature; increasing number of articles re full text, especially commercial databases

- Most studies show full-text gives higher recall, lower precision

Library/information science literature

- Hard to compare studies – do the numbers have meaning independently?

- Some studies use test-bed of documents (laboratory conditions)

- Conclusion of lit review by Rowley: combination of controlled vocabulary and free text needed
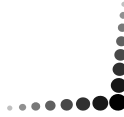
Library/information science literature

- Analysis of why searches fail: differences in terminology between searcher and author

- Research into user needs, information-seeking behaviour
  - ‣ Difficulty of using subject headings
  - ‣ Frustration with long hit lists

Archival literature

- Richard Lytle, "Intellectual access to archives: provenance and content indexing methods of subject retrieval" (American Archivist, Spring and Winter 1980)

- Helen Tibbo, "The epic struggle: subject retrieval from large bibliographic databases" (American Archivist, Spring 1994)

- Avra Michelson, "Description and reference in the age of automation" (AA, Spring 1987)

Archival literature

- Fernanda Ribeiro, "Subject indexing and authority control in archives: the need for subject indexing in archives and for an indexing policy using controlled language" (Journal of the Society of Archivists, 1996)

- Tim Hutchinson, "Strategies for searching online finding aids: a retrieval experiment" (Archivaria, Fall 1997)

What kind of controlled vocabulary?

- Topical subjects often assumed, but also:
  - ‣ Function
  - ‣ Occupation
  - ‣ Form/genre
  - ‣ Geographic

- Enhanced authority files; improved provenance-based access

- Philosophical and practical difficulties of assigning subject headings to archival material

Searching environments

- Remote users vs. mediated access

- Search techniques in automated system
  - Keyword in multiple fields
  - Context searching
  - Easier to pearl grow

- How do users search?
  - Build into system design?

Further research

- Archival finding aids/databases need to be studied separately (library literature not enough)

- User needs/behaviour (e.g. UofT/FIS)

- Multi-level description – interactive searching quite different than for library databases
  - ▸ Somewhat analogous to web/hypertext
  - ▸ How do users navigate?
  - ▸ How is success measured?

Conclusions

- Combination of controlled and uncontrolled vocabulary

- Need good search interfaces and system design to take advantage of controlled access points

- Flexibility, flexibility, flexibility

- Structured text

Conclusions cont'd

- Effectiveness of system depends on user requirements
- Approach to take for CAIN/Canadian archives?