NEEDLE IN HAYSTACK
MY BACKGROUND

- Plant chemist, then “reports librarian”
- Librarian, industrial research laboratory
- With Classification Research Group, 1952-60
- National Lending Library, Boston Spa
- University library
- Research and Consultancy, ASLIB
- Director, UCL SLAIS
- Development of online search aids
CONTEXT OF RETRIEVAL

ACTIVITIES
produces
problem solving
produces
SKILLS
produces
problem solving
produces
ARTEFACTS

WORLD
interaction
models and interprets

PEOPLE
communicative
interaction

PUBLIC KNOWLEDGE

PERSONAL KNOWLEDGE
contributes

INFORMATION
to

recorded and disseminated
1. Each social activity generates and shares its “community of practice” knowledge, expressed in its own special terminology.

2. Part or whole of this knowledge may be put in the public domain, the rest being confidential.

3. The everyday practices of people generate “general knowledge”, expressed in everyday language, and much of this is recorded and available in the public domain.

4. The sources of “public knowledge” are therefore 2 and 3. Some of it is in general language, the rest in a great variety of specialist terminologies.
DEVELOPMENTS - 1

since the 1930s

- Classified card catalogues and alphabetical indexes
- Experiments with mechanical and microfilm selectors
- Development of digital computing
- DBMS for structured data
- Magnetic tapes for current awareness
- Online computer databases for text
- CD-ROM systems, OPACs
DEVELOPMENTS - 2

- Research into computer retrieval and retrieval evaluation
- Research into NLP and AI
- Development of online search aids
- Internet retrieval tools – e.g. FTP, Gopher
- Web browsers
- Search engines
- Research into multimedia, multilanguage
UNCERTAINTY IN RETRIEVAL

DOCUMENTS

Content [Perceived subject(s)] Representations

USERS

Information need [Request] Query

(what we need)

(what we match)
### METHODS of representing subject content

<table>
<thead>
<tr>
<th>Word indexing</th>
<th>Categorising</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective</td>
<td>Enumerative</td>
</tr>
<tr>
<td>(extracted or assigned)</td>
<td>Faceted</td>
</tr>
<tr>
<td>Full text</td>
<td>Weighting</td>
</tr>
<tr>
<td></td>
<td>Scan</td>
</tr>
<tr>
<td>Precoordinate</td>
<td>Ranked output</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Scan</td>
<td>Coordinate (freely faceted)</td>
</tr>
<tr>
<td>Boolean search</td>
<td></td>
</tr>
</tbody>
</table>
PROBLEMS(1)
(and attempted solutions)

In using words

. Word forms (stemming, truncation)
. Synonyms (see or use references)
. Multi-meaning words (scope notes)
. False coordination (proximity indicators, phrase searching, role indicators)
. In controlled vocabularies, need for updating
. Not enough links between terms (use of “term clouds”, topic maps)
. Need for more structured documents, with sections separately indexed and retrievable
. Need for dialogue with user to clarify query (reference librarian, information officer, search intermediary, relevance feedback, online search aid)
ONLINE SEARCH AIDS

Software on local PC aiding search of online databases

Aimed to provide functions such as:

- A friendly user interface
- Simple linguistic analysis of queries
- Prompting of user to clarify queries
- Giving aid in selection of databases and hosts
- Transforming queries into search statements
- Providing means to evaluate output
- Giving aid in reformulation of queries
CURRENT COMPUTER SEARCH TOOLS

• Search engines based on full-text indexing
• Directories based on categorising
• Structured databases using DBMS
• Databases based on selective indexing using thesauri
• Systems based on selective indexing using taxonomies
• Systems based on social tagging
• More?
TREATMENT OF WORDS

Words from title, abstract or text

“Significant” words

Standard terms

Related terms

Linked or coordinated into “subjects”

Categories
TRADITIONAL METHOD

File-time
1. Construct description of DOC (or use DOC metadata)
2. Record location of DOC (e.g. shelf or file)
3. Determine subject(s) of DOC
4. Represent subjects as standard terms or categories
5. Store representations in file

Search-time
1. Receive user request (perhaps after interaction)
2. Formulate query in standard terms or categories
3. Select exact matches to query from file of representations
4. Display descriptions and locations of corresponding DOCs
5. Adjust query if user not satisfied
STATISTICAL METHOD

File-time
1. [Could use DOC metadata as description]
2. Record location of DOC (e.g. URL)
3. Extract all words (except stopwords) and their positions in DOC
4. Stem words and store words and positions in file

Search-time
1. Receive user request (perhaps after interaction)
2. Match stemmed request terms against word file, and rank matching DOCs according to probable relevance to query (assessed by statistical term weighting)
3. Display in ranked sequence snippets (terms in context) [could be metadata] and DOC locations.
4. If user unsatisfied, seek user relevance assessment of top-ranked DOCs, and adjust term weights and/or query terms
CONCLUSIONS FROM TREC
(Karen Spärck Jones, 2005)

Systems with automated indexing of full text, with statistically based term weighting, output ranking and relevance feedback, can deliver reasonable retrieval performance. They can do it as well as minimal human searching of files, though not as well as with expert human query development.

TREC has not been able to make significant, direct comparisons of this full automation with the “Boolean/thesaurus” approach.
1) What is the case for building classifications, thesauri and taxonomies? Is it related to the needs of “communities of practice”?

2) Are the benefits of controlled retrieval languages strong enough to justify the effort and cost of making and using them, and keeping them up-to-date?
QUESTIONS - 2

3) As multidisciplinary and international activity of every kind increases, is there a growing need to harmonise or match terminologies?

4) What is the future for “universal” controlled languages such as general classifications or general ontologies?

5) What needs to be done to improve the performance of full-text, statistical retrieval?
“Seeing how far ritual can be used in place of understanding” (Robert Fairthorne)

Knowledge of documents (what we need)

Understanding of users

Information skills

DIGITAL RETRIEVAL SYSTEMS

Search skills

Network skills

(what we have)
What are the outstanding problems in search?

PN: “In general, we think there are two aspects of it.
One is understanding user’s needs more.
The other is understanding the contents of documents, whether they be Web pages or video.”
RETRIEVER FINDS NEEDLE