Outline of today’s lecture

1. Seeking information
2. Thinking about retrieval
3. Controlled vocabulary and natural language
4. Relevance
5. Other approaches?

1. Information seeking

Another way to describe users of information systems is as information seekers.

Our information needs are inseparable from our:
- goals
- purposes
- objectives

1. Types of searches

Large and associates (2001: 35) argue that there are three common kinds of searches:
- Known item searches
- Specific item searches
- Subject searches

Each of these implies a particular approach to information seeking.

1. Different kinds of searches

Known item searches
- “Where is that website I saw last week?”

Specific item searches
- “How much guano was extracted from Nauru in 1963?”

Subject searches
- “I want to know more about classification”

1. Framing our search

In order to retrieve relevant information, we need to start by knowing something about:
- What we’re looking for (domain knowledge)
- How the relevant information system handles queries
1. Understanding the information retrieval system

Based on her study of online catalogue systems, Borgman (1996: 495) argues that users must possess three kinds of knowledge:

- how to turn ‘an information need into a searchable query’
- ‘how and when to use system features’
- ‘technical skills [required] in executing the query’

2. Thinking about retrieval

‘Information retrieval is a difficult problem because it requires describing information that you do not yet have’.

(Borgman 1996: 494)

A large part of retrieval is bound up with turning ‘an information need into a searchable query’

2. Defining information retrieval

One definition: ‘the art and science of searching’ ...

- ‘for information in documents’
- ‘for documents themselves’
- ‘for metadata which describes documents’
- ‘within databases ... for text, sound, images or data’

2. Purpose of an information retrieval system

- ‘To present `stored data items in a way that a user can – more or less conveniently – inspect the documents and assess their relevance in the given situation’
- ‘To provide a means to access those documents the user is (or might be) interested in.’

(Thiel & Muller 1996: 181)

2. Broad outlines of an IR system (Chowdhury 1999: 4)

2. A variety of information retrieval devices

- Different technologies (eg systems cards to search engines via CD-Roms)
- Different aids to retrieval (eg from abstracts to indexes)
- The challenge of non-textual information
2. Abstracts

‘A concise and accurate representation of the contents of a document, in a style similar to that of the original’

(Rowley 1992: 481)

- Commonly used as guides to academic and scientific papers
- Also usually printed together with the original document
- Often made available for database searching instead of full text

2. Advantages and disadvantages

Abstracts:
- Offer a concise summary of the document
- Can save time in information searching

But also:
- Can vary in quality, type, and detail
- May be a poor substitute when full text is required

2. Indexes

- Traditionally found at the back of certain kinds of books, providing a guide to the page location of contents
- Organised alphabetically, commonly by name and/or topic
- Lend themselves well to use with pre-existing classification schemes

2. Indexing - a business records approach

‘The process of establishing and applying terms or codes to particular records by which they may be retrieved. Appropriate allocation of indexing terms allows retrieval of records across classifications or categories’.


2. Retrieving non-textual information

A range of different approaches have been developed to label non-textual information for storage and retrieval

Green (1984) provides a hierarchical classification of pictures and slides
2. Non-textual information

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>030</td>
<td>People</td>
</tr>
<tr>
<td>030.1</td>
<td>Men</td>
</tr>
<tr>
<td>030.2</td>
<td>Women</td>
</tr>
<tr>
<td>030.3</td>
<td>Children</td>
</tr>
<tr>
<td>030.4</td>
<td>Groups of People</td>
</tr>
<tr>
<td>030.5</td>
<td>People Doing Things</td>
</tr>
<tr>
<td>030.6</td>
<td>People in Scenics</td>
</tr>
<tr>
<td>030.7</td>
<td>Nudes</td>
</tr>
<tr>
<td>030.8</td>
<td>Parades</td>
</tr>
<tr>
<td>030.9</td>
<td>Religious Observances</td>
</tr>
<tr>
<td>030.10</td>
<td>Others</td>
</tr>
</tbody>
</table>

Green (1984: 5)

3. Retrieval - language

Word searches allow us to retrieve relevant information from a variety of documents that hold text. Such searches commonly depend upon

- Controlled vocabulary

- Natural language

Sometimes, however, the two are combined.

3. Controlled vocabulary

**Definition:**

‘a limited set of terms that must be used to represent the subject matter of documents’

(in Chowdhury 1999: 118)

As we have seen, controlled vocabularies underpin many established classification systems

3. Controlled vocabulary (based on Harvey 1999: 197)

**Strengths**

- Eases burden of searching through
  - Controlling synonyms
  - Providing scope notes
  - Displaying broader/narrower/related terms
- Also retrieves related peripheral concepts
- Uses compound terms to overcome syntax problems

**Weaknesses**

- Highly expensive to index at same level as natural language, and high input costs
- Risk of human error: omissions, misunderstandings of source document
- Terms in thesaurus lag behind
- Lack of specificity
- Searcher needs to learn terminology

3. Natural language indexing

**Definition:**

‘An indexing language which is essentially the language of the documents being indexed.’

(Rowley 1992: 481)

Increasingly common for electronic retrieval, popularised in large part by Web search engines

3. Natural language (based on Harvey 1999: 197)

**Strengths**

- ‘High specificity gives precision’
  - Excellent for finding particular words
  - Potential for high recall
  - Uses ‘authors’ words
  - No lag in terminology
  - Searcher can use natural language
  - Low input costs

**Weaknesses**

- ‘Intellectual effort placed on searcher’
- Problems caused by synonyms
- Problems of syntax and false drops
- Exhaustivity may lead to loss of precision
3. Natural language – boolean search logic

Definition:

‘The use of the terms AND, OR and NOT in framing search statements.’

(Rowley 1992: 484)

Used in many databases, as well as popular search engines

---

3. Natural language – boolean searching

- AND

- NOT

- OR

---

3. Natural language – boolean searching

This method has faced a number of criticisms:

- Sometimes hard to formulate a query using AND, OR, NOT
- Hard to predict how many retrieved items will be needed
- Unable to rank retrieved items by relevance

(Chowdhury 1999: 161)

---

4. Retrieving relevant information

When we do find information, we still need to ask:

- Are there other relevant materials to be found? (recall)
- How much of what I’ve found so far is junk? (precision)

Nor is this helped by the likelihood that any definition of relevance is likely to be subjective.
4. Recall

‘the ratio of relevant items retrieved to the total number of relevant items in the database’ - A/B

\[
\text{All items in database} = \text{B}
\]

\[
\text{Retrieved relevant items} = \text{A}
\]

4. Precision

‘the ratio of relevant items retrieved to the total number of items retrieved’ - A/C

\[
\text{All retrieved items} = \text{C}
\]

\[
\text{Retrieved relevant items} = \text{A}
\]

4. Relationship between recall and precision

It is common to suggest an inverse relationship between the two categories.

‘A search designed to achieve high recall tends to retrieve many irrelevant matches too … a search designed to achieve high precision misses many relevant items’

(Borgman 2000: 122)

5. Other approaches?

‘Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path.’

(Borgman 2000: 122)

5. Hypertext as a retrieval mechanism?

With the emergence of the World Wide Web, many enthusiasts hoped that the kinds of document association made possible with hypertext would encourage new and better forms of information retrieval.

In certain ways, Internet search engines represent one way of trying to capitalise on this.
5. Search engines and retrieval – an example

Google utilises software called PageRank:

PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."


5. Metadata

- In the past few decades, metadata have emerged as important tools for information retrieval that ‘describe resources, indicate where they are located, and outline what is required to use them successfully’

(Chowdhury 1999: 434)

6. Further reading

U. Thiel & A. Muller (1996) ‘‘Why was this item retrieved?’’, New Ways to Explore Retrieval Results. in M. Aboul & A. Smeaton (eds.) Information Retrieval and Information Extraction, Boston: Kluwer Academic Publishers.