Introduction to Indexing

Lecture 5
CS 410/510
Information Retrieval on the Internet

Outline

- Basic concepts
- Manual indexing
- Automated indexing

2

Some terminology

- Indexing
  - Creation of a document representation that can be stored and retrieved in electronic form (can be done manually or automatically)
- Index
  - Collection of document representatives that can be used in the retrieval process
- Indexing term
  - Keyword, phrase or word extracted from the text that is used for indexing

3

Some terminology

- Concept
  - Mental model of an object or idea that is represented by one or more terms
- Indexing language
  - Entire collection of terms (assigned keywords, extracted text words or phrases) that can be used to index documents in a collection
    - e.g. all legal strings to form words
    - e.g. all the terms in a controlled vocabulary

4

Indexing alternatives

	Indexing Language	
Indexing Method	Uncontrolled vocabulary	Controlled vocabulary
Automatic	X	
Manual		X

X **most common combinations**

5

Purpose of indexing

- To represent the document to facilitate retrieval
- Two perspectives
  - Representation: represent the content, that is, what the document is about
  - Discrimination: characterize the document so that it can be distinguished from others
    - How is this document different?
    - To what information needs/queries might it be relevant?

6

Issues and challenges

- Relationship between words and concepts
  - Often not one-to-one
  - User information needs relate to concepts
  - Indexing terms are (usually) words
- Synonymy: different words, same meaning (hurry, rush)
- Homonymy – same word, different meanings (bark of tree, bark of dog)*
- Polysemy – same word, related but distinct meaning (opening a door; opening a book)*

*Krovetz and Croft. Lexical ambiguity and information retrieval. Transactions on Information Systems, 10, pp 115-141, 1992.

7

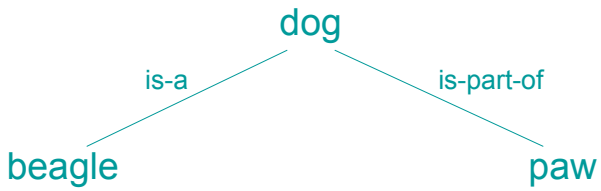
Effects of ambiguity

- Synonymy: mismatch of vocabulary in query and document degrades recall
  - Failure to retrieve documents that use synonyms instead of query term
- Homonymy/polysemy: mismatch in meaning degrades precision
  - Retrieval of documents that don't match the intended concept in the query

8

Issues and challenges

- Broader term/narrower term: mismatch in granularity of terms
  - Failure to retrieve documents if query uses different granularity for same concept
  - May reflect various hierarchical relationships



9

Issues and challenges

- Coordination: combining multiple concepts
  - logical AND
- Post-coordinate indexing
  - Index terms are simple terms
  - Combination occurs at time of **searching**
    - Query: *heart* AND *surgery*
- Pre-coordinate indexing
  - Index terms can represent complex concepts
  - Combination occurs at time of **indexing**
  - Many controlled vocabularies contain pre-coordinated terms
    - Query: *Heart Surgery* (a term in MeSH)

10

Coordination

- Pre-coordination *may* improve search precision by retrieving only documents that match the complex concept
  - Usually only available with controlled vocabulary indexing
- Post-coordination is more flexible

11

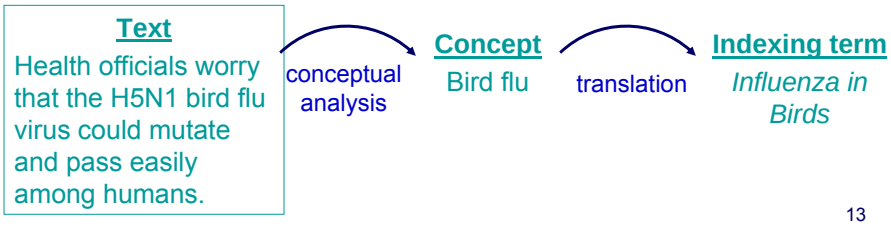
Manual indexing

- Usually by trained indexers
  - Read/scan document
  - Assign terms to represent the document
- Terms usually restricted to a controlled vocabulary (CV)
  - # terms assigned usually << # words in text
  - Are not necessarily words occurring in the text
- Typically used in bibliographic databases

12

Manual indexing process

- Conceptual analysis: determine what document is about, identify important concepts for indexing
- Translation: choose terms to represent concepts



13

Controlled vocabularies*

- Explicit list of terms
  - If same term used for multiple concepts, name must be qualified
  - If multiple terms used for same concept, one must be designated as preferred term and others listed as synonyms or aliases
- Types
  - List
  - Synonym ring
  - Taxonomy
  - Thesaurus

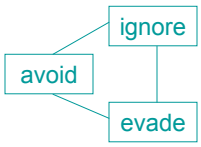
*ANSI/NISO Z39.19-2005 Guidelines for the Construction, Format and Management of Monolingual Controlled Vocabularies. http://www.niso.org/standards/resources/Z39-19-2005.pdf

14

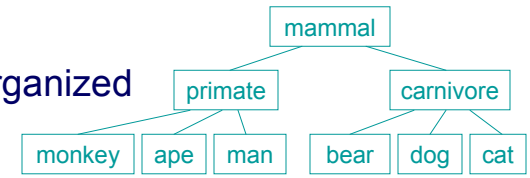
Controlled vocabularies

- List
  - Flat, unstructured list of terms

- Synonym ring
  - List of terms considered equivalent
  - Used for retrieval only, not indexing



- Taxonomy
  - Hierarchically organized



*ANSI/NISO Z39.19-2005 Guidelines for the Construction, Format and Management of Monolingual Controlled Vocabularies. http://www.niso.org/standards/resources/Z39-19-2005.pdf

15

Thesaurus

- Thesaurus relationships
  - Equivalence (Use/Used For)
    - Synonyms, lexical variants, near synonyms
  - Hierarchical (Broader Term/Narrower Term)
    - is-a, instance-of, part-of
  - Association (Related Term)
    - cause/effect, process/agent, action/product, action/target, object/property, etc.

*ANSI/NISO Z39.19-2005 Guidelines for the Construction, Format and Management of Monolingual Controlled Vocabularies. http://www.niso.org/standards/resources/Z39-19-2005.pdf

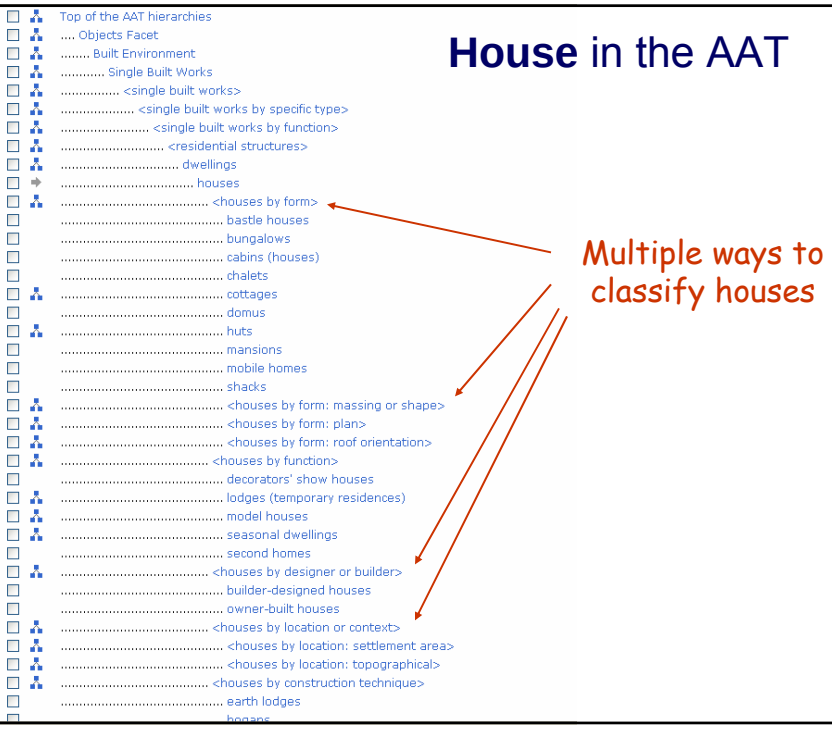
16

Faceted vocabulary

- Multiple hierarchies, generally orthogonal
- Art and Architecture Thesaurus (AAT)
  - Seven facets
    - Associated concepts
    - Physical attributes
    - Styles and periods
    - Agents
    - Activities
    - Materials
    - Objects

17

House in the AAT



18

Advantages of controlled vocabularies

- One “canonical” term represents groups of synonyms
- Terms typically distinguish among different senses of a word; e.g. in MeSH:
  - Common Cold
  - Cold (absence of warmth or heat)
  - Pulmonary Disease, Chronic Obstructive
    - For acronym COLD: Chronic Obstructive Lung Disease, aka emphysema
- Hierarchical relationships are explicit and can be exploited in search

19

Problems with CV indexing

- Exhaustivity
  - Too exhaustive → retrieval of unwanted documents (precision failure)
  - Insufficiently exhaustive → failure to retrieve wanted documents (recall failure)
- Specificity: generally desirable
  - Too specific → failure to retrieve wanted documents if query less specific
  - Insufficiently specific → retrieval of unwanted documents (precision failures)
- Consistency
  - Consistency ≠ quality, but
  - Consistency associated with better retrieval effectiveness

20

Disadvantages of CV indexing

- Usually requires manual indexing
  - Expensive
  - May be inconsistent
- Performance not clearly better in studies (compared to automatic indexing)
- Creation and maintenance of CV is expensive
- End user may be unfamiliar with CV
  - Mitigated by tools to browse CV, view definitions and scope notes
  - Mitigated by tools to automatically match queries to CV terms

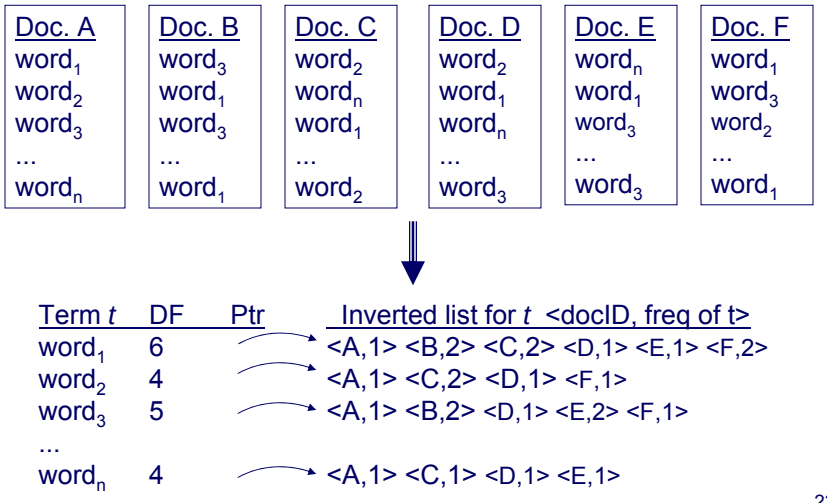
21

Automatic indexing

- Extracting words (and possibly phrases) from the text
  - After initial text preprocessing which may include stemming and stopword removal
- Storing words in data structure to enable retrieval
  - Usually includes recording term frequency and position

22

Inverted file (document-level)



Term *t*	DF	Ptr	Inverted list for *t* <docID, freq of t>
word$_1$	6		<A,1> <B,2> <C,2> <D,1> <E,1> <F,2>
word$_2$	4		<A,1> <C,2> <D,1> <F,1>
word$_3$	5		<A,1> <B,2> <D,1> <E,2> <F,1>
...			
word$_n$	4		<A,1> <C,1> <D,1> <E,1>

23

Inverted file

- Most common data structure for indexes
- Two main elements
  - Vocabulary:
    - List of all the terms
    - Frequencies for each term
    - Pointer to the inverted list for the term
  - Inverted lists:
    - Identifier of each document *d* containing term *t*
    - Frequency of *t* in *d*
    - May have a pointer to word-level inverted list that records position of each occurrence of *t* in *d*

24

Automated indexing

- Implementation of indexing data structures will be covered in next lecture

25

Next: Indexing structures

26