**Information Retrieval**

1. Course aims:
   1. Indexing: Representing the information content of documents through the use of e.g. stopword removal, stemming, and term weight calculation.
   2. Retrieval: Building models that select which information objects are relevant to a user's need. Models will include Boolean model, vector space model, probabilistic model, language model, inference network model, and relevance feedback model.
   3. Evaluation: Implementing and evaluating IR models, mainly with respect to effectiveness aspects.
   4. Tasks other than ad-hoc retrieval: e.g.: Classification, Summarisation
2. What will we cover?
   1. Organisation + Introduction
   2. Indexing and TF-IDF
   3. Retrieval Models I: VSM and BM25
   4. Retrieval Models II: Probabilistic IR, LM, DFR, Theory
   5. Retrieval Models III: Pagerank and others
   6. Retrieval Models IV: Relevance feedback
   7. Evaluation: Precision & Recall
   8. Semantic Search, DB+IR
   9. Classification and Summarisation
   10. Visual Information Retrieval
   11. Social Media Mining
3. Motivation and Applications
   1. There is too much unstructured “Data”
   2. Data is not the same as information   
        
      Data -> Information -> Knowledge
   3. Information is worthless if it cannot be found (extracted from raw data)
   4. Knowledge=Processed information retrieved from data
4. Introduction to Information Retrieval (IR)
   1. Terminology
   2. Information Need
   3. Retrieval Tasks
   4. A Conceptual Model for IR
   5. Document and Document Representation
   6. Queries
   7. Best-match retrieval
   8. History
   9. Topics in IR
   10. Information Retrieval vs Information Extraction vs Web Search
   11. Important forums (Conferences and Journals)
5. Databases vs. IR
   1. Format of data:
      1. DB: Structured data. Clear semantics based on a formal model.
      2. IR: Mostly unstructured. Free text.
   2. Queries:
      1. DB: Formal (like SQL)
      2. IR: often expressed in natural language (keywords search)
   3. Result:
      1. DB: exact result
      2. IR: Sometimes relevant, often not
6. Terminology
   1. General: Information Retrieval, Information Need, Query, Retrieval Model, Retrieval Engine, Search Engine, Relevance, Relevance Feedback, Evaluation, Information Seeking, Human-Computer-Interaction, Browsing, Interfaces, Ad-hoc Retrieval, Filtering.
   2. Related: Document Management, Knowledge Engineering.
   3. Expert: term frequency (TF), document frequency, inverse document frequency (IDF), vector-space model (VSM), probabilistic model, BM25 (Best-Match Version 25), DFR (Divergence from Randomness), page rank, stemming, precision, recall.
7. Information Need
   1. Example of an information need in the context of the world wide web:
      1. Find all documents (information!) about universities in the UK that
         1. offer master’s degrees in information Retrieval, and
         2. are registered with ACM SIGIR. The information (the document!)
      2. Should include full curriculum, fees, student campus, e-mail and other contact details.
   2. Formal representation of an information need = Query
8. Information Retrieval:
   1. An Informal Definition Representation, storage, organisation and access of information (information items, information objects, documents).
   2. Goal of an IR system
      1. Recall: Retrieve all relevant documents (e.g. legal)
      2. Retrieve as few non-relevant documents as possible.
      3. Precision: Retrieve the most relevant documents (e.g. web)
      4. Retrieve relevant documents before non-relevant documents.
9. Information Retrieval / Data Retrieval

|  |  |  |
| --- | --- | --- |
|  | Information Retrieval | Data Retrieval |
| Matching | vague | exact |
| Model | probabilistic | deterministic |
| Query language | natural | artificial |
| Query specification | incomplete | complete |
| Items wanted | relevant | all (matching) |
| Error handling | insensitive | sensitive |

1. What is IR?
   1. Goal: Find the documents most relevant to a given Query
   2. Dealing with notions of:
      1. Collection of documents
      2. Query (User’s information need)
      3. Notion of Relevancy
2. Types of Information Needs
   1. Retrospective (Ad-hoc Querying)
      1. “Searching the past”
      2. Different queries posed against a static collection
   2. Prospective (Filtering)
      1. “Searching the future”
      2. Static query posed against a dynamic collection
      3. Time dependent
3. Retrospective Searches (I)
   1. Topical search
      1. Open-ended exploration Identify positive accomplishments of the Hubble telescope since it was launched in 1991.
      2. Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.
   2. Open-ended exploration:
      1. Who makes the best chocolates?
      2. What technologies are available for digital reference desk services?
4. Retrospective Searches (II)
   1. Known-item search
   2. Question answering
5. Prospective “Searches” •
   1. Filtering
      1. Make a binary decision about each incoming document
   2. Routing / Multi-label Classification
      1. Sort incoming documents into different bins
6. What Types of Documents / Information / Media?
   1. Text (Documents)
   2. XML and structured documents
   3. Images
   4. Audio (sound effects, songs, etc.)
   5. Video
   6. Source code
   7. Applications/Web services
7. The Information Retrieval Cycle:  
     
   Source 🡪 (resource) 🡪 Query formulation 🡪 (query) 🡪 Search 🡪 (ranked list) 🡪 Selection 🡪 (documents) 🡪 Result 🡪 (query reformulation, relevance feedback) 🡪 Query formulation 🡪 …
8. The IR Black Box;
   1. Query 🡪 Representation 🡪 Query Representation 🡪 Comparison function 🡪 Results
   2. Documents 🡪 Representation 🡪 Document Representation 🡪 Index 🡪 Comparison Function 🡪 Results.
9. How do we represent documents?
   1. Remember: computers don’t “understand” anything!
   2. “Bag of words” representation:
      1. Break a document into words
      2. Disregard order, structure, meaning, etc. of the words
      3. Simple, yet effective!
10. Documents and Document Representations :
    1. Documents
       1. Unit of retrieval
       2. A passage of free text
          1. composed of text, strings of characters from an alphabet
          2. composed of natural language: newspaper articles, journal paper, dictionary definition, e-mail messages
          3. size of documents: arbitrary, newspaper article vs journal article vs e-mail
       3. Sub-document can also be a unit of retrieval (passage, XML element, answer to a question)
    2. Document Representation
       1. Free-text representation: extracted directly from text, good performance in broad domains.
       2. Controlled vocabulary representation: most concise representation, good performance in narrow domains with limited number of (expert) users.
       3. Full-text representation: most complete representation, optimal performance, huge resource requirements.
       4. Reduced (partial) content representation: stopwords, stemming, noun phrases, compression.
       5. Structure representation: chapter, section, paragraph.
       6. Semantic representation: actors, employees, workedWith.
11. Queries
    1. Information Need
    2. Simple queries
       1. composed of two or three, perhaps of dozen of keywords
       2. e.g. as in web retrieval
    3. Boolean queries
       1. ‘neural network AND speech recognition’
       2. e.g. as in online catalog and patent search
    4. Context queries
       1. proximity search, phrase queries
       2. e.g. ‘neural’ and ‘network’ distance at most 5 words
12. Best-Match Retrieval
    1. Compare the terms in a document and query
    2. Compute “similarity” between each document in the collection and the query based on the terms they have in common
    3. Sorting the document in order of decreasing similarity with the query
    4. The outputs are a ranked list and displayed to the user – the top ones are more relevant as judged by the system
13. Tasks of IR
    1. Index the documents in the collection (offline)
    2. Process the query
    3. Measure Similarity and compute ranking scores
       1. Find documents most closely matching the query (relevant documents)
    4. Display results
       1. E.g., user may refine the query (feedback)
14. Similarity Models
    1. Boolean model
    2. Vector-space model
    3. Probabilistic model
    4. Language modelling
15. Search Output
    1. What now?
       1. User identifies relevant documents for “delivery”
       2. User issues new query based on content of result set
    2. What can the system do?
       1. Assist the user to identify relevant documents
       2. Assist the user to identify potentially useful query terms
16. Selection Interfaces
    1. One dimensional lists
       1. What to display? title, source, date, summary, ratings, ...
       2. What order to display? similarity score, date, alphabetic, ...
       3. How much to display? number of hits
       4. Other aids? related terms, suggested queries, …
    2. Two+ dimensional displays
       1. Clustering, projection, contour maps, VR
       2. Navigation: jump, pan, zoom
17. Query Expansion / Enrichment
    1. Relevance feedback
       1. User designates “more like this” documents
       2. System adds terms from those documents to the query
    2. Manual reformulation
       1. Initial result set leads to better understanding of the problem domain
       2. New query better approximates information need
    3. Automatic query suggestion
18. Evaluating IR Systems
    1. User-centered strategy
       1. Recruit several users
       2. Observe each user working with one or more retrieval systems
       3. Measure which system works the “best”
    2. System-centered strategy
    3. Given documents, queries, and relevance judgments
    4. Try several variants of the retrieval method
    5. Measure which variant is more effective
19. Good Effectiveness Measures
    1. Capture some aspect of what the user wants
    2. Have predictive value for other situations
    3. Easily replicated by other researchers
    4. Easily compared
20. Topics in IR
    1. Retrieval models (ranking function, learning to rank, machine learning)
    2. Text processing (“Indexing”): NLP / understanding (language models)
    3. Interactivity and users
    4. Efficiency, compression, MapReduce, Scalability
    5. Distributed IR (data fusion, aggregated search, federated search)
    6. Multimedia: image, video, sound, speech
    7. Evaluation including crowd-sourcing
    8. Web retrieval and social media search
    9. Cross-lingual IR (FIRE), Structured Data (XML),
    10. Digital libraries, Enterprise Search, Legal IR, Patent Search, Genomics IR