Lecture 7
Going beyond search

Toine Bogers
Information Search, Retrieval &
Recommendation
September 24th

Outline
• cross-lingual retrieval
• multimedia retrieval
• enterprise search
• expert retrieval
• other retrieval tasks

Cross-lingual retrieval
• deals with retrieving information written in a language different from the user’s query language
• relies heavily on machine translation technology
  - parallel corpora of aligned phrases
  - lookup of most likely translation
  - topic of Text Mining/NLP course
• gaining in popularity the last decade
  - some TREC tracks
  - Cross-Language Evaluation Forum (CLEF)

TREC
• Text Retrieval Conference
  - ongoing series of workshop since 1992
  - sponsored by NIST
• large-scale evaluations of different problems
• different tracks/topics
  - blog track (explore information seeking using blogs)
  - enterprise track (search over organizational data)
  - cross-language track
  - spam track
  - legal track
  - genomics track (search gene sequences and supporting docs)
  - Terabyte track (test scalability of IR systems)
  - video track
  - ...
Cross-lingual retrieval

- two main approaches
  - query translation
    - translate the query into the language of the document
    - perform normal retrieval using the translated query
    - done at run time
    - hard because of ambiguity
    - users might not understand document language
  - document translation
    - translate all the documents into the required languages
    - perform normal retrieval using the original query on the translated collection
    - done at indexing time
    - good for unlocking other languages
    - index size explodes with each new language
Outline

• cross-lingual retrieval
  • multimedia retrieval
    – image retrieval
    – video retrieval
    – music retrieval
  • enterprise search
  • expert retrieval
  • other retrieval tasks

Image retrieval

• tasks
  – query by example
  – context-based image retrieval
  – content-based image retrieval
    • textual
    • visual
  – high-level feature extraction
  – automatic image annotation
    • greatly helped by Flickr and Google Image Labeler!

• evaluation forum
  – ImageCLEF
Video retrieval

- tasks
  - high-level feature extraction
  - object tracking & retrieval
Video retrieval

- **tasks**
  - high-level feature extraction
  - object tracking & retrieval
  - shot boundary detection
  - search

- **evaluation forum**
  - TRECVID, started in 2000 as regular TREC track

Speech retrieval?

- **searching speech data**
  - speech recognition
  - regular text retrieval

Music retrieval

- **tasks**
  - query by singing/humming
  - query by tapping
  - artist identification
  - artist similarity
  - genre classification
  - music mood classification
  - music similarity and retrieval
  - cover song identification
  - real-time audio to score alignment
  - melody extraction
  - beat extraction

- **evaluation forum**
  - MIREX
Music retrieval

Outline

• cross-lingual retrieval
• multimedia retrieval
• enterprise search
  – background
  – W3C collection
  – CERC collection
• expert retrieval
• other retrieval tasks

Background

• search through the electronic text content of an organization
  – company intranet
  – external websites
  – other electronic text
    • e-mail
    • database records
    • documents on fileshares
    • ...
• why? economic penalties because of
  – lost opportunities
  – lost productivity
  • 15–25% spent on non-productive information-related activities

Background

• huge divide between
  – safe, clean laboratory IR
  – IR in a real enterprise environment
• problems
  – different repositories
  – different access rights
  – different data formats
  – combining structured (relational) and unstructured data
  – scalability
W3C collection

- crawl of the w3.org sites in June 2004
  - 5.7 GB
  - 331,037 docs
  - used in TREC 2005 and 2006
- contents
  - regular Web pages
  - mailing lists
  - cooperative work sites
  - project group pages
- tasks
  - e-mail search
  - expert search

CERC collection

- crawl of the CSIRO sites in February/March 2007
  - 6+ GB
  - 370,000+ docs
  - used in the TREC 2007 track
- contains same variation in documents as W3C
- tasks
  - more realistic work tasks
    - creating an overview page
    - created by actual users
  - expert search
  - document search (different document types)

Outline

- cross-lingual retrieval
- multimedia retrieval
- enterprise search
- expert retrieval
  - tasks
  - evidence of expertise
  - motivation
  - problems
  - approaches
  - evaluation
- other retrieval tasks

Tasks
Tasks

- **expert finding**
  - find the experts on a specific topic

- **expert profiling**
  - find out what one expert knows about different topics
  - reverse of expert finding
  - need static set of topics for this!

- some differences with regular IR
  - usually a fixed set of topics is used
  - experts are entities that are not directly represented
    • have to approximate them somehow
  - aboutness is not expertise (although we assume it is)

Evidence of expertise

- e-mail or forum messages
- corporate communications
- shared folders in file systems
- resumes and homepages
- employee databases
- e-mail flow
- bibliographic information
- software library usage
- search and publication history
- project time charges

Example

Motivation

- **discovery**
  - rapidly locate individuals or communities of expertise to accelerate R&D
  - rapid formation of operational or proposal teams
  - support formation of cross-disciplinary teams to respond to new market threats and opportunities

- **assessment**
  - assess enterprise skill sets
  - enable identification of atrophy
  - discover new and emerging skill areas
  - predict effects of skill loss (attrition/retirement) or gain (mergers/acquisitions)

Source: Maybury (2006)
**Problems**

- expert retrieval can be difficult because
  - sparse data
  - skewed data
  - when is someone an expert on document X?
    - mentioned in X?
    - read X?
    - and how much of it, i.e. does author position matter?
  - are all documents created equal?
    - e-mail
    - notes
    - web pages
    - publications
    - course descriptions
    - project descriptions

**Candidate-centric approach**

**Document-centric approach**

**Evaluation**

- relatively new task (since 2005)
- test collections
  - fixed set of topics
  - fixed set of people
  - documents with some kind of personal association
  - extra metadata
- some example collections
  - W3C collection
  - CERC collection
  - UvT Expert Collection
  - new collections?
    - Wikipedia
    - CiteULike
Outline

- cross-lingual retrieval
- multimedia retrieval
- enterprise search
- expert retrieval
- other retrieval tasks

Other retrieval tasks

- some examples
  - source code retrieval
  - patent retrieval
  - geographic IR
  - ...

Questions?