Text-based (image) retrieval

Henning Müller

HES SO//Valais

Sierre, Switzerland

UNIVERSITÉ DE GENÈVE

Overview

- Difference of words and features
  - Weightings instead of distance measures
- Stemming and pre-treatment
- Approaches for multilingual retrieval
- Tools available on the web
  - Lucene, …

Text retrieval (of images)

- Started in the early 1960s … for images 1970s

- Not the main focus of this talk

- Text retrieval is old!!

  – Many techniques in image retrieval are taken from this domain (sometimes reinvented)

- It becomes clear that the combination of visual and textual retrieval has biggest potential

  – Good text retrieval engines exist in Open Source

Problems with annotation (of images)

- Many things are hard to express
  - Feelings, situations, … (what is scary?)
  - What is in the image, what is it about, what does it invoke?
- Annotation is never complete
  - Plus it depends on the goal of the annotation
- Many ways to say the same thing …
  - Synonyms, hyponyms, hypernyms, …
- Mistakes
  - Spelling errors, spelling differences (US vs. UK), weird abbreviations (particularly medical …)

Basics in text retrieval

- Started with boolean search of words in text
  - In combination with AND, OR, NOT
  - No ranking, rather finite list of corresponding documents

- Vector space model to have distance between search terms and documents
  - Each occurring word is a dimension, its difference in frequency can be measured
  - Overall frequency of words as importance for axis

Zipf distribution (wikipedia example)

- X- rank

- Y- number
  of occurrences
  of the word
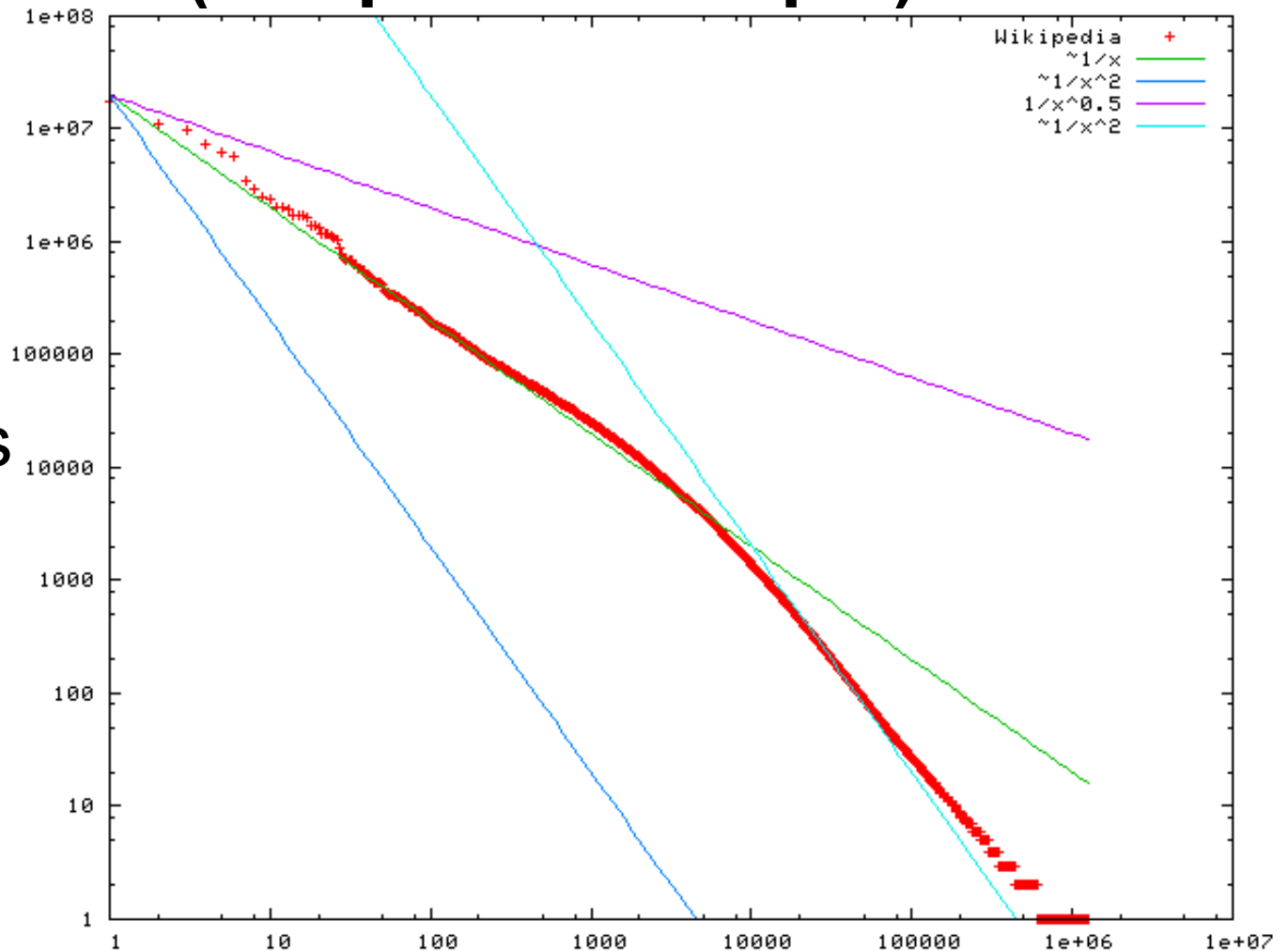
Principle ideas used in text IR

- Words follow basically a Zipf distribution
- Tf/idf weightings
  - A word frequent in a document describes it well
  - A word rare in a collection has a high discriminative power
  - Many variations of tf/idf (see also Salton/Buckley paper)
- Use of inverted files for quick query responses
  - Relevance feedback, query expansion, …

Techniques used in text retrieval

- Bag of words approach
  - Or N-grams can be used
- Stop words can be removed
- Stemming can improve results
- Named entity recognition
- Spelling correction (also umlauts, accents, …)
  - Google had a big success with this
- Mapping of text to a controlled vocabulary/ ontology

Stop word removal

- Very frequent words contain little information and can be removed
  - Automatically in Google et al.
- These words depend on the language
  - Stop word lists exist in many languages
    - Often 40-50% of texts
  - Contains also less frequent words not carrying information
- Or simply remove words above a certain frequency

Stemming - conflation

- Strongly dependent on the language
- Basically suffix stripping based on a set of rules
  - Cats, catty, catlike=cat as root or stem
- Can also create errors or slightly change meaning (errors often reported around ~5%)
- Porter stemmer for English is one of the most well known algorithms with a free implementation

Synonymy, polysemy

- Synonymy
  - Several words can say the same thing: car, automobile

- Polysemy
  - The same word can have several meanings

- Latent semantic Indexing (LSI)
  - Word cooccurences in the entire collection
  - Can reduce effects of synonyms

Query expansion vs. relevance feedback

- Most queries contain only very few keywords
- Add keywords to expand the original query
  - Can be automatic or manual
  - Semantically similar words, synonyms, discriminative words
- Often used in a similar way as relevance feedback but not with entire documents

Medical terminologies

- MeSH, UMLS are frequently used
  - Mapping of free text to terminologies
    - Quality for the first few is very high
  - Links between items can be used
    - Hyponyms, hypernyms, …
  - Several axes exist (anatomy, pathology, …)
    - This can be used for making a query more discriminative
- This can also be used for multilingual retrieval
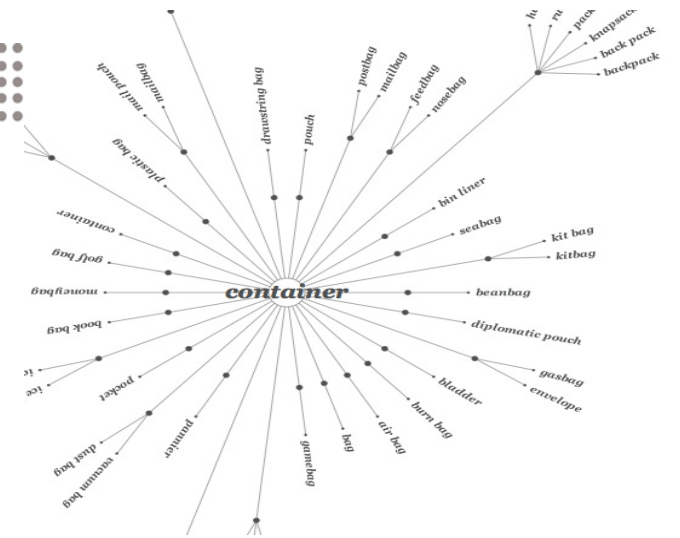
Wordnet

- Hierarchy, links, definitions in English language
  - Maintained in Princeton
- Car, auto, automobile, machine, motorcar
  - motor vehicle, automotive vehicle
    - vehicle
      - conveyance, transport
        » instrumentality, instrumentation
        » artifact, artefact
        » object, physical object
        » entity, something

Apache Lucene

- Open source text retrieval system
  - Written in Java
- Several tools available
  - Easy to use
- Used in many research projects and in industry
- Image retrieval plugin exists
  - LIRE (Lucene Image REtrieval)
  - Using simple MPEG-7 visual features

Multilingual retrieval

- Many collections are inherently multilingual
  - Web, FlickR, medical teaching files, …
- Translation resources exist on the web
  - TrebleCLEF has a survey of such resources in work
  - Translate query into document language
  - Translate documents into query language
  - Map documents and queries onto a common terminology of concepts
- We understand documents in other languages

Cross Language Evaluation Forum (CLEF)

- Forum to compare multilingual retrieval in a variety of domains
  - GeoCLEF
  - QA CLEF
  - Domain-specific CLEF
  - ...
- Proceedings are a very good start for multilingual techniques

Challenges in multi-linguality

- Language pairs have a strongly varying difficulty
  - Families of languages are easier for multilingual retrieval
- Resources available depend strongly on the languages used
  - English has many resources, German, Spanish and French quite a few but rare languages rather little

Multilingual tools

- Many translation tools are accessible on the web
  - Yahoo! Babel fish
  - www.reverso.net
  - Google translate
- Named entity recognition
- Word-sense disambiguation

Current challenges in text retrieval

- Many taken from the WWW or linked to it
- Analysis of link structures to obtain information on potential relevance
  – Also in companies, social platforms, …
- Question of diversity in results
  – You do not want to have the same results show up ten times on the top
- Retrieval in context (domain specific)
- Question answering

Diversity

Conclusions

- Text retrieval is the basis of image retrieval
  - Many techniques come from this domain
- Text has more semantics than visual features
  - But other problems as well
- Text and image features combined have biggest chances for success
  - Use text wherever available
- Multilinguality is an important issue as most of the web is very multilingual
  - And also a part of research

References

- G. Salton and C. Buckley, Term weighting approaches in automatic text retrieval, Information Processing and Management, 24(5):513--523, 1988.

- K. Sparck Jones and C. J. Van Rijsbergen, Progress in documentation, Journal of Documentation}, 32:59--75, 1976.

- J. J. Rocchio, Relevance feedback in information retrieval, The SMART Retrieval System, Experiments in Automatic Document Processing, pages 313--323.

- M. Braschler, C. Peters, Cross-Language Evaluation Forum: Objectives, Results, Achievements, Information Retrieval, 2004.

- J. Gobeill, H. Müller, P. Ruch, Translation by Text Categorization: Medical Image Retrieval in ImageCLEFmed 2006, Springer Lecture Notes in Computer Science (LNCS 4730), pages 706-710, 2007.