Introduction to

Information Retrieval

CS4611 Revision

Professor M. P. Schellekens

Assistant: Ang Gao

Information Retrieval

- **Boolean retrieval**
  - Build inverted index
  - Processing Boolean queries
    - Intersecting, Union
  - query processing order
- **The term vocabulary and postings lists**
  - Stemming: Reduce terms to their “roots” before indexing
  - Lemmatization: Reduce inflectional/variant forms to base form
  - Faster postings merges: skip pointers/skip lists

Information Retrieval

- **Dictionaries and tolerant retrieval**
  - Dictionary data structures for inverted indexes
    - Hashtables
    - Trees
  - Wildcard queries
    - Permuterm index
    - Bigram (k-gram) indexes
    - Processing wild-cards queries
  - Edit distance
    - Construct Levenshtein distance matrix

Information Retrieval

- **Index compression**
  - Why compression
    - Use less disk space (saves money)
    - Keep more stuff in memory (increases speed)
    - Increase speed of transferring data from disk to memory (again, increases speed)
  - Lossy vs. lossless compression
  - Heaps' law Zipf's law
  - Dictionary compression
    - The dictionary is small compared to the postings file.
    - But we want to keep it in memory.
    - Dictionary as a string
    - Dictionary as a string with blocking

Information Retrieval

- Postings compression
    - Store gaps instead of docIDs
    - Gap encoding
        - VB codes
        - Gamma codes
- **Scoring, term weighting and vector space model**
    - Why Ranked retrieval
    - Problem with Boolean search: Feast or famine
    - Jaccard coefficient, What’s wrong with Jaccard?

Information Retrieval

- tf-idf weighting:
  - The tf-idf weight of a term is the product of its tf weight and its idf weight. $\mathrm{w}_{t,d} = (1+\log_{10}\mathrm{tf}_{t,d}) \times \log_{10}(N/\mathrm{df}_t)$
- Ranked retrieval in the vector space model
  - Queries as vectors
    - Queries as vectors in the space
    - Rank documents according to their proximity to the query in this space
  - Cosine similarity between query and document

Dot product

Unit vectors

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \bullet \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Information Retrieval

- **Computing scores in a complete search system**
  - top K document retrieval
- **Evaluation in information retrieval**
  - Evaluation of unranked retrieval sets
    - Precision and recall
    - Precision/recall tradeoff
    - A combined measure: *F*
    - Accuracy, Why accuracy is a useless measure in IR
    - Why harmonic mean?
  - Assessing relevance
    - kappa statistic

Information Retrieval

- **Link analysis**
  - The web as a directed graph
  - Model behind PageRank: Random walk
  - Formalization of random walk: Markov chains
  - PageRank = long-term visit rate = steady state probability.
  - Teleporting
  - Link matrix, Transition probability matrix, Transition matrix with teleporting
  - Formalization of “visit”: Probability vector
  - How do we compute the steady state vector: Power method.
  - PageRank issues and how important is PageRank?

Information Retrieval

Questions ?