COM6115: Text Processing

Information Retrieval: Retrieval models — boolean approach

Mark Hepple

Department of Computer Science
University of Sheffield

Overview

- Definition of the information retrieval problem
- Approaches to document indexing
  - manual approaches
  - automatic approaches
- Automated retrieval models
  - boolean model
  - ranked retrieval methods (e.g. vector space model)
- Term manipulation:
  - stemming, stopwords, term weighting
- Web Search Ranking
- Evaluation

Bag-of-Words Approach

- Standard approach to representing documents (and queries) in IR:
  - ◇ record what words (terms) are present
  - ◇ usually, plus count of term in each document
- Ignores relations between words
  - ◇ i.e. of order, proximity, etc
  - ◇ e.g. rabbit eating = eating rabbit

- Such representations known as **bag of words** approaches
  - ◇ c.f. mathematical structure "bag"
    — like a set (i.e. unordered), but records a count for each element

Information Retrieval: Methods

- Boolean search:
  - binary decision: is document relevant or not?
  - presence of term is necessary and sufficient for match
  - boolean operators are set operations (AND, OR)
- Ranked algorithms:
  - frequency of document terms
  - not all search terms necessarily present in document
  - Incarnations:
    - **The vector space model (SMART, Salton et al, 1971)**
    - **The probabilistic model (OKAPI, Robertson/Spärck Jones, 1976)**
    - **Web search engines**

The Boolean model

- Approach: construct *complex search commands*, by
  - combining *basic* search terms (keywords)
  - using *boolean operators*
- *Boolean Operators*:
  - AND, OR, NOT, BUT, XOR (*exclusive* OR)
- E.g.:

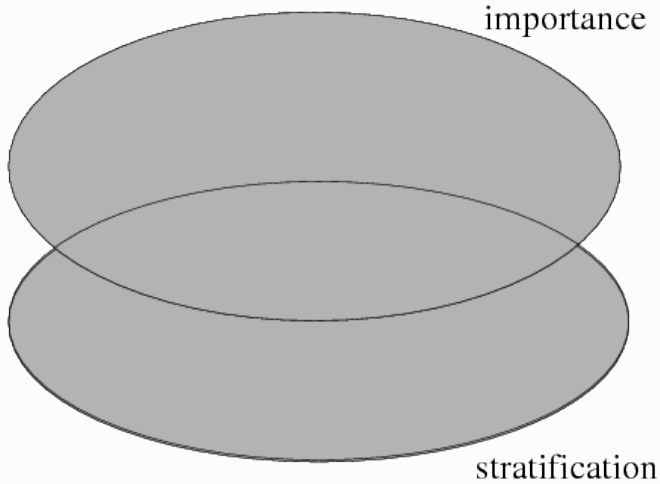
  `Monte-Carlo AND (importance OR stratification) BUT gambling`
- Boolean query provides a simple logical basis for deciding whether any document should be returned, based on:
  - whether basic terms of query do/do not appear in the document
  - the meaning of the logical operators

The Boolean model: set-theoretic interpretation

- Boolean operators have a **set-theoretic interpretation** for **efficient** retrieval
- Overall document collection forms maximal document set
- let $d(E)$ denote the document set for expression E
  - E either a basic term or boolean expression
- Boolean operators map to set-theoretic operations:
  - AND $\mapsto \cap$ (intersection): $d(E_1 \text{ AND } E_2) = d(E_1) \cap d(E_2)$
  - OR $\mapsto \cup$ (union): $d(E_1 \text{ OR } E_2) = d(E_1) \cup d(E_2)$
  - NOT $\mapsto {}^c$ (complement): $d(\text{NOT } E) = d(E)^c$
  - BUT $\mapsto -$ (difference): $d(E_1 \text{ BUT } E_2) = d(E_1) - d(E_2)$

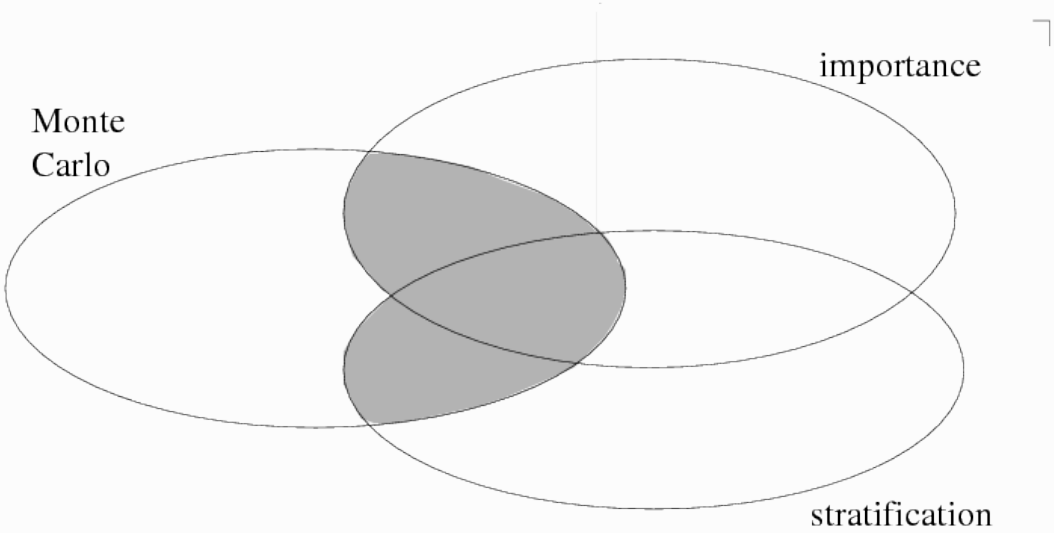
The Boolean model: set-theoretic interpretation (contd)

E.g. `Monte-Carlo AND (importance OR stratification) BUT gambling`



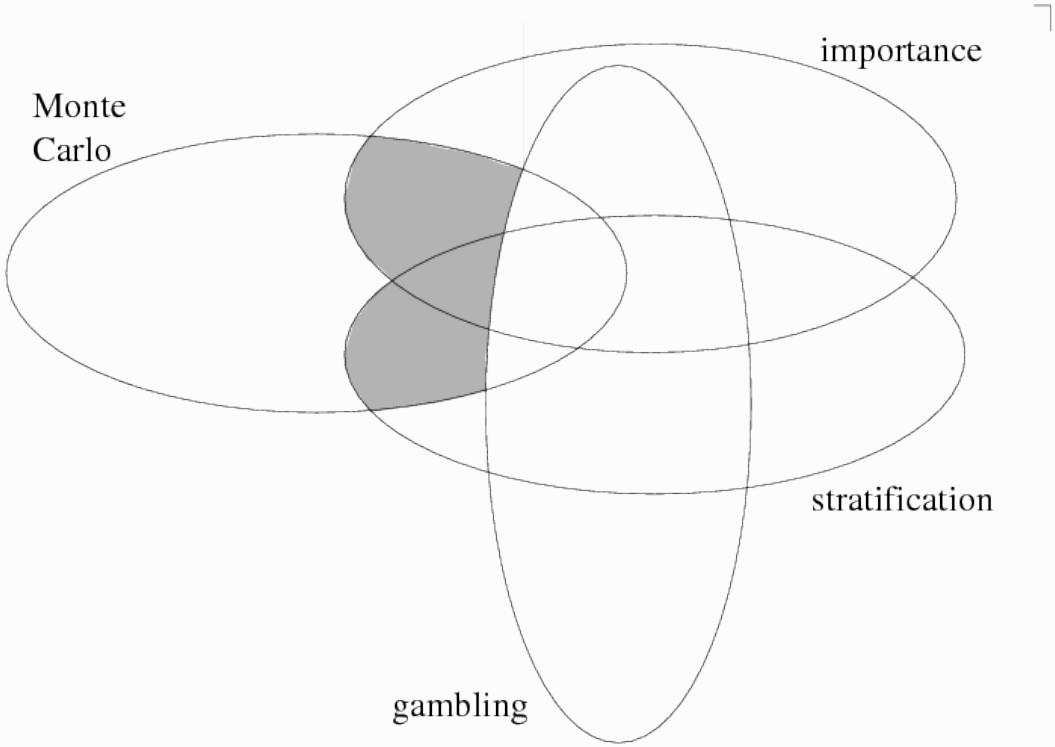
The Boolean model: set-theoretic interpretation (contd)

E.g. Monte-Carlo AND (importance OR stratification) BUT gambling



The Boolean model: set-theoretic interpretation (contd)

E.g. Monte-Carlo AND (importance OR stratification) BUT gambling



Boolean Queries: Complexity

- Question: **Magnetic resonance imaging, magnetic resonance arthrography and ultrasonography for assessing rotator cuff tears in people with shoulder pain for whom surgery is being considered**

- Query: `((Ultrasonography [mh] OR ultrasound [tw] OR ultrasonograph* [tw] OR sonograp*[tw] OR us [sh]) OR (Magnetic Resonance Imaging [mh] OR MR imag*[tw] OR magnetic resonance imag* [tw] OR MRI [tw])) AND (Rotator Cuff [mh] OR rotator cuff* [tw] OR musculotendinous cuff* [tw] OR subscapularis [tw] OR supraspinatus [tw] OR infraspinatus OR teres minor [tw]) AND (Rupture [mh:noexp] OR tear* [tw] OR torn [tw] OR thickness [tw] OR lesion* [tw] OR ruptur* [tw] OR injur* [tw])`

From Lenza, M., Buchbinder, R., Takwoingi, Y., Johnston, R. V., Hanchard, N. C., & Faloppa, F. (2013). Magnetic resonance imaging, magnetic resonance arthrography and ultrasonography for assessing rotator cuff tears in people with shoulder pain for whom surgery is being considered. The Cochrane Library.

The Boolean model: summary

- Documents either match or don't match
  - ◇ Expert knowledge needed to create high-precision queries → OK for expert users
  - ◇ Often used by bibliographic search engines (library)
- Not good for the majority of users
  - ◇ Most users not familiar with writing Boolean queries → not natural
  - ◇ Most users don't want to wade through lists of 1000s unranked results → unless very specific search in small collections
  - ◇ This is particularly true of web search → large set of docs