COM6115: Text Processing

Information Retrieval: Web Search Ranking

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

Web Search Ranking

- Web docs contain info beyond their mere *"textual content"*
  - ◇ *state-of-the-art web search* engines, like Google, exploit this
  - ◇ achieve *much more effective* retrieval than could without it
- HTML contains clues that some terms are *more important*
  - e.g. terms in regions marked as title or headings
  - e.g. terms *emphasised by formatting*: bold / bigger / colour
    - ◇ can use clever term weighting schemes, that add weight to such terms
- Link text — commonly provide *description of target* doc
  - ◇ often a better description than doc provides of *itself*
  - e.g. "Hey, here's a great intro to calculus for beginners – check it out!"
    - ◇ Google treats link text as *part of* target doc
- Link structure of web *more broadly*
  - ◇ if page A *points to* page B, implies B is worth looking at
  - ◇ can be used as a measure of *authority / quality*

Exploiting Link Structure: the PageRank Algorithm

- Key method to exploit link structure of web: PageRank algorithm
  - ◇ named after its inventor: Larry Page (co-founder of Google)
  - ◇ assigns a score to each page on web: its *PageRank score*
    - can be seen to represent the page's *authority* (or *quality*)
- PageRank algorithm — key idea:
  - ◇ link from page A to page B confers **authority** on B
  - ◇ *how much* authority is conferred depends on:
    - the authority (PageRank score) of A, and its number of *out-going links*

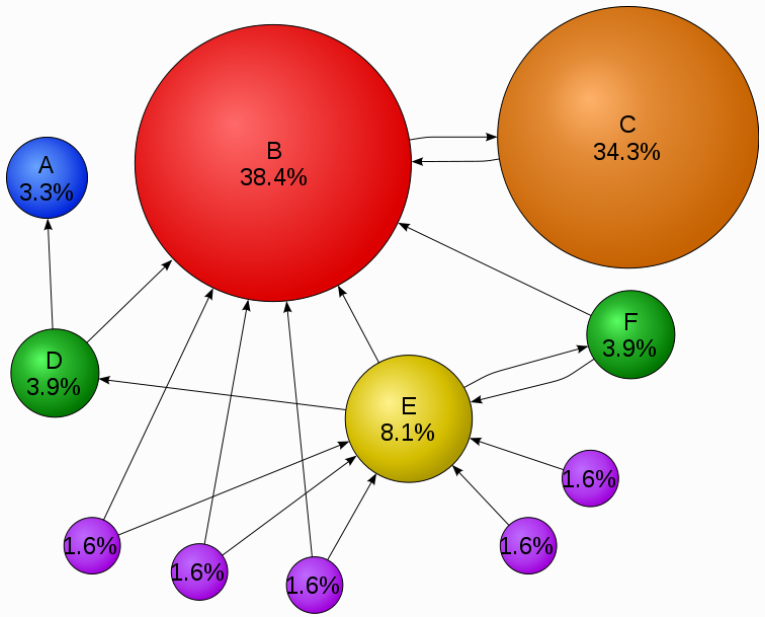
      i.e. A's authority is *shared out* amongst its out-going links
  - ◇ note that this measure is *recursively defined*

    i.e. score of any page depends on score of every other page
- PageRank scores have an alternative interpretation:
  - ◇ probability that a *random surfer* will visit that page

    i.e. one who starts at a random page, clicks randomly-chosen links forward, then (getting bored) jumps to a new random page, and so on ...

Exploiting Link Structure: the PageRank algorithm (ctd)

- Graphical intuition:



- During retrieval, rank score of doc *d* is a *weighted combination* of:
  - ◇ its PageRank score: a measure of its authority
  - ◇ its IR-Score: how well *d* matches the query *q*, based on
    - Vector Space model, TF.IDF, *up-weighting* of important terms, etc