Information Retrieval (IR)

The field of information retrieval deals with the representation, storage, organization of, access to information items.

Search Engines: Search a large collection of documents to find the ones that satisfy an information need, e,g, find relevant documents

- Indexing
- Query representation
- Document representation
- Retrieval Model: compare documents with query
- Evaluation/feedback

IR basics

- Indexing
  - Manual indexing: using controlled vocabularies, e,g, libraries, early version of Yahoo
  - Automatic indexing: indexing program assigns keywords, phrases or other features, e.g. words from text of document
- Popular Retrieval Models
  - **Boolean**: exact match, query using “and or not”, results not ranked.
  - **Vector Space**: best match
  - **Citation analysis models**: best match,
    - frequency, pattern, graphs of citations in articles and books, e.g. CiteSeer, Google Scholar
  - **Probabilistic models**: best match
    - e.g. Naïve Bayes Classifier, BM25
  - **Deep neural network models**

Vector Space Model

Any text object can be represented by a term vector.

Example
Doc1: 0.3, 0.1, 0.4
Doc2: 0.8, 0.5, 0.6

Query: 0.0, 0.2,0.0

Similarity is determined by distance in a vector space.

Vector Space Similarity: Cosine of the angle between the two vectors

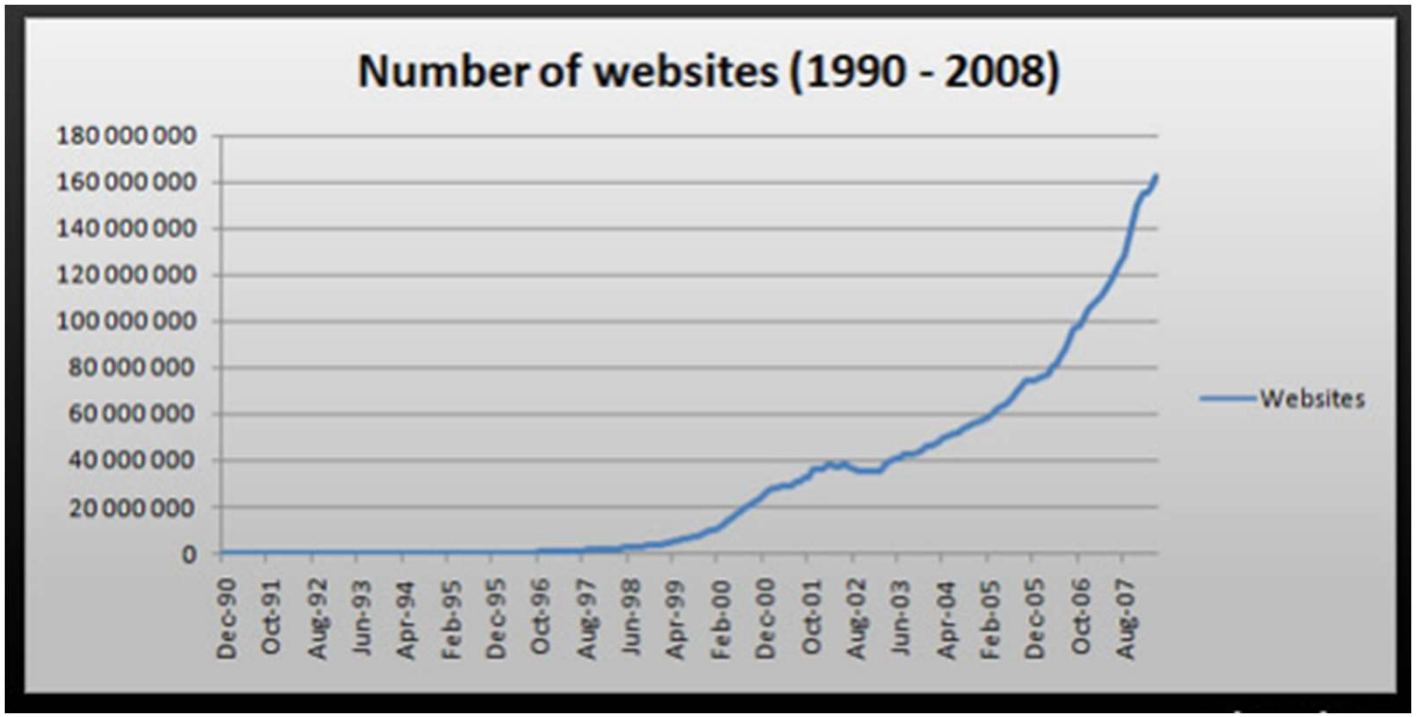
$$\frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}}$$

Google

- Started by 2 students, in a garage
- Released as Google in 1998
- Changed the way people use the Internet
- Designed to handle the expansion of the WWW

Sergey Brin & Lawrence Page

Growth of the Internet



Goals of Google

Accurate Searches

- Search Engines of the time unable to find themselves
- Number of documents matching queries was rapidly increasing
- Humans only interested in the first 10 or so results
- Need some way to recognise better matches

- What is the ground breaking improvement?

Features of Google

PageRank

- Uses citation (link) graph of the web
- Ranking the page
  - Estimate relevance of search results
  - Bring order to the web
    - `PR(A) = (1-d) + d (PR(T1)/C(T1) + … + PR(Tn)/C(Tn))`
- Modelled on human behaviour - Random Surfer

PageRank basics:

- Goal: bring order to the Web, every page has a number

- Measure of relative importance of web pages based on citation analysis
  - Backlinks act like a kind of peer review
  - A collaborative notion of authority or trust
- Link structure of the web
  - Each page is a node, each link is an edge in the graph
  - In general, highly linked pages are more important
  - But simple backlink counts are not accurate

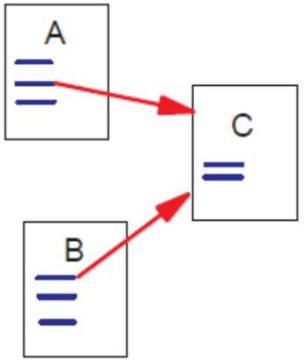


Figure 1: A and B are Backlinks

- PageRank improves on this by considering the importance of the backlinks to the page

PageRank simplified example

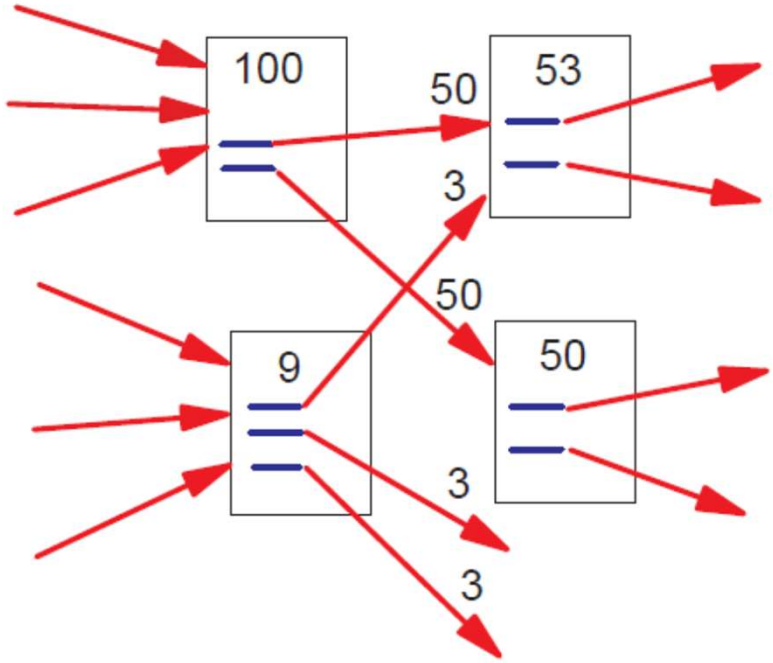


Figure 2: Simplified PageRank Calculation

PageRank Simplified Definition

- PR: *a page has high rank if the sum of the ranks of its backlinks is high*

- u – a webpage
- F_u – {pages u points to}, B_u – {pages pointing to u}
- $N_u = |F_u|$ - number of links from u
- c – factor of normalization (so that total rank of all web pages is constant)
- $R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$

Ranks updated iteratively, until converge

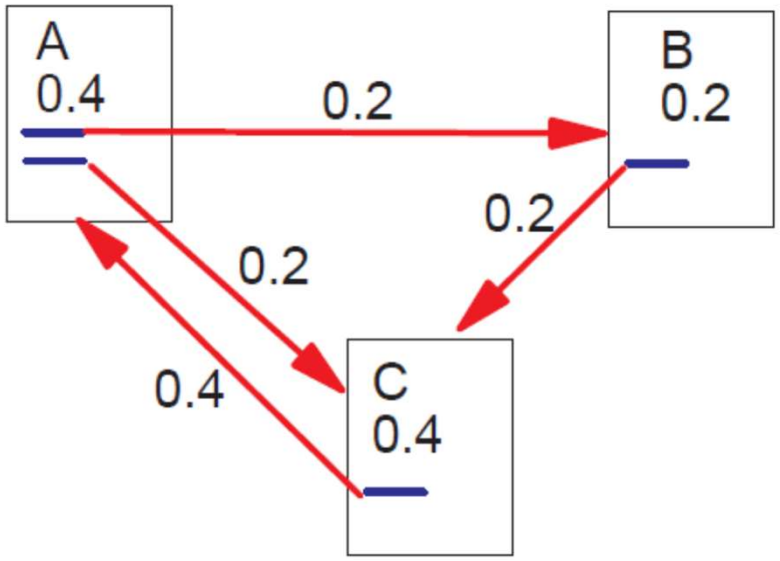
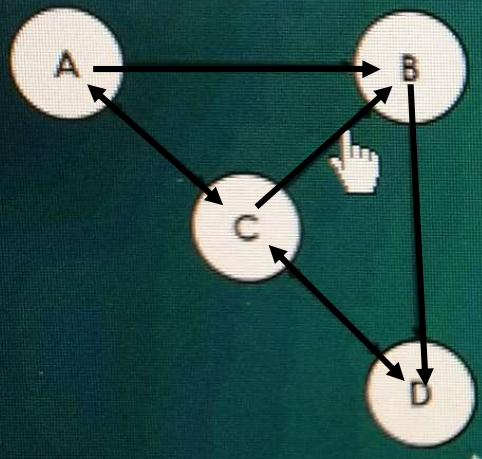


Figure 3: Simplified PageRank Calculation

A simple example

PageRank algorithm



	Iteration 0	Iteration 1	Iteration 2	PageRank
A	1/4	1/12	1.5/12	1
B	1/4	2.5/12	2/12	2
C	1/4	4.5/12	4.5/12	4
D	1/4	4/12	4/12	3

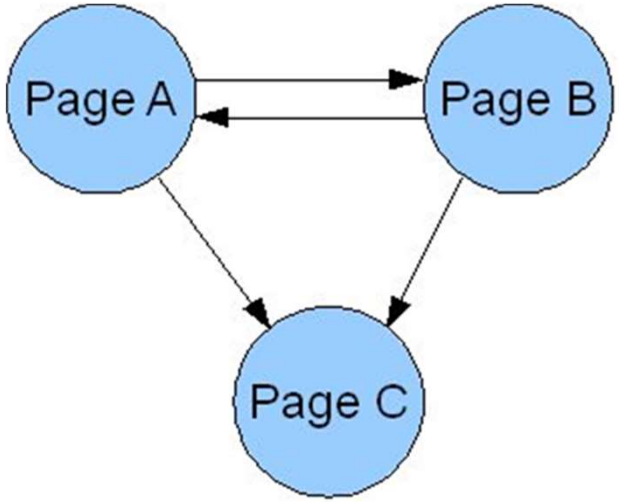
Limitations?

- Any things that may break the system?

Dangling Links

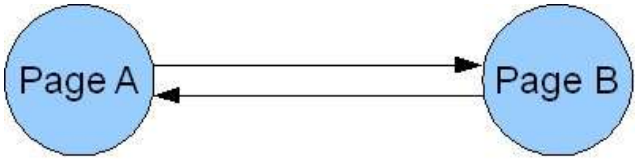
- Links pointing to pages with no outgoing links.
- Affect the PageRank model since it is not clear where their weights should be distributed.
- Dangling links can be pages that have not been downloaded yet.

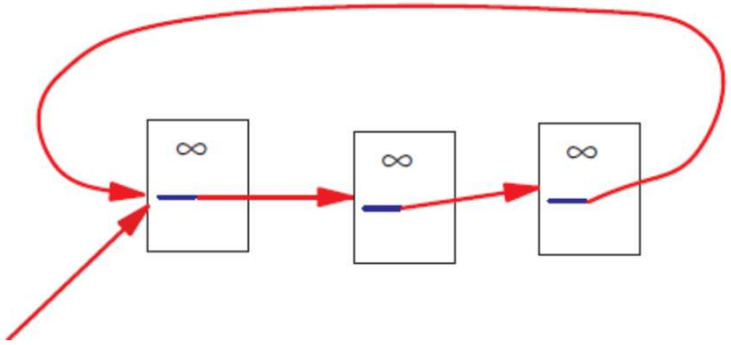
- Solution: Remove dangling links from the system until all PRs are calculated.



Rank Sinks

- Equation is recursive – iterate until it converges
- Rank sink – 2 pages pointing to each other but no other page
- Solution: Introducing a rank source E(u)
  - $R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + c\, E(u)$
- *E(u) is some vector over the web pages that corresponds to a source of rank*





Discussion

- How to get to the first page?

Manipulation by Commercial Interests

- Many features are hard to implement because of possibility of manipulation – like higher ranking for frequently updated pages – this can be abused.

- PR is virtually immune because for any page to get a high PR, it must convince a high PR page or *a lot* of low PR pages to link to it.
- But this can cost a lot of money