Information Retrieval

What is IR

- The process of obtaining relevant information from a large repository (e.g., documents, web pages, databases).
- **Documents**: Textual data (web pages, articles, books)
    - Multimedia & Maps
- **Queries**: User input expressing information need.
- **Indexing**: Creating efficient data structures for fast retrieval.
- **Ranking**: Ordering results by relevance.

Applications

- Search engines
- Document retrieval
- Recommender system
- …

BM25

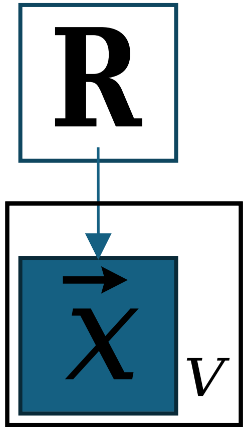
- probabilistic retrieval model based on the Bag-of-Words assumption
- Probability Ranking Principle
  - : document
  - : relevance; {1, 0}
- Interested only in ranking

Binary Independence Model (BIM)

- Given query

Naïve Bayes

- : binary representation of document (Term-existence)
- Naive Bayes conditional independence assumption: presence/absence of a word in a document is independent of the presence/absence of any other word



Binary Independence Model (BIM)

- Assumption:
    - Terms not in query does not impact relevance

Retrieval Status Value

- Retrieval Status Value
- If assume RSVIDF

BM25

- Best match 25
- Words are drawn independently from the vocabulary using a multinomial distribution
- Distribution of term frequencies (tf) follows a binomial distribution – approximated by a Poisson
- Assume that term frequencies in a document follow a Poisson distribution

Poisson Distribution

- Models the probability of the number of events occurring in a fixed interval of time/space, with known average rate
- Examples
  - Number of cars arriving at the toll booth per minute
  - Number of typos on a page
- Also be used to approximate binomial
- Assume that term frequencies in a document follow a Poisson distribution
  - Implies fixed document length
  - Reasonable fit for “general” words, but poor for the topic-specific words

Extensions

- Term is either regular or topic related
- Extend the Poisson as mixture of Poisson
  - Use a simple function to approximate:
  - controls term frequency scaling
- Document length normalization
  - b: a parameter between 0 and 1. 0 means no length normalization

BM25

- Normalize term frequency using document length

- Ranking function

Google

- Google was founded on September 4, 1998, by Larry Page and Sergey Brin.
- Google began in January 1996 as a research project by Larry Page and Sergey Brin while they were both PhD students at Stanford

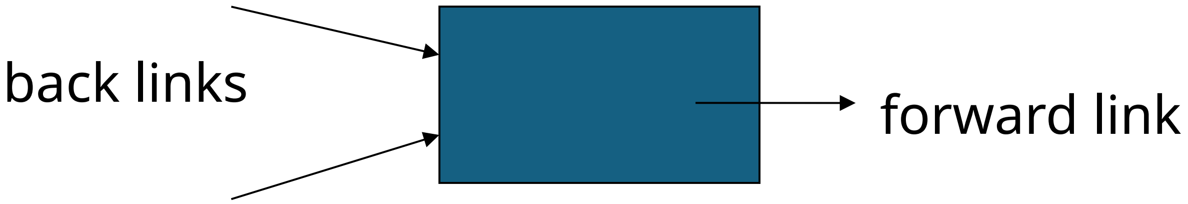
- Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer networks and ISDN systems* 30.1-7 (1998): 107-117.

Ranking

- Webpage corpus is not a controlled collection
    - BM25/tf-idf only consider relevance
    - Reputation?
- Hit rate
    - Rank higher if it is visited more frequently
    - Fake hits
    - Cold start for new pages
- Citation
    - A paper is important if it is cited by many papers
    - Not well-controlled

PageRank

- A page with many links to it is more likely to be useful than one with few links to it
  - Just like citation
- The links from a page that itself is the target of many links are likely to be particularly important
  - This is something new

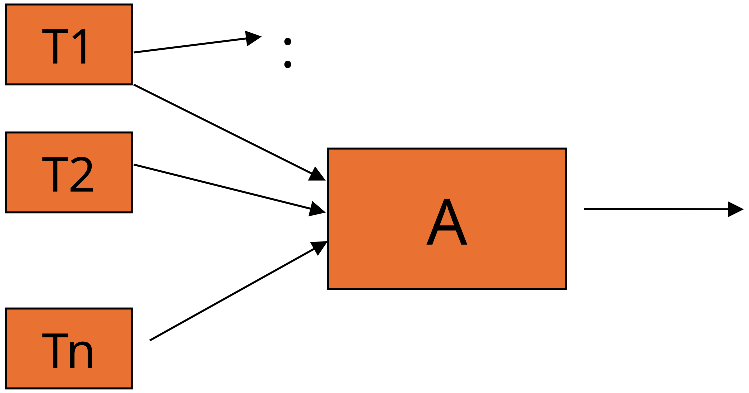


Each link has different weight

PageRank

- Each page is ranked using a value called PageRank (PR)
- A page’s PR depends on the PRs of its back link pages

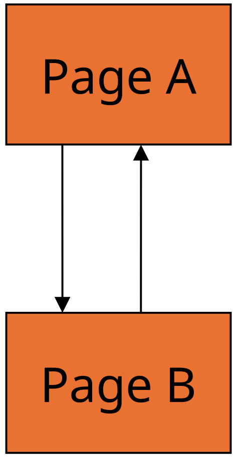
  - : damping factor, normally this is set to 0.85
  - : pages linking to page A
  - : PageRank of page A
  - : PageRank of page pointing to page A
  - : the number of links going out of page



PageRank Example

- Assign each page an initial rank value
  - Could be any number (seed)
- Repeat calculations until converge

Seed = 40
1)
PR(A)= 0.15 + 0.85 * 40 = 34.25
PR(B)= 0.15 + 0.85 * 0.385875 = 29.1775
2)
PR(A)= 0.15 + 0.85 * 29.1775 = 24.950875
PR(B)= 0.15 + 0.85 * 24.950875 = 21.35824375
3) ……



Seed = 0
1)
PR(A)= 0.15 + 0.85 * 0 = 0.15
PR(B)= 0.15 + 0.85 * 0.15 = 0.2775
2)
PR(A)= 0.15 + 0.85 * 0.2775 = 0.385875
PR(B)= 0.15 + 0.85 * 0.385875 = 0.47799375
3)
PR(A)= 0.15 + 0.85 * 0.47799375 = 0.5562946875
PR(B)= 0.15 + 0.85 * 0.5562946875 = 0.622850484375

d= 0.85
PR(A)= (1 – d) + d(PR(B)/1)
PR(B)= (1 – d) + d(PR(A)/1)

Supervised Ranking Methods

- Binary classification

- Learning to rank
  - Pointwise (regression problem)
    - : predicting the real-value or ordinal score of x
    -
  - **Pairwise**
    - : classification problem for a given pair
    - Usually implemented with a scoring function:

  - Listwise

Training loss

- max-margin loss

$$\mathcal{L} = \max(0, \gamma - \Phi(m, e^{+}) + \Phi(m, e^{-})),$$

ranking score

- Cross-entropy loss

$$\mathcal{L} = -(y \log \Phi(m, e) + (1 - y) \log (1 - \Phi(m, e))),$$