# Informational Retrieval on the Web

Mei Kobayashi, Koichi Takeda
IBM Research Japan (2000)

Presenter: Vahid Karimi

# Table of Contents

# Historical Background

- Historical background [Schatz 1997]

- Grand visions
  - Vannevar Bush (1945)
    - Memex
    - Systems for information manipulation
  - Licklider (1961, 1962)
    - Libraries of the future
    - Procognitive systems

# Historical background (Continued)

- Text search (syntax search)
  - Roughly mid 1960s to mid 1980s
  - Bibliographical search
    - Pioneered by medicine
  - Abstract databases
  - Full text
    - Pioneered by law

- Techniques for text search
  - Fundamentally the same as 30 years ago
  - Scope expanded
  - EX:  Inverted index file, stemming
  - From words to phrases (from biblio to full documents)
    - Proximities on inverted index file

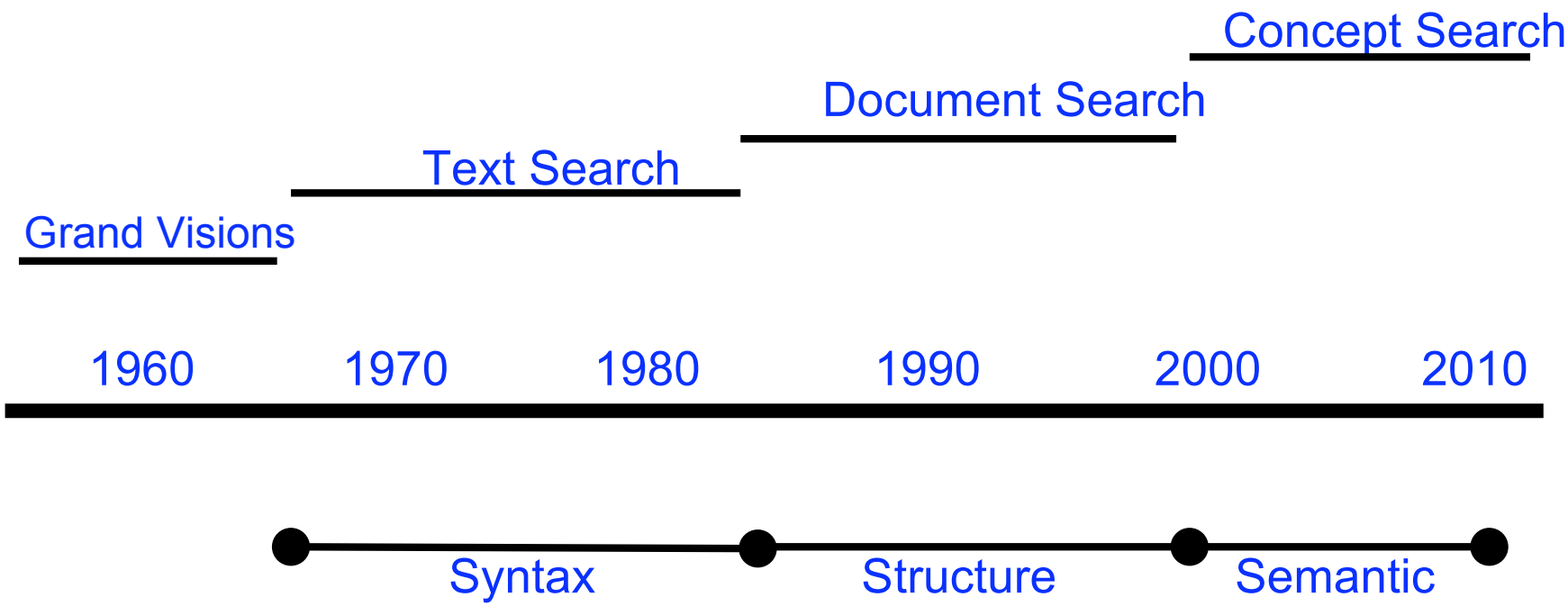
# Historical Background (Continued)

- Document search (structure search)
  - Roughly mid 1980s to 2000
  - Mainframe to distributed workstations
  - Multimedia retrieval
  - Telesophy (wisdom at a distance)
    - At Bell 1985-1986
  - Search on the Internet

- Concept search (Vocabulary Switching)
  - Roughly from 2000

# Historical Background (Continued)

- The following courtesy of [Schatz 1997]



# Classical Information retrieval and Search Engines

- Similarity a lot
- One major difference
  - Collections are not given to search engines
  - Search engines have to find them (Crawling)

- Challenges
  - Dynamic
  - Open and closed web
  - Spamming

# Evaluation of Search Engines

- One measure speed, precision, recall
  Precision = number of relevant documents/ total number of documents retrieved
  Recall = number of relevant documents retrieved / total number of relevant documents

- Ideally precision, and recall must be equal to 1
  - Add disjunctive terms
    - Recall goes up
    - Precision suffers

- Another measure
  - Same as above, but on the first few pages

- Indexing by inverted file (quantity, quality)

# Tools for web-based retrieval and ranking

- Indexing
  - Automatic
  - Manual or human based
  - Using of metadata
- Hyperlink analysis [Henzinger 2001]
  - Mirrored Hosts
  - Web Page Categorization
  - Geographical scope
- Crawling
  - Next week presentations
- Clustering
  - Organizing large databases

# Classification of Ranking Techniques

- Text based models [Dhillon, Fan, Guan 2001][Lee, Chuang, Seamons 1997][Berry 1996]
  - Premise
  - Boolean models
  - Similarity models
    - Vector space model
    - Latent semantic index model
- Link based models
  - Query independent ranking
  - Query dependent ranking

# Text Based Models

- Similarity models
  - Measures the similarity between a document and a query
  - Hence the naming
- Vector space model by Salton
  - A similarity model
  - Represents each document as a vector space
  - Its dimension depends on the document terms (vocabularies)
  - Terms have associated weights
  - To represent the value of the terms

# Vector Space Model

- How it works?
  - Extract all terms ignoring cases
  - Get rid of stop words (a, an, the)
  - Count the number of terms in each document
  - Use heuristics or other algorithms to eliminate low and high frequency words
  - After the above operations, we identified 1 to w terms (words) and 1 to d documents

# Vector Space Model (Continued)

- Then, we need to weigh terms

- Different weighing measures
  - Term frequency weighing
  - $w_{ij} = tf_{i,j} * idf_j$
  - $Tf_{i,,j}$ captures how often a term (j) occurs in a document (i)
  - $idf_j$ captures how often j occurs in the entire collection

# Vector Space Model (Continued)

- Similarity between a query (q) and a document (i):
  - Sim ( Q,Di) = $\Sigma_{j=1}^{v} w_{q,i} \cdot w_{i.j} / (\Sigma_{j=1}^{v} w_{q,i}^2 \cdot \Sigma_{j=1}^{v} w_{i.j}^2 )^{1/2}$
  - Values between 0 and 1
    - The closer the document and the query, the closer to 1
    - Clustering (document and document)
  - The denomination is for normalization so that
  - Two documents one containing (x,x,y,y,z,z)
  - Another containing (x,y,z) gets the same weighing
  - Good or not!

# Vector Space Model (Continued)

- Vector Space model
  - Conceptual
    - Since a document vector is sparse and long
  - Inverted index file

# Latent Semantic Index Model

- Singular value decomposition (SVD)
- Any m * n matrix A can be factored as:
  - $A = V \Sigma U^T$

Where V is an m * m matrix

U is an n * n matrix

$\Sigma$ is of a special form m * n

$$\Sigma = \begin{vmatrix} D & 0 \\ . \ . & \\ & \\ 0 & 0 \end{vmatrix}$$

$$D = \begin{vmatrix} \alpha_1 & 0 \\ . & \\ 0 & \alpha_k \end{vmatrix}$$

$\alpha_1, \ldots \alpha_k$ are all positive real num

& $\alpha_1 \geq \alpha_2 \geq \ldots\ldots \geq \alpha_k \geq 0$

& called the singular value of A

# Latent Semantic Index Model (Continued)

- Choosing k is difficult
- Topic of factor analysis
- But by choosing k, the matrix A (term – document)
- Transforms to $A_k$
- Dimensions are reduced using SVD
- Same operation on matrix B (term – query)
- Using SVD, B transforms to $B_q$
- Then the similarity of $A_K$ to $B_q$ is measured
- An example refer to:
- [Deerwester, Dumais, Furnas, Landauer, Harshman 1990]

# Linked Based Models

- Link based models [Henzinger 2001]
  - Premise (one or both)
    - Recommendation
    - Same topic
  - All major search engines
    - claim to use some form of hyperlink analysis
    - No details

- Query independent models

  1) Carriere, Kazman model (1997)
  2) PageRank (by Brin and Page 1998)
  3) WLRank (Weighted Link Rank] [Baeza, Davis 2004]
  4) Absorbing model by Amati et al. 2003 [Baeza 2005]
  5) Network flow model by Tomlin 2003 [Baeza 2005]

# Query Independent Models

- Query independent models
- Concept

- 1) Carriere, Kazman model (1997)

- 2) PageRank (by Brin and Page, 1998)
  - R(A) = € / n + (1 - €) * Σ R(B) / outdegree (B)
    - A,B ε G
    - € is a constant, usually between 0.1 and 0.2
    - n is the number of nodes (web pages) in G
    - Outdegree B = number of hyperlinks on page B

# Query Independent Models (Continued)

- PageRank model (continued)
  - Hugh set of linear equations
  - Google
  - Based on random surfer model
  - € damping factor, leaving the page

- 3) WLRank (Weighted Link Rank) model [Baeza, Davis 2004]
  - A variant of PageRank
  - Introduced some attributes to give more weights to some links
  - Claimed that precision improved

- 4) Absorbing model by Amati et al.2003 [Baeza 2005]
- 5) Network flow model by Tomlin 2003 [Baeza 2005]

# Query Dependent Models

- Query dependent models
- Concept

1) Carriere and Kazman (1997) neighbourhood graph

2) HITS (hyper-linked induced topic search) (by Kleinberg 1998)

3) Topic Sensitive PageRank (by Haveliwala 2002) [Baeza 2005]

# Query Dependent Models (Continued)

1) Carrier and Kazman model (1997)

- Builds a query-specific graph (neighbourhood graph) as follows:
  - Step 1: Uses a search engine to retrieve results for a query
    - These are root nodes (every document is a node)
  - Step 2: Adds nodes that linked to root nodes in the neighbourhood graph

    Adds nodes that root nodes are linked to in this neighbourhood graph
  - Step 3: Uses either indegree technique to rank neighbourhood graph or PageRank to rank neighbourhood nodes

# Query Dependent Models (Continued)

2) HITS (hyper-linked induced topic search) by Kleinberg 1998

- Based on identifying authority and hub pages
- Using a neighbourhood graph
- An iterative algorithm
- Authorities and hubs converge
- No bound on that
- In practice, converge quickly
- Not used by any search engine
- Topic drifting

# Query Dependent Models

3) Topic Sensitive PageRank by Haveliwala 2002 [Baeza 2005]
   - Use PageRank to rank pages based on ranking at index time
   - At the query time, assign new ranking to pre-ranked topic sensitive

# Comments

- ...

# References


- [Baeza, Davis 2004] R. Baeza, E. Davis, Web Page Ranking Using Link Attributes, 13th international World Wide Web conference, 2004.

- [Baeza 2005] R. Baeza-Yates and C. Castillo, Web Search, *Encyclopedia of Language and Linguistics*, to appear in 2005.

- [Berry 1996] Large Scale Information Retrieval with Latent Semantic Indexing, http://www.cs.utk.edu/~berry/lsi++/node3.html.

- [Deerwester, Dumais, Furnas, Landauer, Harshman 1990] S. Deerwester, S.Dumais, T. Furnas, T. Landauer, R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, 1990.

- [Dhillon, Fan, Guan 2001] *I.S. Dhillon, J. Fan and Y. Guan*, Efficient Clustering of Very Large Document Collections. [ps.gz, pdf] invited book chapter in *Data Mining for Scientific and Engineering Applications*, pages 357-381, Kluwer, 2001.

- [Henzinger 2001] M. Henzinger, Hyperlink analysis for the web, IEEE 2001.

# References (Continued)


- [Kobayashi, Takeda 2000] M. Kobayashi, K. Takeda, Information Retrieval on the Web, ACM Computing Surveys, 32(2), 2000.

- [Lee, Chuang, Seamons 1997] D. Lee, H. Chuang, K. Seamons, Document Ranking and the Vector-Space Model, IEEE, 1997.

- [Schatz 1997] Schatz, Bruce R., Information retrieval in digital libraries: Bringing search to the net, Science, Vol. 275, 1997.