Link Analysis

Hongning Wang
CS@UVa

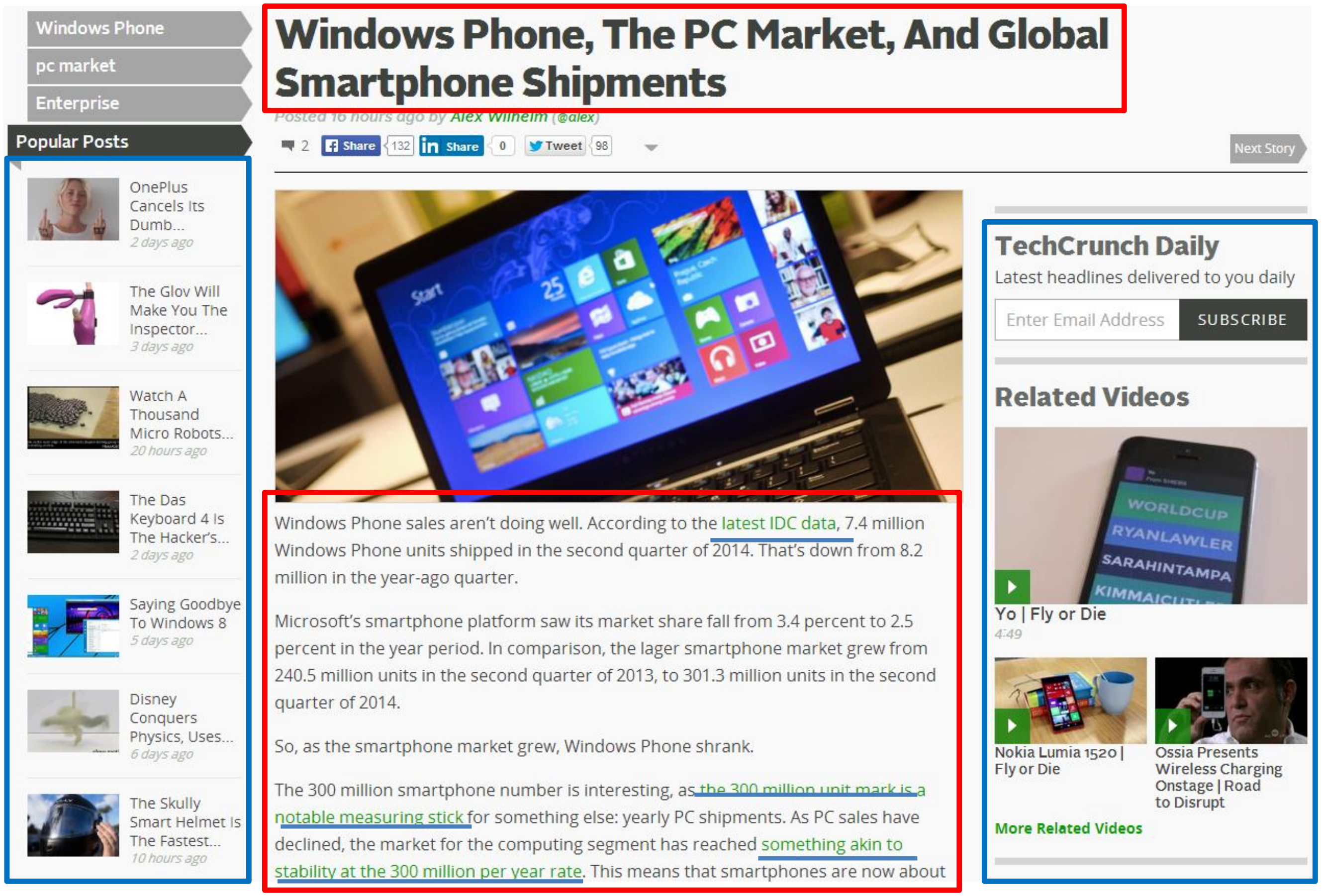
Structured v.s. unstructured data

- Our claim before
  - IR v.s. DB = unstructured data v.s. structured data
- As a result, we have assumed
  - Document = a sequence of words
  - Query = a short document
  - Corpus = a set of documents

However, this assumption is not accurate…

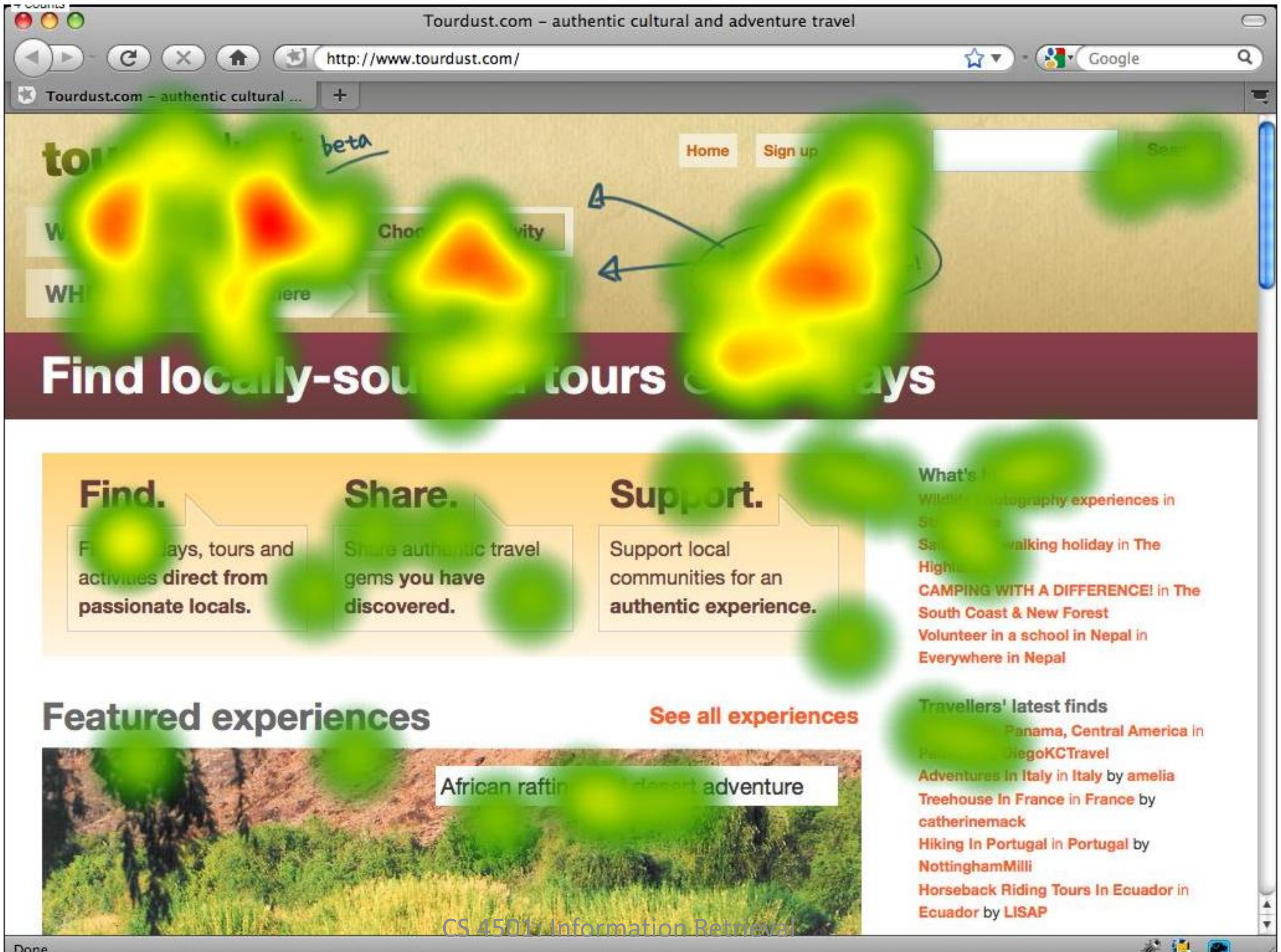
A typical web document has

Title

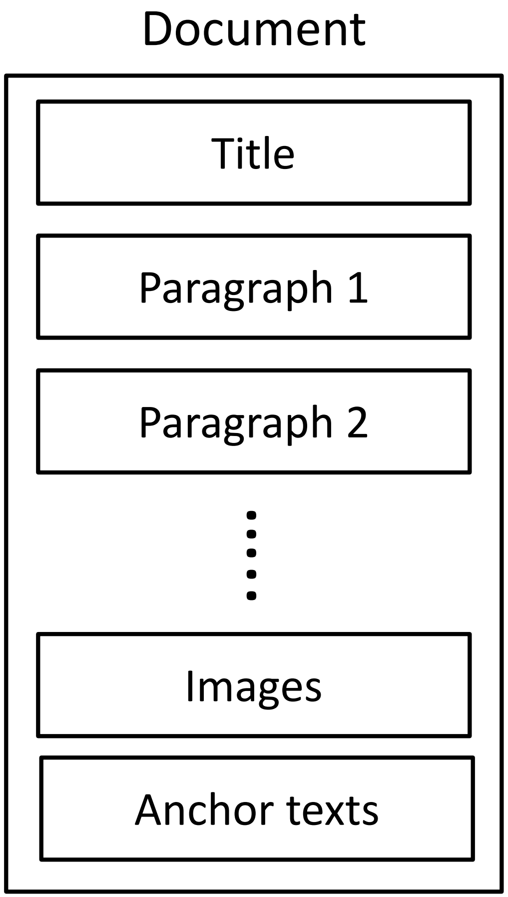


Anchor **Body**

How does a human perceive a document's structure



Intra-document structures



Concise summary of the document

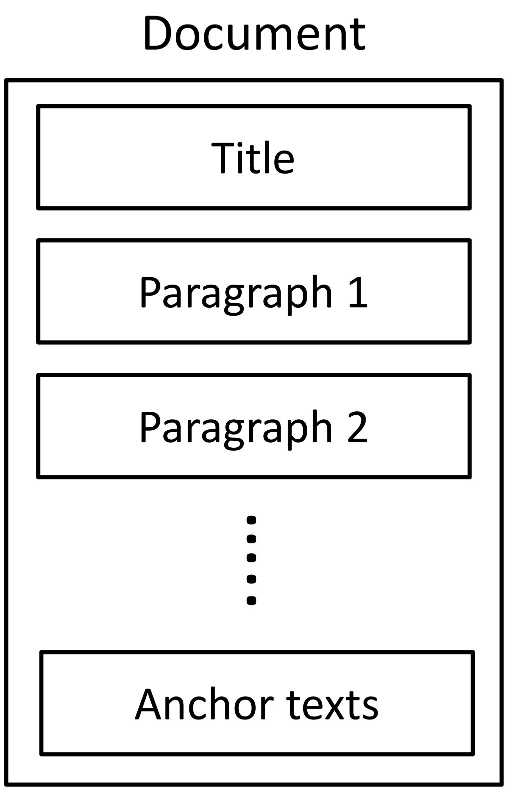
Likely to be an abstract of the document

They might contribute differently for a document's relevance!

Visual description of the document

References to other documents

Exploring intra-document structures for retrieval



Intuitively, we want to give different weights to the parts to reflect their importance

Think about query-likelihood model…

Select D_j and generate a query word using D_j

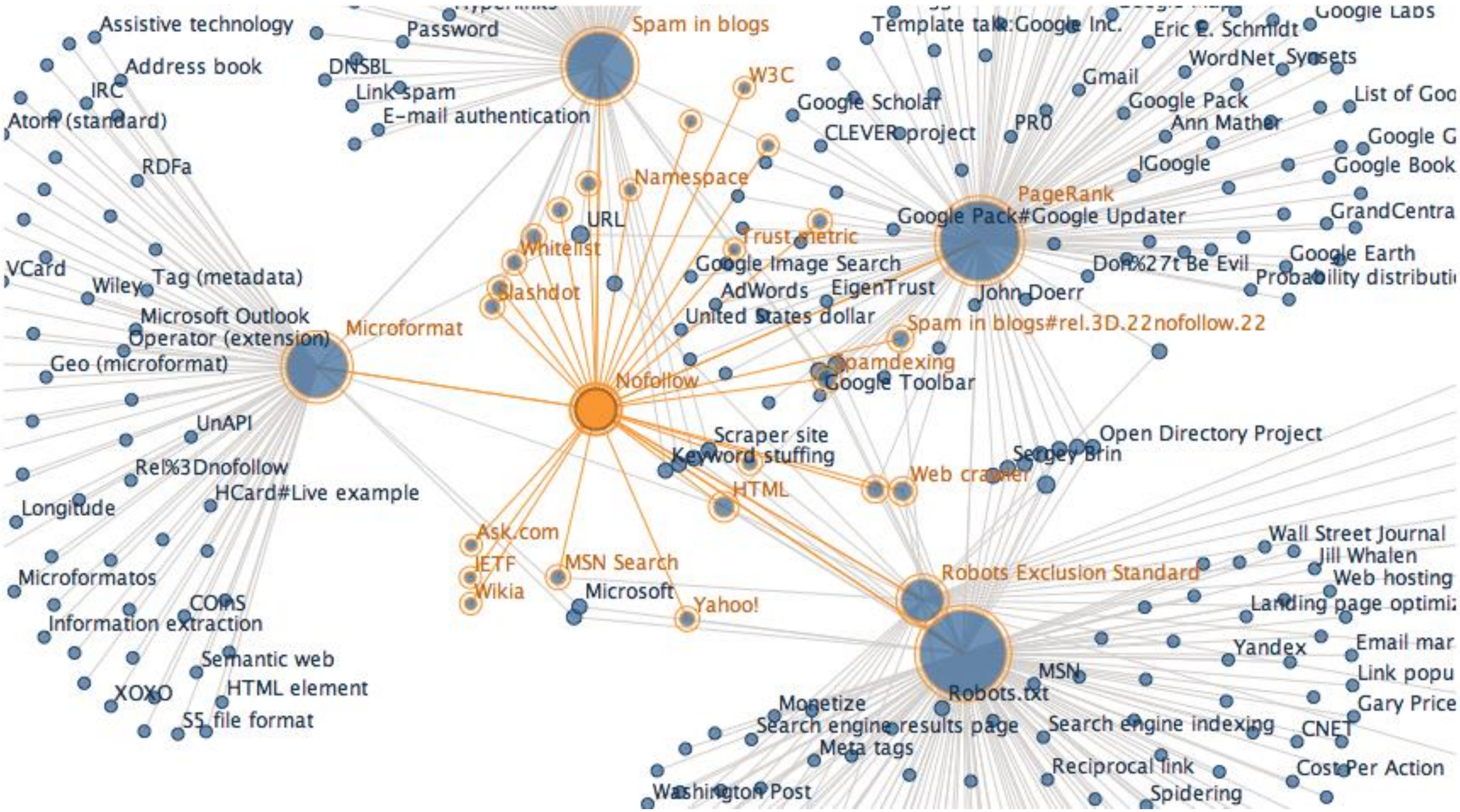
$$p(Q|D,R)=\prod_{i=1}^{n} p(w_i|D,R)$$

$$=\prod_{i=1}^{n}\sum_{j=1}^{k} \underline{s(D_j|D,R)}\, p(w_i|D_j,R)$$

"part selection" prob. Serves as weight for D_j
Can be estimated by EM or manually set

Inter-document structure

- Documents are no longer independent



Source: *https://wiki.digitalmethods.net/Dmi/WikipediaAnalysis*

What do the links tell us?

- Anchor
  - Rendered form

Barack Hussein Obama II (US /bəˈrɑːk huːˈseɪn ɵˈbɑːmə/, UK /ˈbæræk huːˈseɪn ɵˈbɑːmə/; born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the *Harvard Law Review*. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for t Illinois Senate career of Barack Obama sentatives in 2000.

  - Original form

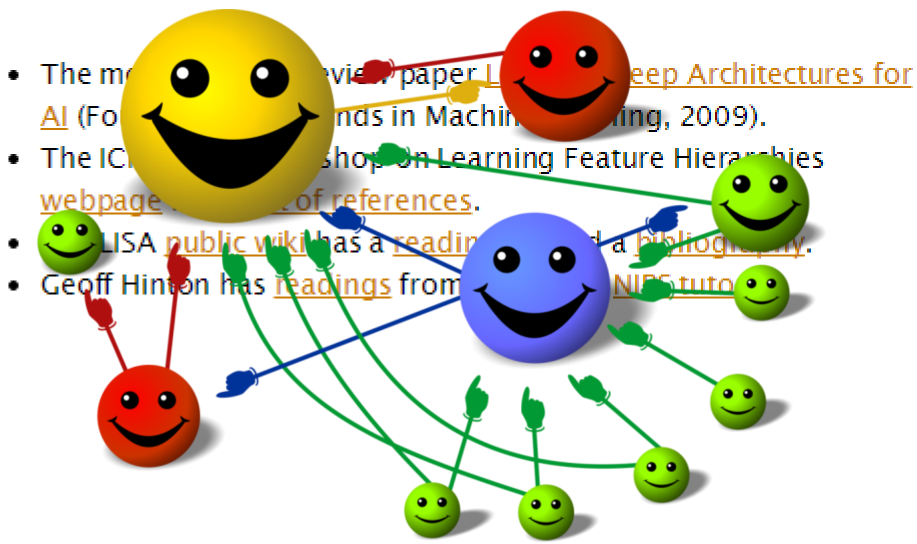
```
" from 1992 to 2004. He "
<a href="/wiki/Illinois_Senate_career_of_Barack_Obama" title="Illinois Senate career of Barack
Obama">served three terms</a>
" representing the 13th District in the "
```

What do the links tell us?

- Anchor text
  - How others describe the page
    - E.g., “big blue” is a nick name of IBM, but never found on IBM’s official web site
  - A good source for query expansion, or directly put into index

What do the links tell us?

- Linkage relation
  - Endorsement from others – utility of the page



"PageRank-hi-res". Licensed under Creative Commons Attribution-Share Alike 2.5 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:PageRank-hi-res.png#mediaviewer/File:PageRank-hi-res.png

Analogy to citation network

- Authors cite others' work because
  - A conferral of authority
    - They appreciate the intellectual value in that paper
  - There is certain relationship between the papers
- Bibliometrics
  - A citation is a vote for the usefulness of that paper
  - Citation count indicates the quality of the paper
    - E.g., # of in-links

Situation becomes more complicated in the web environment

- Adding a hyperlink costs almost nothing
  - Taken advantage by web spammers
    - Large volume of machine-generated pages to artificially increase “in-links” of the target page
    - Fake or invisible links
- We should not only consider the count of in-links, but the quality of each in-link
  - PageRank
  - HITS

Link structure analysis

- Describes the characteristic of network structure
- Reflect the utility of the web document in a general sense
- An important factor when ranking documents
  - For learning-to-rank
  - For focused crawling

Recall how we do internet browsing

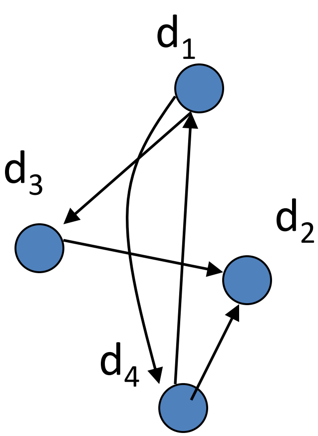
1. Mike types a URL address in his Chrome's URL bar;
2. He browses the content of the page, and follows the link he is interested in;
3. When he feels the current page is not interesting or there is no link to follow, he types another URL and starts browsing from there;
4. He repeats 2 and 3 until he is tired or satisfied with this browsing activity

PageRank

- A random surfing model of internet
  1. A surfer begins at a random page on the web and starts random walk on the graph
  2. On current page, the surfer uniformly follows an out-link to the next page
  3. When there is no out-link, the surfer uniformly jumps to a page from the whole page
  4. Keep doing Step 2 and 3 forever

PageRank

- A measure of web page popularity
  - Probability of a random surfer who arrives at this web page
  - Only depends on the linkage structure of web pages



Transition matrix

$$M = \begin{pmatrix} 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \end{pmatrix}$$

$$p_t(d) = \alpha M p_{t-1}(d) + \frac{(1-\alpha)}{N} p_{t-1}(d)$$

α: probability of random jump

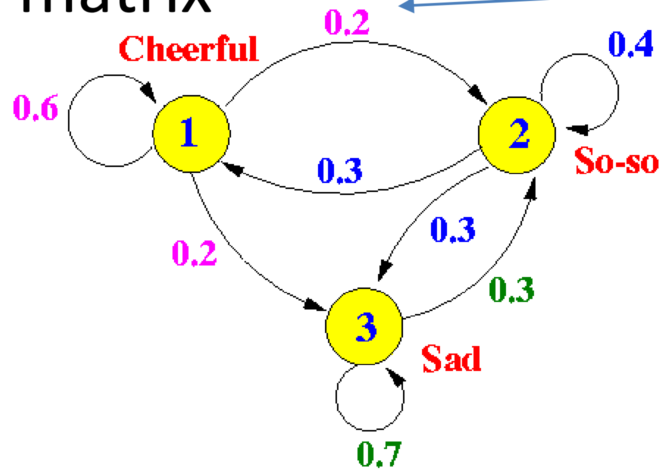
N: # of pages

Random walk

Theoretic model of PageRank

- Markov chains
  - A discrete-time stochastic process
    - It occurs in a series of time-steps in each of which a random choice is made
  - Can be described by a directed graph or a transition matrix

P(So-so|Cheerful)=0.2



$$M = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$

A first-order Markov chain for emotion

Markov chains

- Markov property
  - $P(X_{n+1}|X_1, \dots, X_n) = P(X_{n+1}|X_n)$
    - Memoryless (first-order)
- Transition matrix
  - A stochastic matrix
    - $\forall i, \sum_j M_{ij} = 1$
  - Key property
    - It has a principal left eigenvector corresponding to its largest eigenvalue, which is one

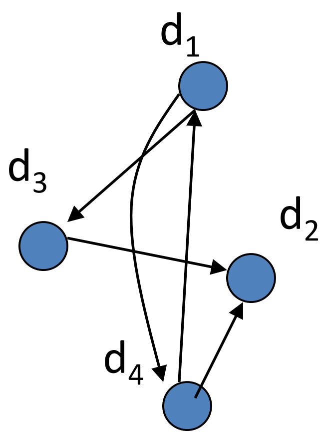
$$M = \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix}$$

Idea of random surfing

Mathematical interpretation of PageRank score

Theoretic model of PageRank

- Transition matrix of a Markov chain for PageRank



$$A = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

1. Enable random jump on dead end

$$A' = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

2. Normalization

$$A'' = \begin{pmatrix} 0 & 0 & 0.5 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0 & 1 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \end{pmatrix}$$

3. Enable random jump on all nodes

$\alpha = 0.5$

$$M = \begin{pmatrix} 0.125 & 0.125 & 0.375 & 0.375 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.125 & 0.625 & 0.125 & 0.125 \\ 0.375 & 0.375 & 0.125 & 0.125 \end{pmatrix}$$

Steps to derive transition matrix for PageRank

1. If a row of A has no 1's, replace each element by 1/N.
2. Divide each 1 in A by the number of 1's in its row.
3. Multiply the resulting matrix by $1 - \alpha$.
4. Add α/N to every entry of the resulting matrix, to obtain M.

A: adjacent matrix of network structure;
α: dumping factor

PageRank computation becomes

- $p_t(d) = Mp_{t-1}(d)$
  - Assuming $p_0(d) = \left[\frac{1}{N}, \dots, \frac{1}{N}\right]$
  - Iterative computation (forever?)
    - $p_t(d) = Mp_{t-1}(d) = \cdots = M^t p_0(d)$
  - Intuition: after enough rounds of random walk, each dimension of $p_t(d)$ indicates the frequency of a random surfer visiting document d
  - Question: will this frequency converges to certain fixed, steady-state quantity?

Stationary distribution of a Markov chain

- For a given Markov chain with transition matrix M, its stationary distribution of π is

$$\left.\begin{aligned} &\forall i \in S, \pi_i \geq 0 \\ &\sum_{i \in S} \pi_i = 1 \end{aligned}\right\} \text{A probability vector}$$

$$\pi = M\pi \longrightarrow \text{Random walk does not affect its distribution}$$

  - Necessary condition
    - Irreducible: a state is reachable from any other state
    - Aperiodic: states cannot be partitioned such that transitions happened periodically among the partitions

Markov chain for PageRank

- Random jump operation makes PageRank satisfy the necessary conditions
  1. Random jump makes every node is reachable for the other nodes
  2. Random jump breaks potential loop in a sub-network
- What does PageRank score really converge to?

Stationary distribution of PageRank

- For any irreducible and aperiodic Markov chain, there is a unique steady-state probability vector π, such that if $c(i,t)$ is the number of visits to state i after t steps, then

$$\lim_{t\to\infty} \frac{c(i,t)}{t} = \pi_i$$

  - PageRank score converges to the expected visit frequency of each node

Computation of PageRank

- Power iteration
  - $p_t(d) = Mp_{t-1}(d) = \cdots = M^t p_0(d)$
    - Normalize $p_t(d)$ in each iteration
    - Convergence rate is determined by the second eigenvalue
  - Random walk becomes series of matrix production
  - Alternative interpretation of PageRank score
    - Principal left eigenvector corresponding to its largest eigenvalue, which is one

$$M \times \pi = 1 \times \pi$$

Computation of PageRank

- An example from Manning's text book

$$M = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}$$

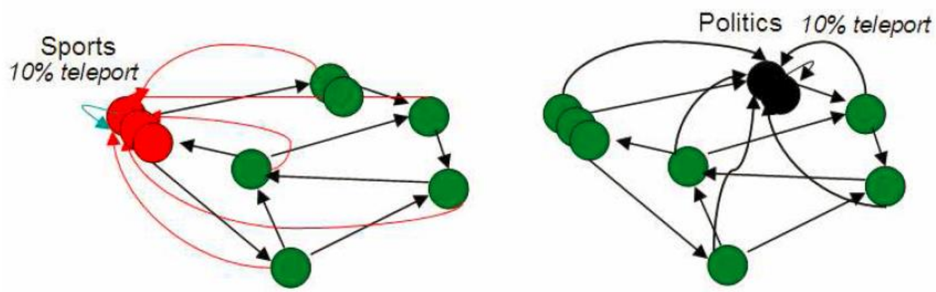
$\vec{x}_0$	1	0	0
$\vec{x}_1$	1/6	2/3	1/6
$\vec{x}_2$	1/3	1/3	1/3
$\vec{x}_3$	1/4	1/2	1/4
$\vec{x}_4$	7/24	5/12	7/24
...	...	...	...
$\vec{x}$	5/18	4/9	5/18

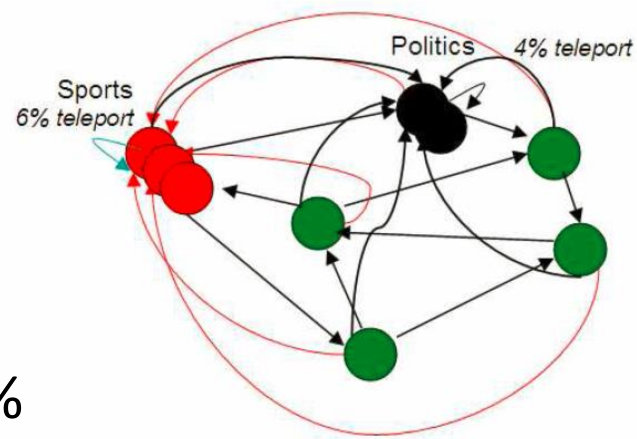
Variants of PageRank

- Topic-specific PageRank
  - Control the random jump to topic-specific nodes
    - E.g., surfer interests in Sports will only randomly jump to Sports-related website when they have no out-links to follow
  - $p_t(d) = [\alpha M + (1-\alpha)\vec{e}p^T(d)]p_{t-1}(d)$
    - $p^T(d) > 0$ iff d belongs to the topic of interest
    - $\vec{e}$ is a column vector of ones

Variants of PageRank

- Topic-specific PageRank
  - A user's interest is a mixture of topics





User's interest: 60%
Sports, 40% politics
Damping factor: 10%

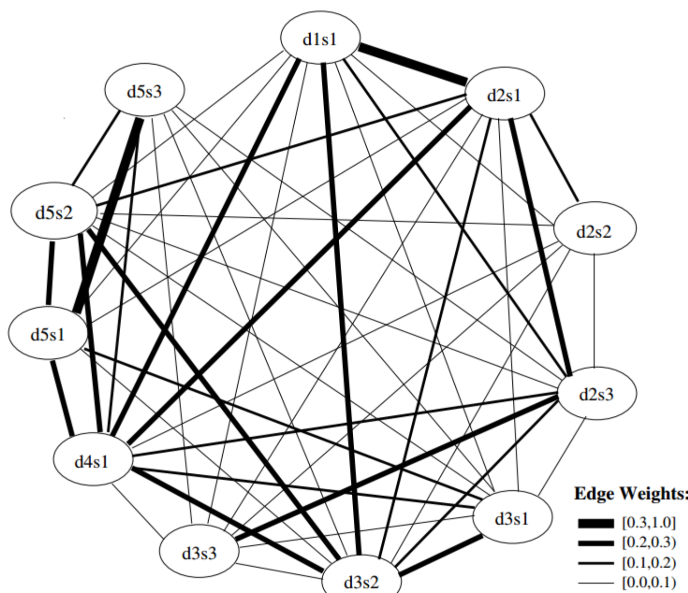
Compute it off-line

$$\pi = \sum_k p(T_k|user) \boxed{\pi_{T_k}}$$

Manning, "Introduction to Information Retrieval", Chapter 21, Figure 21.5

Variants of PageRank

- LexRank
  - *A sentence is important if it is similar to other important sentences*
  - PageRank on sentence similarity graph

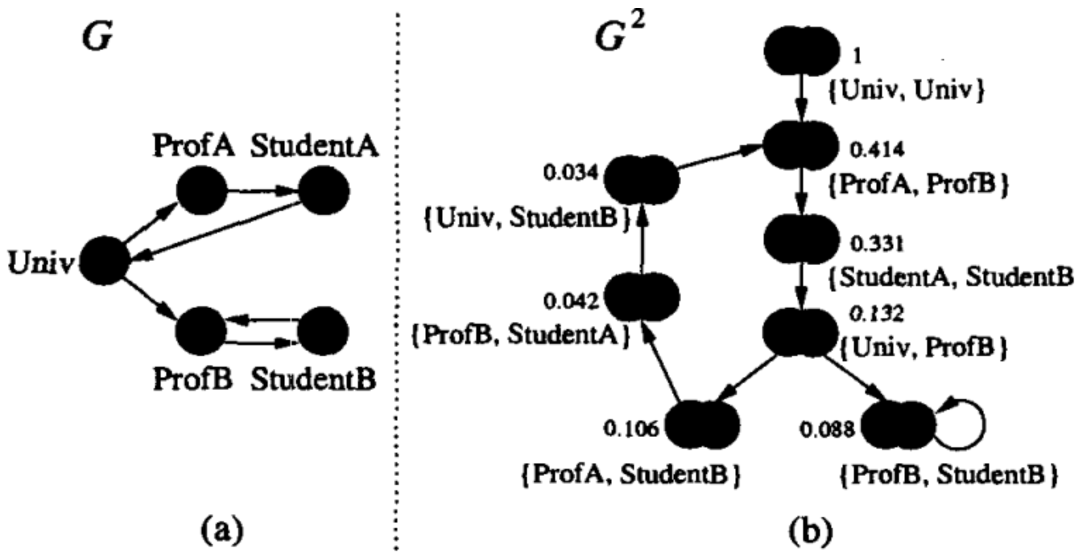


Centrality-based sentence salience ranking for document summarization

Erkan & Radev, JAIR'04

Variants of PageRank

- SimRank
  - *Two objects are similar if they are referenced by similar objects*
  - PageRank on bipartite graph of object relations



Measure similarity between objects via their connecting relation

Glen & Widom, KDD'02

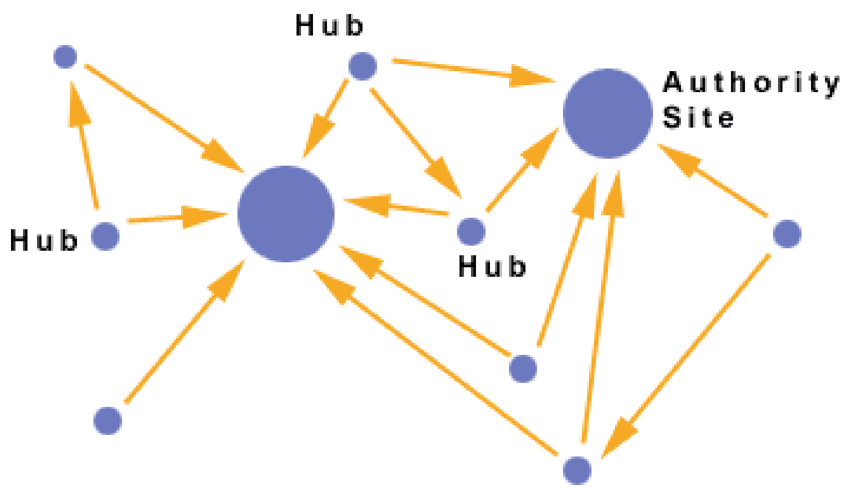
HITS algorithm

- Two types of web pages for a broad-topic query
  - Authorities – trustful source of information
    - UVa-> University of Virginia official site
  - Hubs – hand-crafted list of links to authority pages for a specific topic
    - Deep learning -> deep learning reading list
      - The monograph or review paper Learning Deep Architectures for AI (Foundations & Trends in Machine Learning, 2009).
      - The ICML 2009 Workshop on Learning Feature Hierarchies webpage has a list of references.
      - The LISA public wiki has a reading list and a bibliography.
      - Geoff Hinton has readings from last year's NIPS tutorial.

HITS algorithm

HITS=Hyperlink-Induced Topic Search

- Intuition
  - Using hub pages to discover authority pages
- Assumpt
  - A good ... many good authori
  - A good ... inted to by many g ... score
- Recursiv ... ve algorithm



Computation of HITS scores

- Two scores for a web page for a given query
  - Authority score: $a(d)$
  - Hub score: $h(d)$

$v \rightarrow d$ means there is a link from v to d

$$a(d) \leftarrow \sum_{v \rightarrow d} h(v)$$

$$h(d) \leftarrow \sum_{d \rightarrow v} a(v)$$

With proper normalization (L_2-norm)

Important HITS scores are query-dependent!

Computation of HITS scores

- In matrix form
  - $\vec{a} \leftarrow A\vec{h}$ and $\vec{h} \leftarrow A^T\vec{a}$
  - That is $\vec{a} \leftarrow AA^T\vec{a}$ and $\vec{h} \leftarrow A^TA\vec{h}$
  - Another eigen-system

$$\vec{a} = \frac{1}{\lambda_a}AA^T\vec{a}$$

$$\vec{h} = \frac{1}{\lambda_h}A^TA\vec{h}$$

Power iteration is applicable here as well

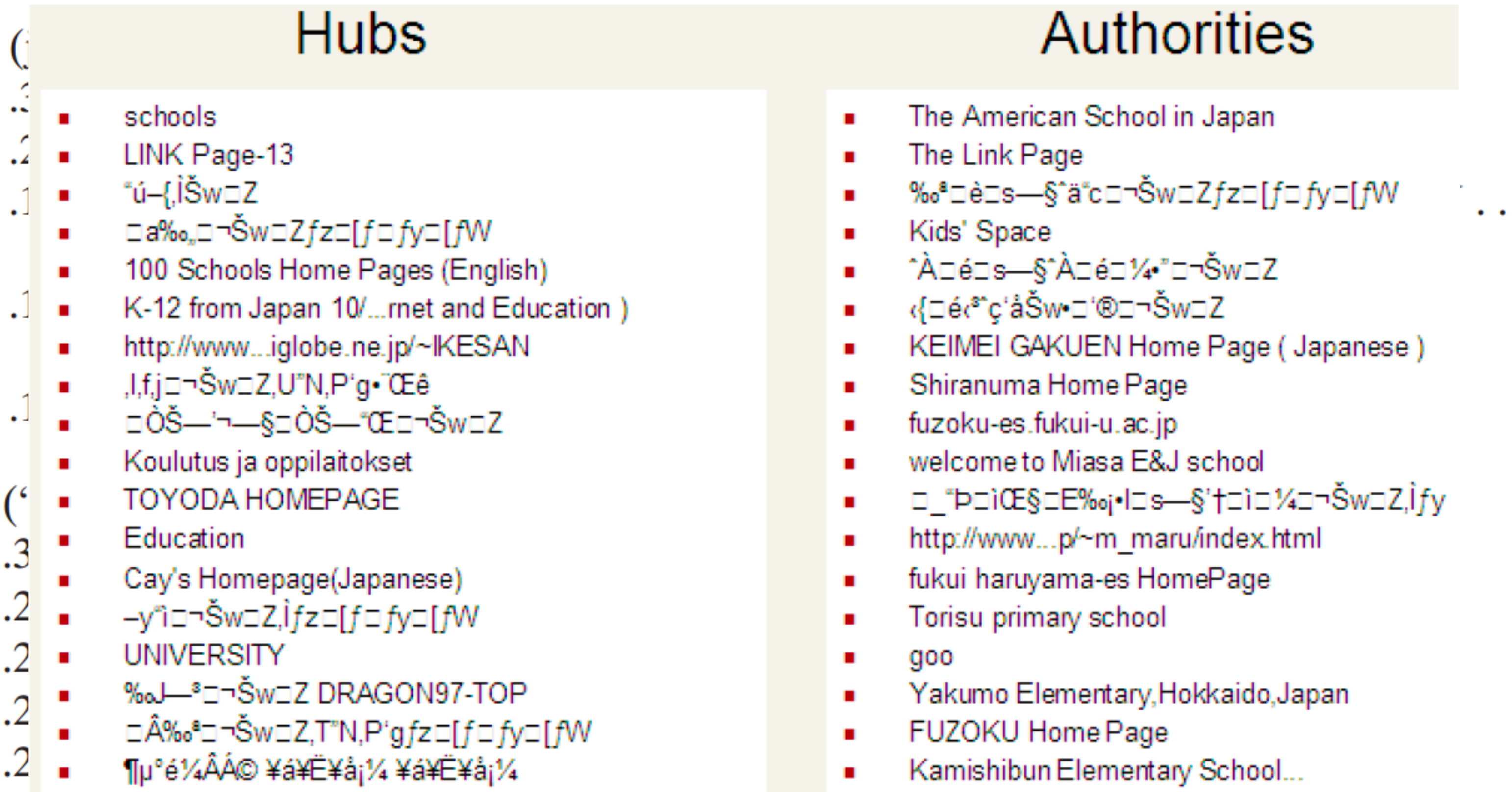
Constructing the adjacent matrix

- Only consider a subset of the Web
  1. For a given query, retrieve all the documents containing the query (or top *K* documents in a ranked list) – root set
  2. Expand the root set by adding pages either linking to a page in the root set, or being linked to by a page in the root set – base set
  3. Build adjacent matrix of pages in the base set

Constructing the adjacent matrix

- Reasons behind the construction steps
  - Reduce the computation cost
  - A good authority page may not contain the query text
  - The expansion of root set might introduce good hubs and authorities into the sub-network

Sample results



Kleinberg, JACM'99

Manning, “Introduction to Information Retrieval”, Chapter 21, Figure 21.6

References

- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: Bringing order to the web." (1999)
- Haveliwala, Taher H. "Topic-sensitive pagerank." In Proceedings of the 11th international conference on World Wide Web, pp. 517-526. ACM, 2002.
- Erkan, Günes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." J. Artif. Intell. Res.(JAIR) 22, no. 1 (2004): 457-479.
- Jeh, Glen, and Jennifer Widom. "SimRank: a measure of structural-context similarity." In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 538-543. ACM, 2002.
- Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) 46, no. 5 (1999): 604-632.