CS4051

Information Retrieval

Week 14

Muhammad Rafi

May 08, 2024

Link Analysis

Chapter No. 21

Today's Agenda

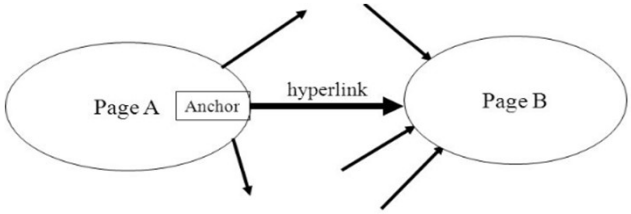
- Link Analysis
- Web as a Graph
- Page Rank & Markov Chain
- Page Rank Computation
- HITS Algorithm
- HITS vs. PageRank
- Conclusion

Link Analysis

- The analysis of hyperlinks and the graph structure of the Web has been instrumental in the development of web search.
- In this chapter we focus on the use of hyperlinks for ranking web search results.
- Such link analysis is one of many factors considered by web search engines in computing a composite score for a web page on any given query.
- Web Page – <In-Link> and <Out-Link>
  - Journal of the ACM.
  - Anchor Text
  - Extended Anchor Text

The Web as a graph

- Web pages are connected with in- and out- links
- Our study of link analysis builds on two intuitions:
  - The anchor text pointing to page B is a good description of page B.
  - The hyperlink from A to B represents an endorsement of page B, by the creator of page A.



The Web as a graph

- Assumption 1 : A hyperlink between pages denotes a conferral of authority (quality signal)
- Assumption 2: The text in the anchor of the hyperlink describe the target page(Context/ Textual description of a page).
- The Web is full of instances where the page B does not provide an accurate description of itself.
  - For example, at the time of the writing of this book the home page of the IBM corporation (http://www.ibm.com)did not contain the term computer anywhere in its HTML code, despite the fact that IBM is widely viewed as the world's largest computer maker.
  - Thus, there is often a gap between the terms in a web page, and how web users would describe that web page.

Page Rank

- Our first technique for link analysis assigns to every node in the web graph a numerical score between 0 and 1, known as its PageRank.
- The PageRank of a node will depend on the link structure of the web graph.
- Given a query, a web search engine computes a composite score for each web page that combines hundreds of features such as cosine similarity and term proximity, together with the PageRank score.

Page Rank

- Consider a random surfer who begins at a web page (a node of the web graph) and executes a random walk on the Web as follows.
  - At each time step, the surfer proceeds from his current page A to a randomly chosen web page that A hyperlinks to.
  - As the surfer proceeds in this random walk from node to node, he visits some nodes more often than others; intuitively, these are nodes with many links coming in from other frequently visited nodes.
  - In the teleport operation the surfer jumps from a node to any other node in the web graph. This could happen because he types an address into the URL bar of his browser.
  - Teleporting is uniformly performed.

Teleporting

- In the teleport operation the surfer jumps from a node to any other node in the web graph.
- In assigning a PageRank score to each node of the web graph, we use the teleport operation in two ways:
  - When at a node with no out-links, the surfer invokes the teleport operation.
  - At any node that has outgoing links, the surfer invokes the teleport operation with probability $0 < \alpha < 1$ and the standard random walk (follow an out-link chosen uniformly at random with probability $1 - \alpha$, where α is a fixed parameter chosen in advance. Typically, α might be 0.1.

Markov Chain

- A Markov chain is a discrete-time stochastic process: a process that occurs in a series of time-steps in each of which a random choice is made.
- A Markov chain consists of N states. Each web page will correspond to a state in the Markov chain we will formulate.
- A Markov chain is characterized by an $N \times N$ transition probability matrix P each of whose entries is in the interval [0, 1]; the entries in each row of P add up to 1.

Markov Chain

- The Markov chain can be in one of the N states at any given timestep; then, the entry Pij tells us the probability that the state at the next timestep is j, conditioned on the current state being i.
- Each entry Pij is known as a transition probability and depends only on the current state i; this is known as the Markov property.

Markov Chain as Stochastic Matrix

- A matrix with non-negative entries that satisfies is known as a stochastic matrix.
- A key property of a stochastic matrix is that it has a principal left eigenvector corresponding to its largest eigenvalue, which is 1.
- An N-dimensional probability vector each of whose components corresponds to one of the N states of a Markov chain can be viewed as a probability distribution over its states
- Markov Properties

$$\forall i, j, P_{ij} \in [0, 1] \qquad \forall i, \sum_{j=1}^{N} P_{ij} = 1.$$

Markov Chain Probability Matrix

- How to get the probability matrix?
  - If a row of A has no 1's, then replace each element by 1/N.
  - For all other rows proceed as follows:
  - Divide each 1 in A by the number of 1's in its row. Thus, if there is a row with three m's, then each of them is replaced by 1/m.
  - Multiply the resulting matrix by $1 - \alpha$.
  - Add α/N to every entry of the resulting matrix, to obtain the required matrix P.

Ergodic Markov chain

- A Markov chain is said to be ergodic if there exists a positive integer T0 such that for all pairs of states i, j in the Markov chain, if it is started at time 0 in state i then for all $t > T0$, the probability of being in state j at time t is greater than 0.
- The random walk with teleporting results in a unique distribution of steady-state probabilities over the states of the induced Markov chain.
- This steady-state probability for a state is the PageRank of the corresponding web page.

The PageRank computation

- The N entries in the principal eigenvector ~π are the steady-state probabilities of the random walk with teleporting, and thus the PageRank values for the corresponding web pages.

We consider the web graph in Exercise 21.6 with $\alpha = 0.5$. The transition probability matrix of the surfer's walk with teleportation is then

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

Imagine that the surfer starts in state 1, corresponding to the initial probability distribution vector $\vec{x}_0 = (1\ 0\ 0)$. Then, after one step the distribution is

$$\vec{x}_0 P = \begin{pmatrix} 1/6 & 2/3 & 1/6 \end{pmatrix} = \vec{x}_1.$$

The PageRank computation

$\vec{x}_0$	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

After two steps it is

$$\vec{x}_1 P = \begin{pmatrix} 1/6 & 2/3 & 1/6 \end{pmatrix} \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix} = \vec{x}_2.$$

A small Web Example

- Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 → 2, 3 → 2, 2 → 1, 2 → 3. Write down the transition probability matrices for the surfer's walk with teleporting, for the following three values of the teleport probability:
  - (a) $\alpha = 0$;
  - (b) $\alpha = 0.5$ and
  - (c) $\alpha = 1$.

Advantage/Disadvantage PageRank

Advantages of PageRank

1. The algorithm is robust against Spam since its not easy for a webpage owner to add inlinks to his/her page from other important pages.
2. PageRank is a global measure and is query independent.

Disdvantages of PageRank

1. The major disadvantage of PageRank is that it favors the older pages, because a new page, even a very good one will not have many links unless it is a part of an existing site.
2. PageRank can be easily increased by the concept of "link-farms" as shown below. However, while indexing, the search actively tries to find these flaws.

HITS

- Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg.
- Hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages.

HITS

- In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs
- The algorithm assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

Example

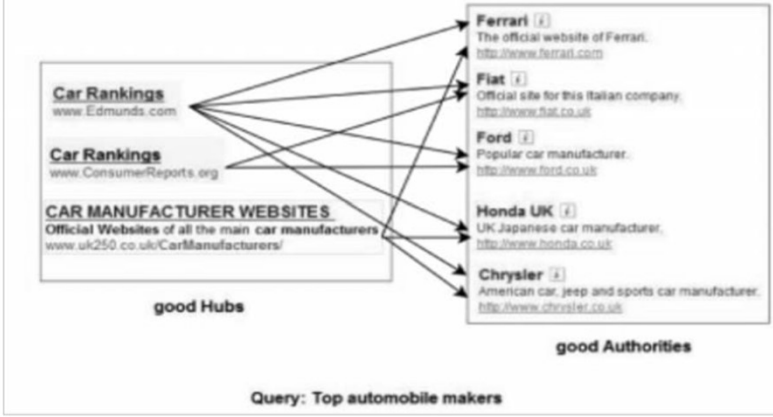


Figure 1: An example of authorities and hubs

HITS Algorithm

The HITS Algorithm can be described as follows:

1) Given a search query Q, collect the top 200 webpages that contain the highest frequency of query Q.

2) Add the the collection the webpages that point to or are pointed by these top 200 webpages. Create Adjacency Matrix A among these webpages.

3) Initialize the hub and authority column vectors U and V with values of 1.

4) For a set k number of iterations, do the following:

```
a) Update the authority scores through the authority matrix V
b) Update the hub scores through the hub matrix U
c) Normalize the hub matrix and authority matrix U and V
```

5) Rank the webpages according to the authority score as reflected through authority matrix V

HITS Algorithm

The algorithm performs a series of iterations, each consisting of two basic steps:

- **Authority Update**: Update each node's *Authority score* to be equal to the sum of the *Hub Scores* of each node that points to it. That is, a node is given a high authority score by being linked to pages that are recognized as Hubs for information.
- **Hub Update**: Update each node's *Hub Score* to be equal to the sum of the *Authority Scores* of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

HITS Algorithm

$\forall p$, we update $\text{auth}(p)$ to be the summation:

$$\text{auth}(p) = \sum_{i=1}^{n} \text{hub}(i)$$

where n is the total number of pages connected to p and i is a page connected to p. That is, the Authority score of a page is the sum of all the Hub scores of pages that point to it.

$\forall p$, we update $\text{hub}(p)$ to be the summation:

$$\text{hub}(p) = \sum_{i=1}^{n} \text{auth}(i)$$

where n is the total number of pages p connects to and i is a page which p connects to. Thus a page's Hub score is the sum of the Authority scores of all its linking pages

HITS Algorithm

$$a_j^{(k)} \leftarrow \sum_{(i,j)\in E} h_i^{(k-1)}, \qquad h_i^{(k)} \leftarrow \sum_{(i,j)\in E} a_j^{(k)},$$

$$\mathbf{a}^{(k)} \leftarrow \mathbf{A}^{\mathbf{T}} \cdot \mathbf{h}^{(k-1)}. \qquad \mathbf{h}^{(k)} \leftarrow \mathbf{A} \cdot \mathbf{a}^{(k)}.$$

An Example (HITS)

Hubs

- schools
- LINK Page-13
- "ú–{,ÌŠw□Z
- □a‰,□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education)
- http://www.…iglobe.ne.jp/~IKESAN
- ,l,f,j□¬Šw□Z,U"N,P'g•"Œê
- □ÒŠ—'¬—§□ÒŠ—"Œ□¬Šw□Z
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- –y"ì□¬Šw□Z,Ìƒz□[ƒ□ƒy□[ƒW
- UNIVERSITY
- ‰J—ª□¬Šw□Z DRAGON97-TOP
- □Â‰ª□¬Šw□Z,T"N,P'gƒz□[ƒ□ƒy□[ƒW
- ¶µ°é¼ÂÁ© ¥á¥Ë¥å¡¼ ¥á¥Ë¥å¡¼

Authorities

- The American School in Japan
- The Link Page
- ‰ª□è□s—§"ä"c□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
- Kids' Space
- "À□é□s—§"À□é□¼•"□¬Šw□Z
- ‹{□é‹³"ç'åŠw•□'®□¬Šw□Z
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- □_"Þ□ìŒ§□E‰l•l□s—§'†□ì□¼□¬Šw□Z,Ìƒy
- http://www...p/~m_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

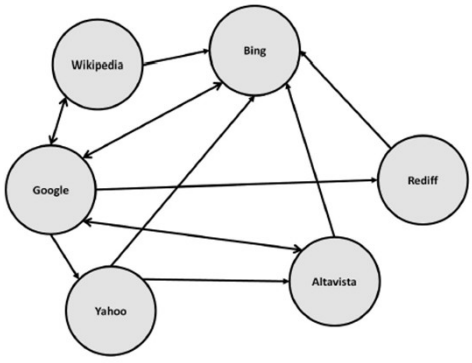
► **Figure 21.6** A sample run of HITS on the query japan elementary schools.

HITS-Issues

- **It is query dependent, that is, the (Hubs and Authority) scores resulting from the link analysis are influenced by the search terms;**
- **As a corollary, it is executed at query time, not at indexing time, with the associated hit on performance that accompanies query-time processing.**
- **It is not commonly used by search engines. (Though a similar algorithm was said to be used by Teoma, which was acquired by Ask Jeeves/Ask.com.)**
- **It computes two scores per document, hub and authority, as opposed to a single score;**
- **It is processed on a small subset of 'relevant' documents (a 'focused subgraph' or base set), not all documents as was the case with PageRank.**

Example:

- A subset of graph with selected Hub & Authority status.



- This is a result of resultant search result on "q"

Adjacency Matrix

	Wiki	Google	Bing	Yahoo	Altavista	Rediff
Wikipedia	0	1	1	0	0	0
Google	1	0	1	1	1	1
Bing	0	1	0	0	0	0
Yahoo	0	0	1	0	1	0
Altavista	0	1	1	0	0	0
Rediffmail	0	0	1	0	0	0

Iterative calculation of Hub & Authority

$$\mathbf{a}^{(1)} = \mathbf{A}^{\mathbf{T}} \cdot \mathbf{h}^{(0)}$$

$$= \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}^{\mathbf{T}} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Iterative calculation of Hub & Authority

$$= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 3 \\ 5 \\ 1 \\ 2 \\ 1 \end{bmatrix}$$

Normalized

$$\mathbf{a}^{(1)} = \begin{bmatrix} \frac{1}{\sqrt{1^2+3^2+5^2+1^2+2^2+1^2}} \\ \frac{3}{\sqrt{1^2+3^2+5^2+1^2+2^2+1^2}} \\ \frac{5}{\sqrt{1^2+3^2+5^2+1^2+2^2+1^2}} \\ \frac{1}{\sqrt{1^2+3^2+5^2+1^2+2^2+1^2}} \\ \frac{2}{\sqrt{1^2+3^2+5^2+1^2+2^2+1^2}} \\ \frac{1}{\sqrt{1^2+3^2+5^2+1^2+2^2+1^2}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\sqrt{41}} \\ \frac{3}{\sqrt{41}} \\ \frac{5}{\sqrt{41}} \\ \frac{1}{\sqrt{41}} \\ \frac{2}{\sqrt{41}} \\ \frac{1}{\sqrt{41}} \end{bmatrix}$$

$$= \begin{bmatrix} 0.15617 \\ 0.46852 \\ 0.78087 \\ 0.15617 \\ 0.312348 \\ 0.15617 \end{bmatrix}$$

HITS vs. PageRank

HITS	PageRank
It gives 2 scores Hub and Authority for each page.	It gives one score e.g. PageRank.
It is executed at query time	It is executed at indexing time.
Not robust against spams.	Robust against web-spams.
Never favor pages, but can be manipulated.	Favor old pages.
It is query dependent	It is query independent