CS317

Information Retrieval

Week 14

Muhammad Rafi

May 02, 2019

Link Analysis

Chapter No. 21

Web as Graph

- Link analysis of Web (as graph) is based on the following two assumptions:
  - 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.
  - This is not always the case; for instance, many links amongst pages within a single website stem from the user of a common template. (company pages to copy right page links)

Web as Graph

- The Web is full of instances where the page B does not provide an accurate description of itself.
- Thus, there is often a gap between the terms in a web page, and how web users would describe that web page.
- Web pages is a composition of text, graphics and images. Standard IR approach does not support searching with these rich contents.
- The window of text surrounding anchor text (sometimes referred to as extended anchor text) is often usable in the same manner as anchor text itself;

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

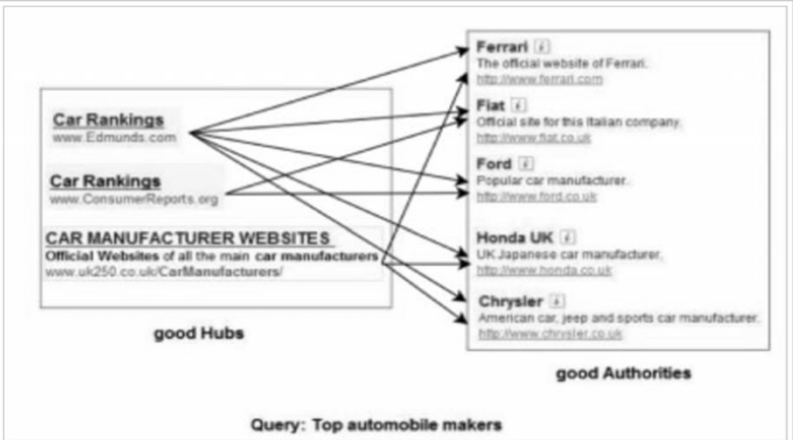


Figure 1: An example of authorities and hubs

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

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

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

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

$$\mathbf{h}^{(k)} \leftarrow \mathbf{A} \cdot \mathbf{a}^{(k)}.$$

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—³□¬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‹³ˆç'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...

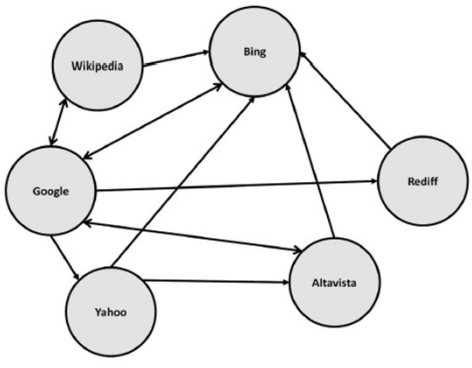
► **Figure 21.6** A sample run of HITS on the query japan elementary schools.

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

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

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

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

Example

Consider a segment of web graph for link analysis based on HITS algorithm, containing four pages n1, n2, n3 and n4; n1 is connected to n2, n3 and n4; and n2 is connected n3 and n4; n3 is connected to n1 and n4, n4 is connected to n4; using A and H as column metrics. Produce two iterations of HITS algorithm and updates on A and H. Identify one page as the best hub and authority. Use L2 normalization for A and H both. [15]

$$\vec{h} \leftarrow A\vec{a}$$
$$\vec{a} \leftarrow A^T\vec{h},$$

Let A be the connectivity matrix for the given graph. We know that

$h^1 = A \cdot a^0$ and $a^1 = A^T \cdot h^0$

A =

0	1	1	1
0	0	1	1
1	0	0	1
0	0	0	1

A^T =

0	0	1	0
1	0	0	0
1	1	0	0
1	1	1	1

h =

1
1
1
1

t=

1	1	1	1

a^1	<1 1 2 4> normalizing we get < 0.213 0.213 0.426 0.852>
h^1	<3 2 2 1> normalizing we get < 0.707 0.471 0.471 0.235 >
a^2	< 0.199 0.298 0.495 0.792>
h^2	< 0.623 0.543 0.445 0.356 >

Best Hub is n1; Best Authority is n4