

CS47300: Web Information Search and Management

Using Graph Structure for Retrieval

Prof. Chris Clifton

23 September 2020

Material adapted from slides created by Dr. Rong Jin (formerly Michigan State, now at Alibaba)



Ad-Hoc Retrieval: Beyond the Words

- Web is a graph
 - Each web site correspond to a node
 - A link from one site to another site forms a directed edge

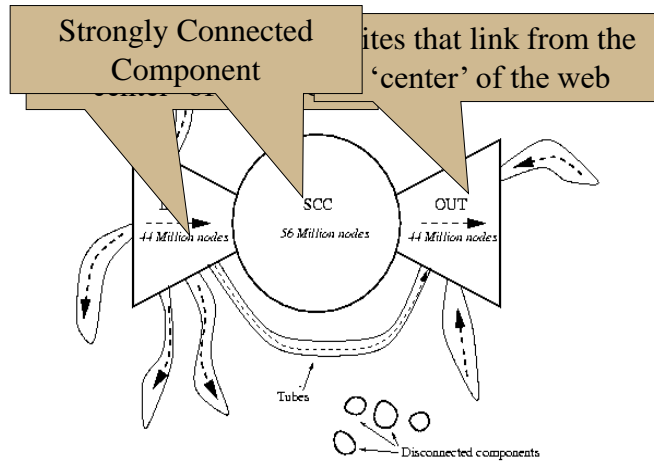
Citation Analysis

The screenshot shows a Microsoft Internet Explorer browser window displaying a page from CiteSeer. The page title is "Probabilistic Latent Semantic Analysis (1999)" by Thomas Hofmann. The page includes an abstract, a list of related documents, and a list of documents that cite this work. The abstract states: "Probabilistic Latent Semantic Analysis is a novel statistical technique for the analysis of two-mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas. Compared to standard Latent Semantic Analysis which stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived from a latent class... (Update)". The page also lists several related documents and documents that cite this work, including "A Hierarchical Model for Clustering and - Categorizing Documents" by Gaussier (2002) and "Indexing by latent semantic analysis" by Deerwester, Dumais et al. (1990).

Ad-Hoc Retrieval: Beyond the Words

- Web is a graph
 - Each web site correspond to a node
 - A link from one site to another site forms a directed edge
- What does it look like?
 - Web is small world
 - The diameter of the web is 19
 - e.g. the average number of clicks from one web site to another is 19

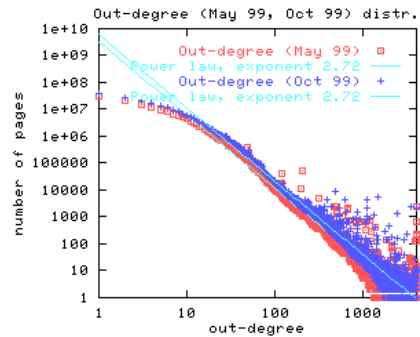
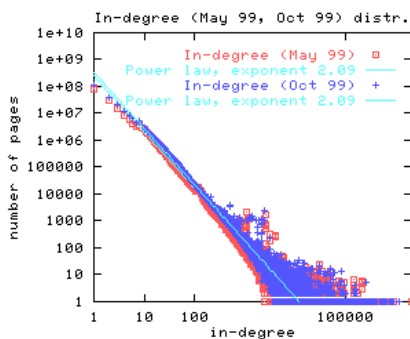
Bowtie Structure



Broder et al., 2001

Inlinks and Outlinks

- Both degrees of incoming and outgoing links follow power law



Broder et al., 2001

So what good is Link Structure?

- When you can't find something, do you:
 - Keep looking in the same place?
 - Look somewhere else?
 - Give up?
 - Ask for help?
- *Other people may already know the answer!*
 - Links: Reflect human judgement

10

Early Approaches

- Basic Assumptions
 - Hyperlinks contain information about the human judgment of a site
 - The more incoming links to a site, the more it is judged important
- Bray 1996
 - The visibility of a site is measured by the number of other sites pointing to it
 - The luminosity of a site is measured by the number of other sites to which it points
 - Limitation: failure to capture the relative importance of different parents (children) sites

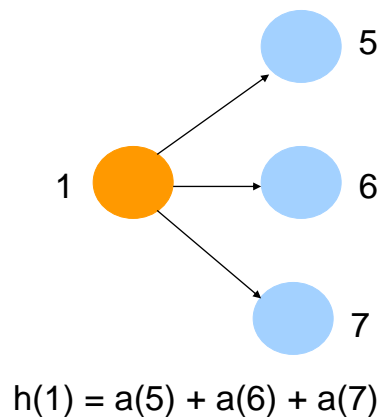
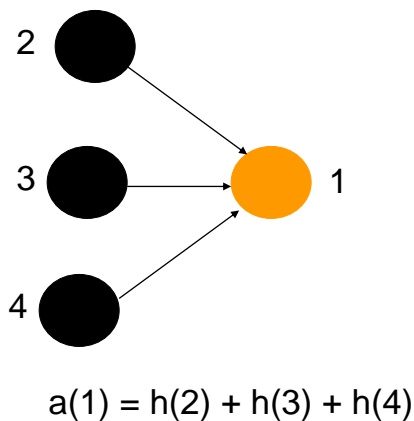
12

HITS - Kleinberg's Algorithm

- HITS – Hypertext Induced Topic Selection
- For each vertex $v \in V$ in a subgraph of interest:
 - $a(v)$ - the authority of v
 - $h(v)$ - the hubness of v
- A site is very authoritative if it receives many citations.
 - Citation from important sites weight more than citations from less-important sites
- Hubness shows the importance of a site.
 - A good hub is a site that links to many authoritative sites

14

Authority and Hubness



15

Authority and Hubness: Version 1

Recursive dependency

$$a(v) = \sum_{w \in pa[v]} h(w)$$

$$h(v) = \sum_{w \in ch[v]} a(w)$$

HubsAuthorities(G)

```

1  1 ← [1,...,1] ∈ R|V|
2  a0 ← h0 ← 1
3  t ← 1
4  repeat
5      for each v in V
6          do at(v) ← ∑w ∈ pa[v] ht-1(w)
7              ht(v) ← ∑w ∈ ch[v] at-1(w)
8  t ← t + 1
9  until || at - at-1 || + || ht - ht-1 || < ε
10 return (at, ht)

```

Problems ?

16

Authority and Hubness: Version 2

Recursive dependency

$$a(v) = \sum_{w \in pa[v]} h(w)$$

$$h(v) = \sum_{w \in ch[v]} a(w)$$

+ Normalization

$$a(v) = \frac{a(v)}{\sum_w a(w)}$$

$$h(v) = \frac{h(v)}{\sum_w h(w)}$$

HubsAuthorities(G)

```

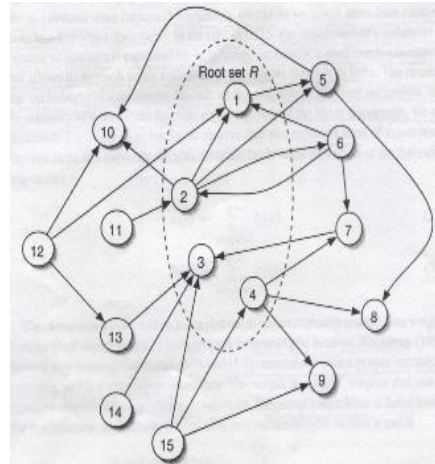
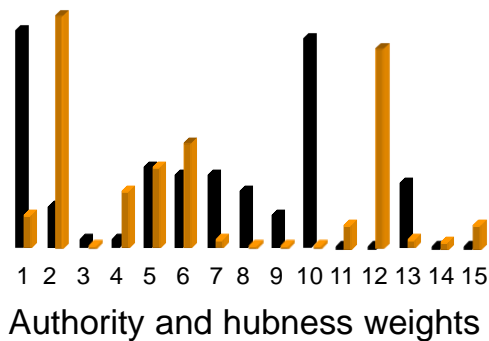
1  1 ← [1,...,1] ∈ R|V|
2  a0 ← h0 ← 1
3  t ← 1
4  repeat
5      for each v in V
6          do at(v) ← ∑w ∈ pa[v] ht-1(w)
7              ht(v) ← ∑w ∈ ch[v] at-1(w)
8          at ← at / || a ||
9          ht ← ht / || h ||
10 t ← t + 1
11 until || at - at-1 || + || ht - ht-1 || < ε
12 return (at, ht)

```

17

HITS Example Results

■ Authority
■ Hubness



18

Authority and Hubness

- Authority score
 - Not only depends on the number of incoming links
 - But also the 'quality' (e.g., hubness) of the incoming links
- Hubness score
 - Not only depends on the number of outgoing links
 - But also the 'quality' (e.g., hubness) of the outgoing links

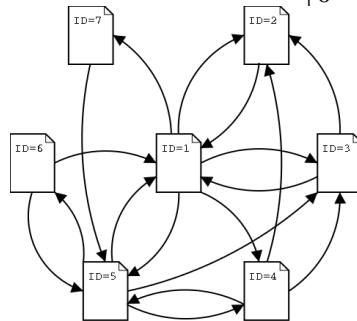
19

Authority and Hub

- Column vector \mathbf{a} : a_i is the authority score for the i -th site
- Column vector \mathbf{h} : h_i is the hub score for the i -th site

- Matrix \mathbf{M} :

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{the } i\text{th site points to the } j\text{th site} \\ 0 & \text{otherwise} \end{cases}$$



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

20

Authority and Hub

- Vector \mathbf{a} : a_i is the authority score for the i -th site
- Vector \mathbf{h} : h_i is the hub score for the i -th site

- Matrix \mathbf{M} :

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{the } i\text{th site points to the } j\text{th site} \\ 0 & \text{otherwise} \end{cases}$$

- Recursive dependency:

$$a(v) \leftarrow \sum_{w \in \text{pa}[v]} h(w)$$

$$h(v) \leftarrow \sum_{w \in \text{ch}[v]} a(w)$$

21

Authority and Hub

- Column vector \mathbf{a} : a_i is the authority score for the i -th site
- Column vector \mathbf{h} : h_i is the hub score for the i -th site

- Matrix \mathbf{M} :

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{the } i\text{th site points to the } j\text{th site} \\ 0 & \text{otherwise} \end{cases}$$

- Recursive dependency:

$$\begin{aligned} a(v) &\leftarrow \sum_{w \in \text{pa}[v]} h(w) \\ h(v) &\leftarrow \sum_{w \in \text{ch}[v]} a(w) \end{aligned} \quad \Rightarrow \quad \begin{aligned} \mathbf{a} &= \mathbf{M}^T \mathbf{h} \\ \mathbf{h} &= \mathbf{M} \mathbf{a} \end{aligned}$$

22

Authority and Hub

- Column vector \mathbf{a} : a_i is the authority score for the i -th site
- Column vector \mathbf{h} : h_i is the hub score for the i -th site
- Matrix \mathbf{M} :

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{the } i\text{th site points to the } j\text{th site} \\ 0 & \text{otherwise} \end{cases}$$

- Recursive dependency:

$$\begin{aligned} a(v) &\leftarrow \sum_{w \in \text{pa}[v]} h(w) \\ h(v) &\leftarrow \sum_{w \in \text{ch}[v]} a(w) \end{aligned} \quad \Rightarrow \quad \begin{aligned} \mathbf{a}_t &= \alpha_t \mathbf{M}^T \mathbf{h}_t \\ \mathbf{h}_t &= \beta_t \mathbf{M} \mathbf{a}_t \end{aligned}$$

Normalization Procedure

23

Authority and Hub

$$\left. \begin{array}{l} \mathbf{a}_t = \alpha_t \mathbf{M}^T \mathbf{h}_t \\ \mathbf{h}_t = \beta_t \mathbf{M} \mathbf{a}_t \end{array} \right\} \rightarrow \begin{array}{l} \mathbf{a}_t = \alpha_t \beta_t \mathbf{M}^T \mathbf{M} \mathbf{a}_t \\ \mathbf{h}_t = \alpha_t \beta_t \mathbf{M} \mathbf{M}^T \mathbf{h}_t \end{array}$$

- Apply SVD to matrix \mathbf{M}

$$\mathbf{M} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T = \sum_i \lambda_i \mathbf{u}_i \mathbf{v}_i^T \longrightarrow \mathbf{a} = \mathbf{u}_1, \mathbf{h} = \mathbf{v}_1$$