PURDUE UNIVERSITY | Department of Computer Science

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)

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

Citation Analysis

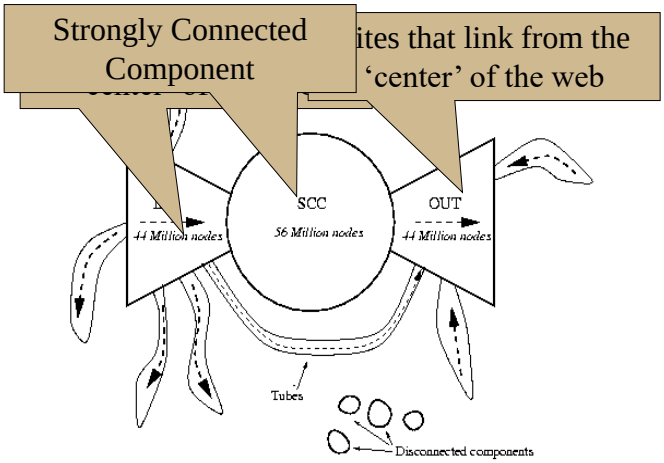
PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

Bowtie Structure



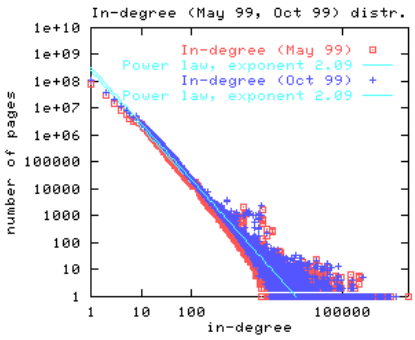
Broder et al., 2001

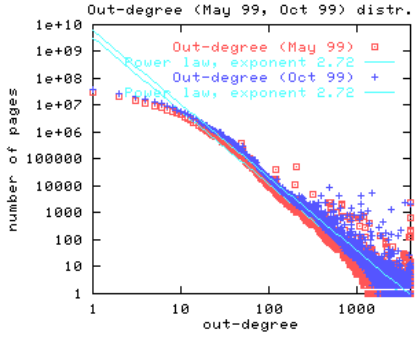
5

PURDUE UNIVERSITY
Department of Computer Science

Inlinks and Outlinks

- Both degrees of incoming and outgoing links follow power law





Broder et al., 2001

9

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

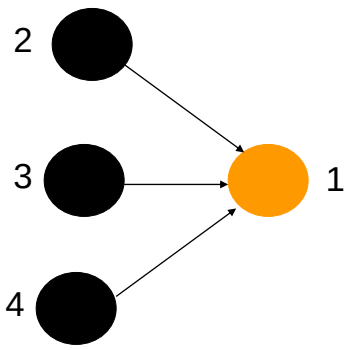
HITS - Kleinberg's Algorithm

- HITS – Hypertext Induced Topic Selection
- For each vertex v ϵ 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

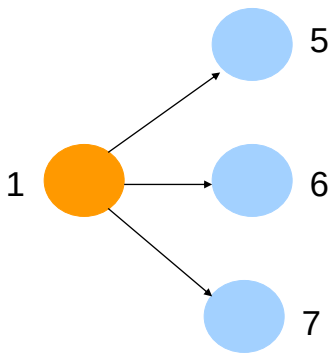
14

PURDUE UNIVERSITY
Department of Computer Science

Authority and Hubness



a(1) = h(2) + h(3) + h(4)



h(1) = a(5) + a(6) + a(7)

15

PURDUE UNIVERSITY
Department of Computer Science

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] Є $R^{|V|}$
2 a_0 ← h_0 ← **1**
3 t ← 1
4 repeat
5 for each v in V
6 do $a_t(v)$ ← $\Sigma_{w \in pa[v]} h_{t-1}(w)$
7 $h_t(v)$ ← $\Sigma_{w \in pa[v]} a_{t-1}(w)$
8 t ← t + 1
9 until $\| a_t - a_{t-1} \| + \| h_t - h_{t-1} \| < \varepsilon$
10 return (a_t, h_t)

Problems ?

PURDUE UNIVERSITY
Department of Computer Science

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

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

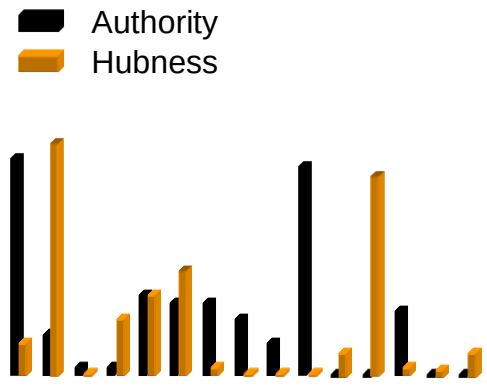
$$h(v) = \frac{h(v)}{\sum_w h(w)}$$

HubsAuthorities(G)

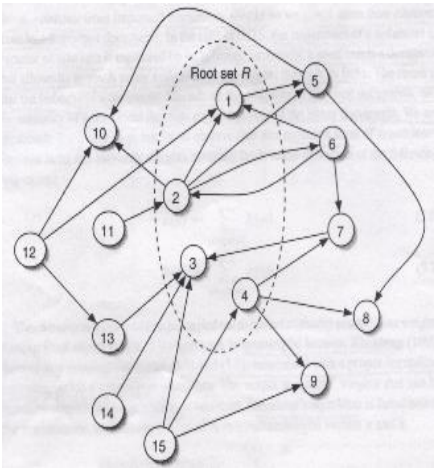
1 **1** ← [1,…,1] Є $R^{|V|}$
2 a_0 ← h_0 ← **1**
3 t ← 1
4 repeat
5 for each v in V
6 do $a_t(v)$ ← $\Sigma_{w \in pa[v]} h_{t-1}(w)$
7 $h_t(v)$ ← $\Sigma_{w \in pa[v]} a_{t-1}(w)$
8 a_t ← a_t / || a ||
9 h_t ← h_t / || h ||
10 t ← t + 1
11 until $\| a_t - a_{t-1} \| + \| h_t - h_{t-1} \| < \varepsilon$
12 return (a_t, h_t)

PURDUE UNIVERSITY
Department of Computer Science

HITS Example Results



Authority and hubness weights



PURDUE UNIVERSITY
Department of Computer Science

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

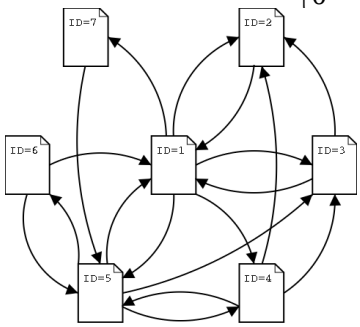
PURDUE UNIVERSITY
Department of Computer Science

Authority and Hub

- Column vector ***a***: a_i is the authority score for the i-th site
- Column vector ***h***: h_i is the hub score for the i-th site

- Matrix **M**:

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



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

PURDUE UNIVERSITY
Department of Computer Science

Authority and Hub

- Vector ***a***: a_i is the authority score for the i-th site
- Vector ***h***: h_i is the hub score for the i-th site

- Matrix **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:

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

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

PURDUE UNIVERSITY
Department of Computer Science

Authority and Hub

•Column vector ***a***: a_i is the authority score for the i-th site
•Column vector ***h***: h_i is the hub score for the i-th site

•Matrix **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:

$$a(v) \leftarrow \Sigma_{w \in pa[v]}\ h(w)$$
$$h(v) \leftarrow \Sigma_{w \in ch[v]}\ a(w)$$

$$\Rightarrow \quad \mathbf{a} = \mathbf{M}^T\mathbf{h}, \quad \mathbf{h} = \mathbf{M}\mathbf{a}$$

22

PURDUE UNIVERSITY
Department of Computer Science

Authority and Hub

- Column vector ***a***: a_i is the authority score for the i-th site
- Column vector ***h***: h_i is the hub score for the i-th site

- Matrix **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:

$$a(v) \leftarrow \Sigma_{w \in pa[v]}\ h(w)$$
$$h(v) \leftarrow \Sigma_{w \in ch[v]}\ a(w)$$

$$\Rightarrow \quad \mathbf{a}_t = \alpha_t \mathbf{M}^T\mathbf{h}_t, \quad \mathbf{h}_t = \beta_t \mathbf{M}\mathbf{a}_t$$

Normalization Procedure

23

PURDUE UNIVERSITY
Department of Computer Science

Authority and Hub

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

- Apply SVD to matrix **M**

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

24