SS-ZG548: ADVANCED DATA MINING

13 Topics in Web Mining

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Oct 26, 2021 ONLINE (WILP @ BITS-Pilani July-Dec 2021)

http://ktiwari.in/adm

Statistics

There were 100 images in a box. 30 of them were containing lion. I asked Bob to separate all the pics of lion. He showed me 60 but, lion was not in 40 of them.

- True positives (TP): 20
- True negatives (TN): 30
- T1-Error: False positives (FP): 40
- T2-Error: False negatives (FN): 10

Confusion Matrix

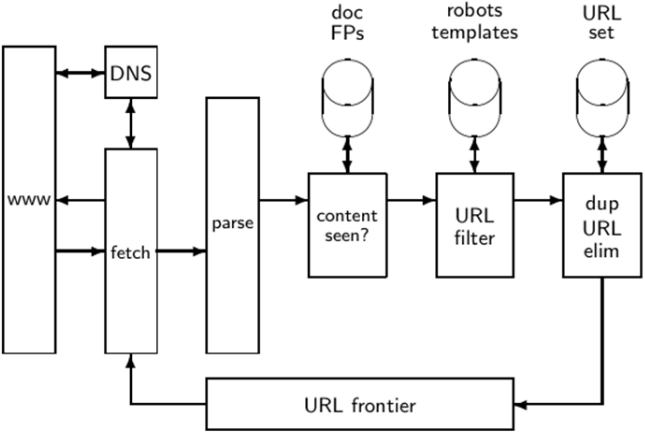
		Experiment	
		T	F
Ground Truth	T	20	10
	F	40	30

Accuracy: ((20+30)/100)*100%,
Precision: (20/60)*100%,
Recall (true positive rate or Sensitivity): (20/(20+10))*100%,
Specificity (true negative rate): (30/(40+30))*100%,
F Score: (Precision+Recall)/2,
F1 Measure: Harmonic mean of Precision and Recall

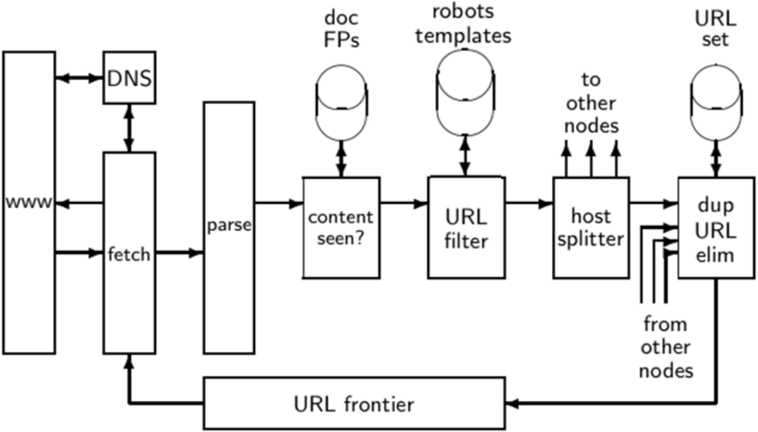
Crawling: Web Searching

- Process by which we gather pages (Also referred as spider)
  - Quickly and efficiently gather as many useful web pages as possible
  - Together with link structure
- Initialize queue with URLs of known seed pages
- Repeat
  - Take URL from queue
  - Fetch and parse page
  - Extract URLs from page
  - Add URLs to queue
- Fundamental assumption: The web is well linked
- Issues: de-duplication link and content, distribute, Spam and spider traps, Politeness and Freshness
- robots.txt (nih.gov)
  Disallow: /news/information/knight/
  Disallow: /nidcd/

Basic Crawl Architecture



Distributed Crawler Architecture



Link Analysis

- Address questions like
  - Do the links represent a conferral of authority to some pages? Is this useful for ranking?
  - How likely is it that a page pointed to by the CERN home page is about high energy physics
- Application involves to the Web, Email, Social networks
- Assumption 1: A hyperlink between pages denotes a conferral of authority (quality signal)
- Assumption 2: The text in the anchor of the hyperlink describes the target page (textual context)
- Anchor text can also be used for indexing, weighting/filtering links in the graph

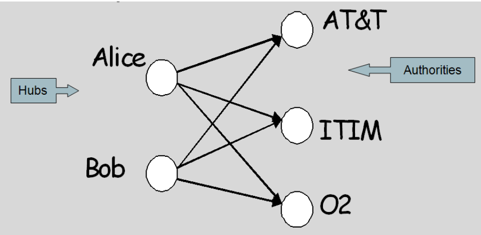
Page Rank

- A page has high rank if the sum of the ranks of its backlinks is high
- Covers both
  - A page has many backlinks
  - A page has a few highly ranked backlinks
- Let u be a web page.
  F_u the set of pages u points to.
  B_u the set of pages that point to u.
  $N_u = |F_u|$ be the number of the links from u
  Let c be a factor used for normalization
- Page rank (simplified Rank function)

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

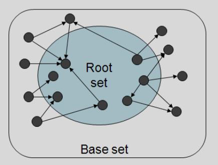
Hyperlink-Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find two sets of inter-related pages:
  - Hub pages are good lists of links on a subject. e.g., “Bob’s list of songs”
  - Authority pages occur recurrently on good hubs for the subject
- Best suited for “broad topic” queries rather than for page-finding queries. Gets at a broader slice of common opinion
- Thus, a good hub page for a topic points to many authoritative pages for that topic. A good authority page for a topic is pointed to by many good hubs for that topic



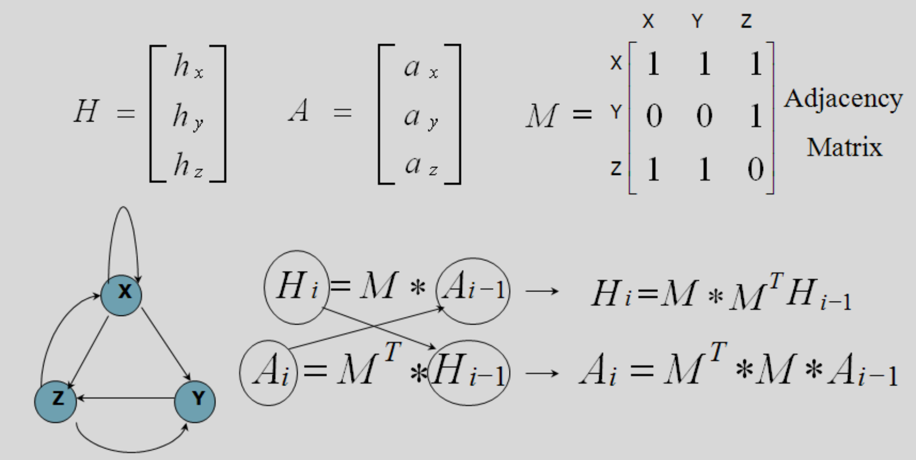
Hyperlink-Induced Topic Search (HITS)

- Construct a base set that could be good hubs or authorities
- From these, identify a small set of top hub and authority pages
- Given text query, use a text index to get all pages containing browser. Call this the root set of pages
- Add in any page that either points to a page in the root set, or is pointed to by a page in the root set. Call this the base set



- For each page x in the base set, compute hub score h(x) and authority score a(x).
  - Initialize: h(x)=1; a(x)=1; for all x
  - Iteratively update all $h(x) = sum_{y \mapsto x} a(y)$; $a(x) = sum_{x \mapsto y} h(y)$;
- Output pages with highest h() scores as top hubs, and highest a() scores as top authorities.

Example: Mini Web



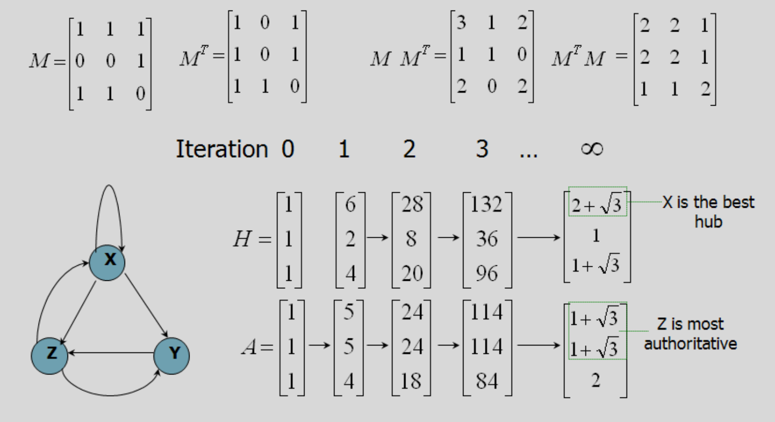
$$H = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} \qquad A = \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} \qquad M = \begin{array}{c|ccc} & X & Y & Z \\ \hline X & 1 & 1 & 1 \\ Y & 0 & 0 & 1 \\ Z & 1 & 1 & 0 \end{array}$$

Adjacency Matrix

$$H_i = M * A_{i-1} \rightarrow H_i = M * M^T H_{i-1}$$

$$A_i = M^T * H_{i-1} \rightarrow A_i = M^T * M * A_{i-1}$$

Example: Mini Web



$$M=\begin{bmatrix}1&1&1\\0&0&1\\1&1&0\end{bmatrix}\quad M^T=\begin{bmatrix}1&0&1\\1&0&1\\1&1&0\end{bmatrix}\quad M\ M^T=\begin{bmatrix}3&1&2\\1&1&0\\2&0&2\end{bmatrix}\quad M^T M=\begin{bmatrix}2&2&1\\2&2&1\\1&1&2\end{bmatrix}$$

Iteration 0 1 2 3 ... ∞

$$H=\begin{bmatrix}1\\1\\1\end{bmatrix}\ \begin{bmatrix}6\\2\\4\end{bmatrix}\rightarrow\begin{bmatrix}28\\8\\20\end{bmatrix}\rightarrow\begin{bmatrix}132\\36\\96\end{bmatrix}\longrightarrow\begin{bmatrix}\boxed{2+\sqrt{3}}\\1\\1+\sqrt{3}\end{bmatrix}$$

X is the best hub

$$A=\begin{bmatrix}1\\1\\1\end{bmatrix}\rightarrow\begin{bmatrix}5\\5\\4\end{bmatrix}\rightarrow\begin{bmatrix}24\\24\\18\end{bmatrix}\rightarrow\begin{bmatrix}114\\114\\84\end{bmatrix}\longrightarrow\begin{bmatrix}\boxed{1+\sqrt{3}}\\\boxed{1+\sqrt{3}}\\2\end{bmatrix}$$

Z is most authoritative

- To prevent the h() and a() values from getting too big, can scale down after each iteration.
- Claim: relative values of scores will converge soon: $\sim$ 5 iterations
- Ranking is based on h() and a() values

Thank You!

Thank you very much for your attention!

Queries ?