DATA WAREHOUSING AND MINING

T.E. CSE-DS, Sem V

Web Mining: Web Structure Mining

Poonam Pangarkar

Web Structure Mining

- Web structure mining can be viewed as creating a model of the Web organization.
- This can be used to classify Web pages or to create similarity measures between documents.
- Web Structure mining discovers the structure information from the web.
- This type of mining can be performed either at the intra page document level or at the inter page level.

Techniques for Web Structure Mining

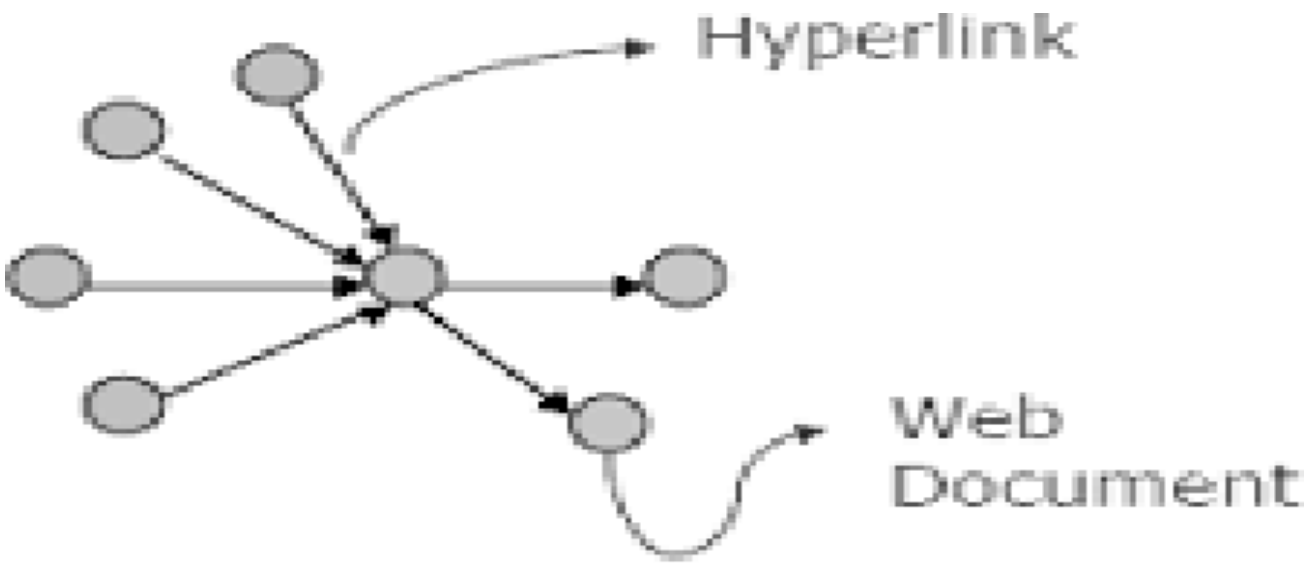
- Page Rank
- CLEVER

Techniques for Web Structure Mining

- The PageRank technique was designed to both increase the effectiveness of search engines and improve their efficiency.
- PageRank is used to measure the importance of a page and to prioritize pages returned from a traditional search engine using keyword searching.
- The PageRank value for a page is calculated based on the number of pages that point to it.
- This is actually a measure based on the number of backlinks to a page.
- A backlink is a link pointing to a page rather than pointing out from a page.
- The measure is not simply a count of the number of backlinks because a **weight** is used to provide more importance to backlinks coming from important pages.

Web Structure Mining

The Structure of a typical Web Graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.



Web Graph Structure

- In links - The hyperlinks pointing to a page
- Out links – The hyperlinks found in a page

Usually, the larger the number of in links, the better a page

Given a page p, we use B_p to be the set of pages that point to p, and F_p to be the set of links out of p.

The PageRank of a page p is defined as

$$\mathrm{PR}(p) = c \sum_{q \in B_p} \frac{\mathrm{PR}(q)}{N_q}$$

Here Nq =| Fq | .

The constant **c** is a value between 0 and 1 and is used for normalization.

CLEVER

- Developed by IBM – aims at finding best source of information.
- Finds authoritative pages and Hubs
- Authority is the best source for requested information.
- **A hub is a page that contains links to authoritative pages.**
- The Clever system identifies authoritative pages and hub pages by creating weights.
- A search can be viewed as having a goal of **finding the best hubs and authorities.**

CLEVER - HITS

- Relevant Pages with factual errors.
- Pages with higher quality contents are authoritative pages
- **H**yperlink-**I**nduced **T**opic **S**earch (HITS) finds **hubs** and **authoritative** pages.

- The HITS technique contains two components:

  - ✓ Based on a given set of keywords (found in a query), a set of relevant pages (perhaps in the thousands) is found.
  - ✓ Hub and authority measures are associated with these pages. Pages with the highest values are returned.

HITS vs Page Rank

- The difference between PageRank and HITS is that HITS is related to query and PageRank is a kind of query unrelated algorithm.
- PageRank algorithm gives each page a rank which is unique and unrelated to query keyword.

Web Usage Mining

- Web usage mining performs mining on Web usage data, or Web logs.
- A Web log is a listing of page reference data.
- Sometimes it is referred to as clickstream data because each entry corresponds to a mouse click.
- These logs can be examined from either a client perspective or a server perspective.
- When evaluated from a server perspective, mining uncovers information about the sites.
- It can be used to Improve the design of the sites.
- By evaluating a client's sequence of clicks, information about a user (or group of users) is detected.
- This could be used to perform prefetching and caching of pages.

Example

- The webmaster at ABC Corp. learns that a high percentage of users have the following pattern of reference to pages: (A, B,A, C).
- This means that a user accesses page A, then page B, then back to page A, and finally to page C.
- Based on this observation, he determines that a link is needed directly to page C from page B. He then adds this link.