Data Mining

Web Crawling (Part B)

Dr. Jason T.L. Wang, Professor

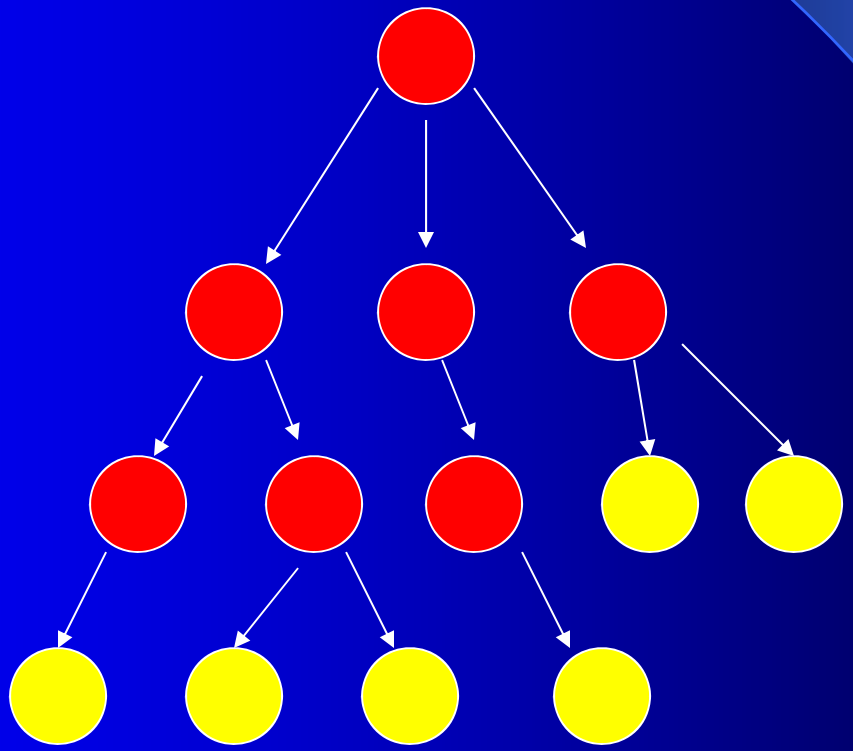
Department of Computer Science

New Jersey Institute of Technology

Where am I?

- Part A introduces Web crawling algorithms and architecture.
- Part B presents importance measures used in Web crawling algorithms.

Red Nodes: Visited Pages
Yellow Nodes: Front Pages



Importance Measures

- Document similarity measure
- Hyperlink measure
- URL measure

Document Similarity Measures

sim(P,Q) is defined as the textual similarity between P and Q, where P is a Web page and Q is the query used as the search criterion that drives the crawler. For example, the criterion can be “Find all documents related to data mining”.

To compute similarities, a vector model developed for information retrieval can be used.

P and Q are represented as t-dimensional vectors, where t is the total number of index terms in the system. The degree of similarity of P and Q is the correlation between the vectors P and Q

$$sim(P, Q) = \frac{P \cdot Q}{|P| \, |Q|}$$

If the vectors are identical, the similarity is one.

On the other hand, if the vectors have no terms in common (i.e. they are orthogonal), the similarity is zero.

Hyperlink Measure

Given a Web page P, there are two types of links with respect to P:

- *Back-links* or *in-links* are hyperlinks that are linked to P.
- *Forward-links* or *out-links* are hyperlinks that are linked from P.
- Back-link count: the number of back-links to P.
- Forward-link count: the number of forward-links from P.
- Importance of P: the number of links to P.
- Weighted importance: a page with many back-links from important pages should be important.

URL Measure

Importance is a function of its hyperlink location (URL), rather than its text contents.

E.g. URLs ending with ".com" have different weights from those with ".net".

E.g. A URL starting with "www" and "home" may be more important since it represents the homepage of a Website.

Topic-Oriented Web Crawling (TOWC)

TOWC algorithms start with a seed-set (or root-set) of relevant pages, and then attempt to efficiently seek out relevant documents based on a combination of link structure and page content analysis, using criteria such as:

- in-degree link count,
- out-degree link count,
- authority score,
- hub score,
- topic-relevant score.

Topic-relevant scoring functions use techniques such as:

- Text classification: to classify and rank the relevance of a Web page to a given topic and build a topic hierarchy;
- Machine learning: to classify document vectors in multi-dimensional space;
- PageRank measures: to compute a weighted hyperlink measure, which is intended to be proportional to the quality of the page containing the hyperlink.

TOWC has been successfully used to efficiently generate indexes for search portals and user groups.

End of

Web Crawling Module (Part B)