Data Mining

Web Mining II – Web Structure Mining (Part C)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Where am I?

- Part A introduces the definitions, terms and concepts for connectivity analysis and Web structure mining.

- Part B presents the HITS algorithm for connectivity analysis.

- Part C presents a summary of mining Web's link structures.

Mining Web's Link Structures

- Finding authoritative Web pages
  - Retrieving pages that are not only relevant, but also of high quality, or authoritative on the topic.
- Hyperlinks can infer the notion of authority.
  - The Web consists not only of pages, but also of hyperlinks pointing from one page to another.
  - These hyperlinks contain an enormous amount of latent human annotation.
  - A hyperlink pointing to another Web page – this can be considered as the author's endorsement of the other page.

Mining Web's Link Structures

- Problems with the Web linkage structure
  - Not every hyperlink represents an endorsement.
    - Other purposes are for navigation or for paid advertisements.
    - If the majority of hyperlinks are for endorsement, the collective opinion will still dominate.
  - One authority will seldom have its Web page point to its rival authorities in the same field.
  - Authoritative pages are seldom particularly descriptive.
- Hub
  - Set of Web pages that provides collections of links to authorities

HITS (Hyperlink-Induced Topic Search)

- Explore interactions between hubs and authoritative pages.
- Use an index-based search engine to form the root set.
  - Many of these pages are presumably relevant to the search topic.
  - Some of them should contain links to most of the prominent authorities.
- Expand the root set into a base set.
  - Include all of the pages that the root-set pages link to, and all of the pages that link to a page in the root set, up to a designated size cutoff.
- Apply weight-propagation
  - An iterative process that determines numerical estimates of hub and authority weights

Systems Based on HITS

Output a short list of the pages with large hub weights, and the pages with large authority weights for the given search topic.

- Systems based on the HITS algorithm
  - CLEVER, Google: achieve better quality search results than those generated by term-index engines such as AltaVista and those created by human ontologists such as Yahoo!.
- Difficulties from ignoring textual contexts
  - Drifting: when hubs contain multiple topics
  - Topic hijacking: when many pages from a single Web site point to the same single popular site

End of Web Structure Mining Module (Part C)