Data Mining

Web Mining I – Web Usage Mining (Part A)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Mining the World-Wide Web

- The WWW is a huge, widely distributed, global information service centre for
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.;
  - Hyper-link information;
  - Access and usage information.
- WWW provides rich sources for data mining.
- Challenges
  - Too huge for effective data warehousing and data mining
  - Too complex and heterogeneous: no standards and structure

Mining the World-Wide Web

- Web is growing and changing very rapidly and user communities are very diverse.
- Only a small portion of the information on the Web is truly relevant or useful.
  - 99% of the Web information is useless to 99% of Web users.
  - How can we find high-quality Web pages on a specified topic?

Web Mining: a challenging task

- Searches for
  - Web access patterns
  - Web structures
  - Regularity and dynamics of Web contents
- Problems
  - The “abundance” problem
  - Limited coverage of the Web: hidden Web sources, majority of data in DBMS
  - Limited query interface based on keyword-oriented search
  - Limited customization to individual users

Web Mining Taxonomy

Web content mining -- automatic discovery of Web document content patterns

- Web Page Content Mining
- Search Result Mining

Web usage mining -- automatic discovery of Web server access patterns

- General Access Pattern Tracking
- Customized Usage Tracking

Web structure mining -- automatic discovery of hypertext/linking structure patterns

Web Content Mining

Web content mining is the process of analyzing text and graphic contents on the Web. It has roots in information retrieval and natural language processing.

Previously unknown or hidden patterns can be extracted using this process. Text content mining is similar to the text mining discussed earlier and graphics content mining is similar to multimedia content-based retrieval.

Web Usage Mining

Web usage mining is the process of analyzing Web access information available on Web servers. It consists of the automatic discovery of user access patterns from the large collections of access logs, which are daily generated by Web servers.

Analyzing Web usage data and access logs can help organizations study customers' Web browsing patterns to facilitate e-commerce specific processing such as the following:

Web structure management – for designing a better Website

Promotional campaigns – for building customized advertisements

Marketing decisions – for making better strategic decisions

Web Structure Mining

Web structure mining is the process of discovering hypertext/linking structure patterns as well as hub and authoritative web pages in the Web.

End of
Web Usage Mining Module (Part A)