WEB MINING

PRESENTED BY:
KOTLA SAIKIRAN
18B81A0436
ECE

Web Mining

- Web Mining is the use of the data mining techniques to automatically discover and extract information from web documents/services
- Discovering useful information from the World-Wide Web and its usage patterns
- My Definition: Using data mining techniques to make the web more useful and more profitable (for some) and to increase the efficiency of our interaction with the web

Web Mining

- The WWW is 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 of data for data mining

Web Mining

- Data Mining Techniques
  - Association rules
  - Sequential patterns
  - Classification
  - Clustering

Classification of Web Mining Techniques

- Web-Structure Mining
- Web-Usage Mining
- Web-Content Mining

Web-Structure Mining

- Generate *structural summary* about the Web site and Web page

Depending upon the hyperlink, ‘Categorizing the Web pages and the related Information @ inter domain level

Discovering the Web Page Structure.

Discovering the nature of the hierarchy of hyperlinks in the website and its structure.

Web-Usage Mining

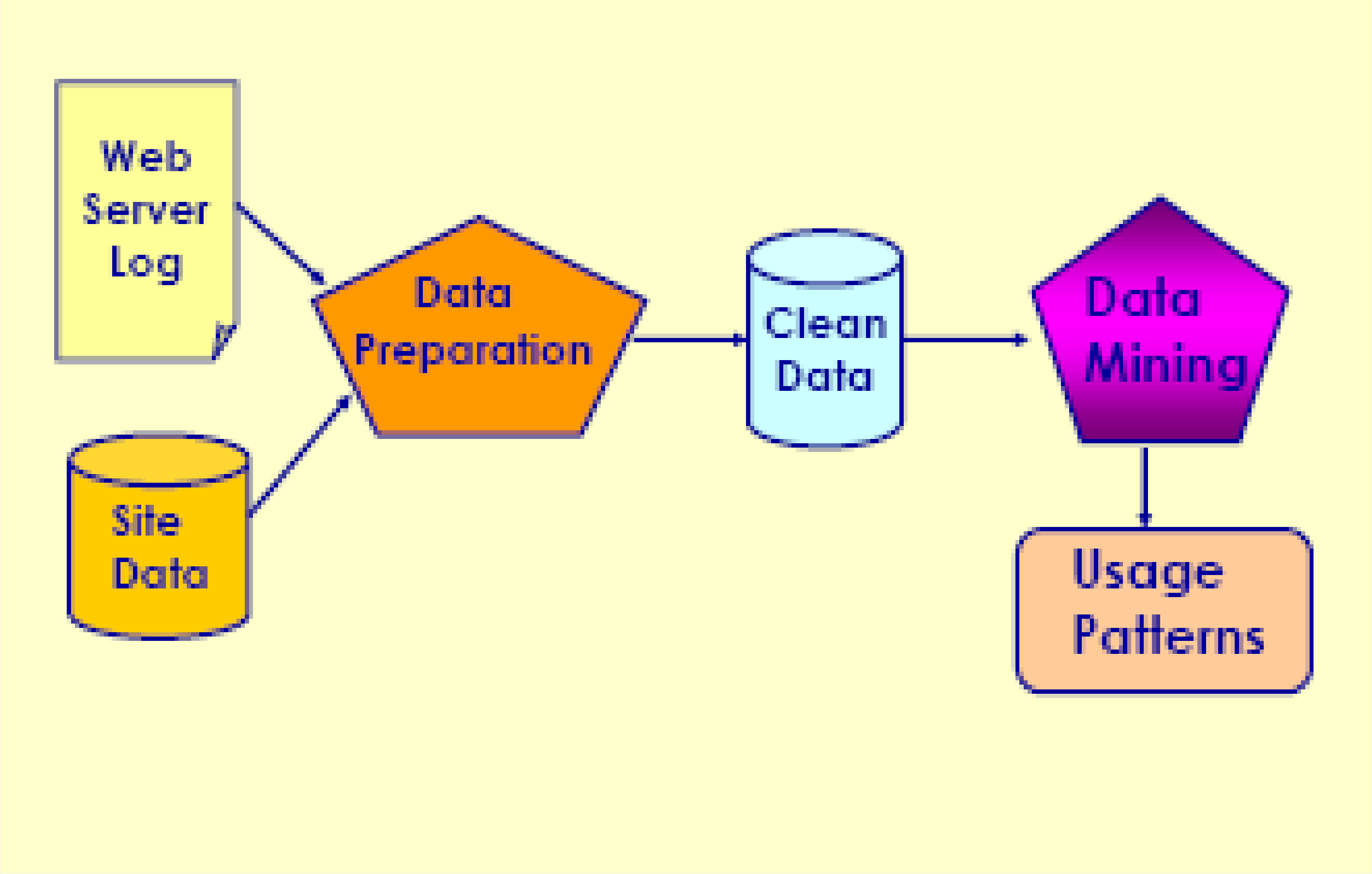
- What is Usage Mining?

Discovering user 'navigation patterns' from web data.

Prediction of user behavior while the user interacts with the web.

Helps to Improve large Collection of resources.

Web Usage Mining Process



Web Usage Mining

- Search Engines
- Personalization
- Website Design

Web Content Mining

- ***'Process of information'*** or resource discovery from content of millions of sources across the World Wide Web
  - E.g. Web data contents: text, Image, audio, video, metadata and hyperlinks
- Goes beyond key word extraction, or some simple statistics of words and phrases in documents.

Why Mine the Web?

- Enormous wealth of information on Web
    - Financial information (e.g. stock quotes)
    - Book/CD/Video stores (e.g. Amazon)
    - Restaurant information (e.g. Zagats)
    - Car prices (e.g. Carpoint)

- Lots of data on user access patterns
    - Web logs contain sequence of URLs accessed by users

- Possible to mine interesting nuggets of information
    - People who ski also travel frequently to Europe
    - Tech stocks have corrections in the summer and rally from November until February

User Profiling

- Important for improving customization
  - Provide users with pages, advertisements of interest
  - Example profiles: on-line trader, on-line shopper
- Generate user profiles based on their access patterns
  - Cluster users based on frequently accessed URLs
  - Use classifier to generate a profile for each cluster
- Engage technologies
  - Tracks web traffic to create anonymous user profiles of Web surfers
  - Has profiles for more than 35 million anonymous users

Problems with Web Search Today

- Today’s search engines are plagued by problems:
  - the *abundance* problem (99% of info of no interest to 99% of people)
  - *limited coverage* of the Web (internet sources hidden behind search interfaces)

    Largest crawlers cover < 18% of all web pages
  - *limited query* interface based on keyword-oriented search
  - *limited customization* to individual users

Problems with Web Search Today(cont.)

- Today’s search engines are plagued by problems:
  - Web is highly dynamic
    - Lot of pages added, removed, and updated every day
  - Very high dimensionality

Web Mining Issues

- Size
  - Grows at about 1 million pages a day
  - Google indexes 9 billion documents
  - Number of web sites
    - Netcraft survey says 72 million sites

    (http://news.netcraft.com/archives/web_server_survey.html)
- Diverse types of data
  - Images
  - Text
  - Audio/video
  - XML
  - HTML

Web Mining Applications

- E-commerce (Infrastructure)
  - Generate user profiles
  - Targetted advertizing
  - Fraud
  - Similar image retrieval
- Information retrieval (Search) on the Web
  - Automated generation of topic hierarchies
  - Web knowledge bases
  - Extraction of schema for XML documents
- Network Management
  - Performance management
  - Fault management

Retrieval of Similar Images

- Given:
  - A set of images
- Find:
  - All images similar to a given image
  - All pairs of similar images
- Sample applications:
  - Medical diagnosis
  - Weather predication
  - Web search engine for images
  - E-commerce

Conclusion

- **Major limitations of Web mining research:**
  - Lack of suitable test collections that can be reused by researchers.
  - Difficult to collect Web usage data across different Web sites.
- **Future research directions**:
  - Multimedia data mining: a picture is worth a thousand words.
  - Multilingual knowledge extraction: Web page translations
  - Wireless Web: WML and HDML.
  - The Hidden Web: forms, dynamically generated Web pages.
  - Semantic Web

References

- Mining the Web: Discovering Knowledge from Hypertext Data by Soumen Chakrabarti (Morgan-Kaufmann Publishers)
- Web Mining :Accomplishments & Future Directions by Jaideep Srivastava
- The World Wide Web: Quagmire or goldmine by Oren Entzioni
- http://www.galeas.de/webmining.html

THANK YOU