# Web mining and knowledge discovery of usage patterns - A survey

**CS748**

**Yan Wang**

# Introduction

- Web data mining
- Usage mining on the Web
- WebSIFT: a usage mining system
- Personalization vs. User navigation pattern
- Privacy on the Web
- Research issues
- Conclusion

# Web data mining

- Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services.
- Subtasks: resource discovery, information extraction, generalisation, analysis.

# Web Data mining

- **Web content mining**: mining the data on the Web
- **Web structure mining**: mining the Web structure data
- **Web usage mining**: mining the Web log data

## Web Content Mining

- The technique, involving mining web data contents, describes the automatic search of information resource available online
- Web data contents: text, image, audio, video, metadata and hyperlinks.
- Most of the Web documents are non-structured text data, some of them are semi-structured like HTML files, or more structured like the table data and database generated HTML pages

# Web content mining

- Information retrieval techniques are taken to deal with the unstructured and semi-structured data mining.
- Database view of Web data is used to have the better information management and querying on the Web, the mining always tries to infer the structure of the Website to transform  a Website to become a database.

# Web Structure Mining

- Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level to generate structural summary about the Website and Web page.
- Direction 1: based on the hyperlinks, categorizing the Web pages and generated the information.
- Direction 2: discovering the structure of Web document itself.
- Direction 3: discovering the nature of the hierarchy or network of hyperlinks in the Website of a particular domain.

# Web Usage Mining

- Web usage mining is the application of data mining techniques to discover usage patterns from the secondary data derived from the interactions of the users while surfing on the web, in order to understand and better serve the needs of Web-based applications.
- Three distinctive phases: preprocessing, pattern discovery, and pattern analysis

# Web Usage Mining-preprocessing

- Preprocessing is the process to convert the raw data into the data abstraction necessary for the further applying the data mining algorithm.
- Resources: server-side, client-side, proxy servers, or database.
- Raw data: Web usage logs, Web page descriptions, Web site topology, user registries, and questionnaire.
- Conversion: Content converting, Structure converting, Usage converting

# Web Usage Mining-preprocessing

- User: The principal using a client to interactively retrieve and render resources or resource manifestations.
- Page view: Visual rendering of a Web page in a specific client environment at a specific point of time
- click stream: a sequential series of page view request
- User session: a delimited set of user clicks (click stream) across one or more Web servers.

# Web Usage Mining-preprocessing

- Server session (visit): a collection of user clicks to a single Web server during a user session.
- Episode: a subset of related user clicks that occur within a user session.

# Web Usage Mining-preprocessing

- Content Preprocessing: the process of converting text, image, scripts and other files into the forms that can be used by the usage mining.
- Structure Preprocessing: The structure of a Website is formed by the hyperlinks between page views, the structure preprocessing can be done by parsing and reformatting the information.

# Web Usage Mining-preprocessing

- Usage Preprocessing is arguably the most difficult task in the usage mining processes, the data cleaning techniques are always necessary to eliminate the impact of the irrelevant items to the analysis result.
- The input data may include the Web server logs, referral logs, registration files, index server logs, and optionally usage statistics topology, and page classification.

# Web Usage Mining-Pattern Discovery

- Pattern Discovery is the key component of the Web mining, which converges the algorithms and techniques from data mining, machine learning, statistics and pattern recognition etc research categories.
- Separate subsections: statistical analysis, association rules, clustering, classification, sequential pattern, dependency Modelling.

# Web Usage Mining-Pattern Discovery

- Statistical Analysis: the analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file, the statistical techniques are the most powerful tools in extracting knowledge about visitors to a Web site.
- Association Rules: refers to sets of pages that are accessed together with a support value exceeding some specified threshold. This technique can be used to discover unordered correlation between items found in a database of transactions.

# Web Usage Mining-Pattern Discovery

- Clustering: a technique to group together users or data items (pages) with the similar characteristics. It can facilitate the development and execution of future marketing strategies.
- Classification: the technique to map a data item into one of several predefined classes, which help to establish a profile of users belonging to a particular class or category.

# Web Usage Mining-Pattern Discovery

- Sequential Pattern: this technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions of episodes. It helps web marketer to predict the future trend.
- Dependency Modelling: this technique provides a theoretical framework for analyzing the behaviour of users, and is potentially useful for predicting future Web resource consumption.

# Web Usage Mining-Pattern Analysis

- Pattern Analysis is the final stage of the Web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process.
- Analysis methodologies and tools: query mechanism like SQL, OLAP, Visualization etc.
- This is a very fertilized research area.

# WebSIFT

- Web Site Information Filter System (WebSIFT) is a Web usage mining framework, that uses the content and structure information from a Web site, and finally identify the interesting results from mining usage data.
- WebSIFT divides the Web usage mining process into three principal parts that are corresponding to the three phases of usage mining: preprocessing, pattern discovery, pattern analysis

# WebSIFT

- Input of the mining process: server logs (access, referrer, and agent), HTML files, optional data.
- Preprocessing process: construct a user session file with the input data to derive a site topology and to classify the pages of a site. The user session file will be converted to the transaction file and output to next phase.

# WebSIFT

- Pattern discovery process: use the techniques such as statistics, association rules, clustering , sequential to generate rules and patterns
- Pattern analysis process: the site topology and page classification are fed into the information filter, the output of pattern discovery will be analyzed by using procedural SQL and OLAP.

# Personalization

- Personalization: the provision to the individual of tailored products, services, information or information relating to products or service.
- Goal: to provide users with what they need or want without explicit indication.
- Categories: manual decision rule systems, collaborative filtering system, and content-based filtering system

# Personalization (applications)

- Customizing access to information sources
- filtering news or e-mails
- recommendation services for the browsing process
- tutoring systems
- Search
- more ...

# User Navigation Pattern

- User navigation pattern discovery is the technique to learn the user behavior pattern when navigating within a web site.
- Goal: personalize web page for individule user; improving the site’s static structure of the underlying hypertext system.
- Two aspects: the interests of the users and the accessed information, solved by constructing user profiles; the way of accessing the information, solved by analyzing web server logs.

# Privacy

- Most critical issues in the internet society.
- Unavoidable conflict between web data mining requirement and user privacy desires.
- Solutions: privacy legislation and technique development

# Research Issues

- multimedia data content mining
- reduce the subjectiveness of the profile data
- eliminate the out of date user data
- temporal analysis such as trend analysis, change point detection or similarity analysis
- pattern analyzing tools

# Conclusions …