Research Trends In AI

1) Web Mining: Information and Pattern discovery on the World Wide Web
   R. Cooley, B. Mobasher, and J. Srivastava

2) Web Mining Research: A Survey
   R. Kosala, H. Blockeel

Presented By: Saadiya Waheed Raza
              02030050
What is Web Mining?

- Web Content Mining
  - Automatic search of information resources online

- Web Usage Mining
  - The discovery of user access patterns from Web servers
Web Usage Mining Architecture

Data Cleaning

Transaction Identification

Data Integration

Transformation

Pattern Discovery

Pattern Analysis

Server Log Data

Clean Log

Transaction Data

Formatted Data

Path Analysis

Association Rules

Sequential Patterns

Clusters & Classification Rules

OLAP/Visualization Tools

Knowledge Query Mechanism

Intelligent Agents

Document and Usage Attributes

Database Query Language
Data Cleaning

- Eliminate irrelevant items from server logs
  - Outlier Analysis

- Identify distorted user traversal patterns
  - Ignore users who cannot be uniquely identified?
Transaction Identification

- A user session or transaction is all of the page references made by a user during a single visit to the site
  - The problem here is again due to proxy servers
  - Use heuristics based on a combination of IP address, machine name, browser agent and temporal information.
Data Integration

- Combining multiple logs
  - Non-scalable
  - Impractical
- Currently, knowledge mined from various logs is integrated.
  - This is not a pre-processing task
  - Trade – off between Micro & Macro Trends
Pattern Discovery

- Path Analysis
  - Frequent traversal patterns
  - Large referral sequences

Example:
65% of clients left the site after 4 or less page references

70% of clients who accessed /company/product2 did so by starting at /company and proceeding through /company/new and /company/product1
Pattern Discovery Cont’d

- Association Rules
  - A set of items in one transaction implies the presence of other items

Example:
40% of clients who accessed the web page /company/product1 also accessed /company/product2
Pattern Discovery Cont’d

- Sequential Patterns
  - Traversal pattern such that one traversal is usually followed by another with reference to their time stamp.

Example:
30% of clients who visited /company/products had done a search in Yahoo, within the past week on keyword \( w \).
Pattern Discovery Cont’d

- Classification
  - Develop a profile of items belonging to a particular group

Example:
Clients from government agencies tend to be interested in /company/product1
Pattern Analysis

- OLAP
  - Size of server logs
- Intelligent Agents
  - Improve topology
  - Identify trends
- High-Level Query Language

```
SELECT association-rules(A*B*C*)
FROM   log.data
WHERE  date >= 970101 AND domain = "edu"
       AND support = 1.0 AND confidence = 90.0
```
Comparison of the two surveys

- Data Cleaning
- Transaction Identification
- Data Integration
- Transformation
- Pattern Discovery
- Information Extraction
- Information Retrieval
- Pattern Analysis
End Note

- Questions
- Thank-You
References


