A RESEARCH SUPPORT SYSTEM FRAMEWORK FOR WEB DATA MINING

Jin Xu, Yingping Huang, Gregory Madey
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556

WSS'03: WI/IAT 2003 Workshop on Applications, Products of Web-based Support Systems
October 13, 2003, Halifax

This research was partially supported by NSF, CISE/IIS-Digital Society and Technology
OUTLINE

- INTRODUCTION
- FRAMEWORK OVERVIEW
- INFORMATION RETRIEVAL
- DATA MINING TECHNIQUES
- CASE
- CONCLUSIONS & FUTURE WORK
INTRODUCTION

- **World Wide Web**
  - Abundant information
  - Important resource for research

- **Web Data Features**
  - Semi-structured
  - Heterogeneous
  - Dynamic

- **A Research Support System for Web Data Mining**
FRAMEWORK

Source Identification → Content Selection → Information Retrieval → Data Mining
INFORMATION RETRIEVAL

- Searching Tools
  - Directory
  - Search engine
- Web Crawler
  - URL access method
  - Web page parser
    - Table extractor
    - Link extractor – absolute links/relative links
    - Word extractor
DATA MINING FUNCTIONS

- Association Rules
  - Find interesting association or correlation relationship among data items

- Classification
  - Predict classes
  - Two steps – build model, apply model

- Clustering
  - Find natural groups of data
OPEN SOURCE SOFTWARE

- Open Source Software (OSS)
  - Apache, Perl, Linux
  - Developed by part time contributors
- SourceForge Developer Site
  - Sponsored by VA Software
  - Largest OSS development site
    - 70,000 projects
    - 90,000 developers
    - 700,000 users
DATA COLLECTION

- **Data sources**
  - Statistics, forums

- **Project statistics**
  - 9 fields – project ID, lifespan, rank, page views, downloads, bugs, support, patches and CVS

- **Developer statistics**
  - Project ID and developer ID
DATA COLLECTION (Cont.)

- Web Crawler
  - Perl and CPAN
  - LWP – fetch pages
  - HTML parser – parse pages
  - HTML::TableExtract – extract information
  - Link extractor – extract links
DATA MINING

- Association Rules
  - “all tracks”, “downloads” and “CVS” are associated

- Classification
  - Predict “downloads”
  - Naïve Bayes – Build Time 30 sec, accuracy 9%
  - Adaptive Bayes Network - Build Time 20 min, accuracy 63%

- Clustering
  - K-means: User specified number of clusters
  - O-cluster: Automatically detect the number of clusters
CONCLUSIONS

- Conclusions
  - Build a framework
  - Describe procedures
  - Discuss techniques
  - Provide a case study

- Future Work
  - Exploratory study
  - Implement all stages