Web Data Mining

Ming-Syan Chen

Data Mining in the Web

- Applications are ported to the Web at rapid pace
- On-line services, such as America Online (AOL), CompuServe (merged to AOL), Prodigy, are anxious to know user access patterns; not just “search” in the Web
- How Amazon does it
Data Mining in the Web (cont’d)

• Important Web applications that can benefit from data mining
  – Web advertising
  – Electronic Commerce
  – On-line transaction/trading

• As the number of WWW applications increases, there is a strong interest to capture user behavior
Web Warehousing

• An approach to the building of computer systems which have as their primary functions the identification, cataloging, (possibly) storage, and analysis of information (in their multimedia objects) through the use of Web technology in order to help individuals to find information they need and analyze it effectively.

Capture Web User Behavior

• Understanding Web user behavior is important
  – improving Web page organization
  – increasing Web server performance
  – exploiting Web advertising
  – increasing business opportunity
More Information Desired

- Collect statistical information (page hits) only, which is insufficient since:
  - the hit freq. of a page depends not only on its content but also on its location
  - the number of users accessing a page is not available
  - info. on what pages accessed together is not available
Data mining in the Web

- Data mining in the Web
  - *Web access pattern collection*
  - Web user pattern mining
  - Web data manipulation and query

Web Access Pattern Collection

- Server-based data collection
  - Who are visiting a given Web site and what are they doing
- Client-based data collection
  - What are the Web sites a particular user has visited?
- Example: the use of an Access Pattern Collection Server (APCS)
Server-Based Data Collection

- Examine the logs collected by HTTPd
  - access log, referral log, agent log, error log...
pending on browser type
  - combining some of them for our use if necessary
  - Session identification from access log is difficult because
    - the use of proxy servers
    - the effect of caching.

File Structure

- File structure of HTTPd
  - Subdirectories under httpd:
    - cgi-bin;
    - cgi-src;
    - htdocs;
    - src;
    - conf;
    - logs
Web Log

- flea.cs.kobe-u.ac.jp - - [12/Apr/1999:11:31:40 +0800] "GET /dasfaa99/New.gif HTTP/1.0" 200 2572
- 209.185.143.91 - - [12/Apr/1999:11:31:51 +0800] "GET / HTTP/1.0" 200 2357
- flea.cs.kobe-u.ac.jp - - [12/Apr/1999:11:32:41 +0800] "GET /dasfaa99/final_program.html HTTP/1.0" 200 12246
- 163.221.174.24 - [12/Apr/1999:11:32:59 +0800] "GET /dasfaa99/New.gif HTTP/1.0" 200 2572
- alpha047.aist-nara.ac.jp - - [12/Apr/1999:11:33:36 +0800] "GET /dasfaa99/ HTTP/1.0" 304 -
Access Log: a b d e c
Referred Log: (a,b) (b,d) (b,e) (a,c)
Possible Traversal Path: a b d b e b a c

On Server-Based Collection

- Greedy methods devised for path traversal patterns in the Web
- Have to be done in accordance with technology advances:
  - e.g., the use of Active Server Pages (session ID available), HTTP 1.1, etc
- Can only capture the user behavior when they are within this site
Client-Based Data Collection

- Understanding individual Web behavior needs client-based data mining
- Results are useful for better personalized service, improved Web page organization, better pricing policies, etc

Methods to Collect Web Data

- Applets can only read/write files in their source servers
  - a big security constraint
- Using Active components and PlugIn’s
  - APCS (access pattern collection server)
    ICTAI99
Use of an APCS

- Design and implementation of an Access Pattern Collection Server (APCS)
- Enciphering module
  - use encryption and decryption to ensure the devised mechanism not to be bypassed
- Deciphering: decoding; user pattern collection and analysis

---

Use of an APCS (cont’d)

- Page conversion
  - Convert index.html to index.apc (enciphered Web page)
- Software component
  - Active control (ocx) or Plugin
  - to decode, report and display
Enciphering Module

Origional HTML file

index.html

RC4 cipher of origional file

index.apc

the new index.html contains reference to index.apc

index.apc wrapper which replaces origional HTML file

index.html

Origional Web Page

Enciphering Module

APCS Server

Deciphering Module

Enciphered Web Page

Web Server

Welcome!

Web Browser
Web Server 1 access log

proxy.net - "GET index.html HTTP/1.0" -
proxy.net - "GET index.html HTTP/1.0" -

APCS 1 access log

browser1.net - "GET http://ws1.net/index.apc HTTP/1.0" -
browser2.net - "GET http://ws1.net/index.apc HTTP/1.0" -
Web Server 2 access log
proxy.net - "GET index1.html HTTP/1.0" -
proxy.net - "GET index2.html HTTP/1.0" -

Web Server 3 access log
proxy.net - "GET index1.html HTTP/1.0" -
proxy.net - "GET index2.html HTTP/1.0" -

APCS 2 access log
browser3.net – “GET http://ws2.net/index1.apc HTTP/1.0” -
browser3.net - “GET http://ws3.net/index1.apc HTTP/1.0” -
browser3.net - “GET http://ws2.net/index2.apc HTTP/1.0” -
browser3.net - “GET http://ws3.net/index2.apc HTTP/1.0” -
Remarks on Client-Based Data Collection

- Very difficult to do for non-registered users in the current Web environment
  - have to be conducted with users’ consent
- Very dependent upon available Web technologies

Data mining in the Web

- Data mining in the Web
  - Web access pattern collection
    - *Web user pattern mining*
  - Web data manipulation and query
Web User Pattern Mining

- Web user pattern mining is to discover user access patterns in Web servers
- Pattern discovery and analysis tools
  - some existing Web tools provide mechanisms for reporting user activity in the servers and conduct data processing, e.g., www.webtrends.com, www.openmarket.com, www.netgen.com, etc

WebTrends

- **Products**
  - WebTrends for Windows NT/95/98, UNIX, Lotus Domino, Proxy Server, Firewall, Novell NetWare
- **Functions**
- **Sample Reports**
  - Summary Report, Advertising Report, Proxy Analysis, Real Server Profile ...
- **URL**
Open Market eBusiness Solutions

- **Products**
  - IPS (Internet Publishing System): an e-publishing platform which includes design and development, content management, publication management, content acquisition, security, and process control/workflow
  - Transact: an EC software for transaction processing, subscription administration, order management, and customer service
  - ShopSite: an online store creation system that allows merchants to build and maintain catalogs of products or services to sell over the Internet
  - LiveCommerce: an e-catalog system which makes it easy for customers to find the products they want to purchase

- **URL**

net.Genesis

- **Product**
  - net.Analysis: it gathers data about online customers by using the APS (Automated Publishing System). Web managers can automatically retrieve, import, compress, and store Web server logs on a daily, weekly, or monthly basis. APS offers end-to-end automation of data collection and analysis to report on production and distribution.

- **URL**
Pattern Discovery and Analysis Tools

- Tools for user pattern discovery employ techniques from database, AI, psychology, and information theory, to discover knowledge from collected data
  - WebViz [Pitk94], (Dyresons) [Dyre97], WEBMINER [Moba96], WebLogMiner [Zaia98a] [Zaia98b]

WEBMINER

- Development of a Web mining architecture
- Modelling user tx’s, each of which consists of multiple log entries
- Clustering algorithms to group log entries into transactions
WebLogMiner

- Database construction from server log file:
  - data cleaning
  - data transformation
- Multi-dimensional Web log data cube construction and manipulation
- Data mining on Web log data cube and Web log database

Mining for Web Transactions

- A data mining capability which provides knowledge discovery for Web transactions generated from electronic commerce services
- To capture Web customer buying behavior [compsac00]
  - not just market basket transaction for the set of items bought by a customer in a single purchase
  - not just Web user travel patterns
  - extension from path traversal patterns
Web Transaction

Web-Transaction

<table>
<thead>
<tr>
<th>Node</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>3, 4</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>G</td>
<td>-</td>
</tr>
<tr>
<td>H</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>S</td>
<td>7, 8</td>
</tr>
<tr>
<td>J</td>
<td>-</td>
</tr>
<tr>
<td>L</td>
<td>9</td>
</tr>
<tr>
<td>S</td>
<td>-</td>
</tr>
<tr>
<td>I</td>
<td>-</td>
</tr>
</tbody>
</table>

Web-Transaction Record

Web-Transaction-Record

<table>
<thead>
<tr>
<th>Path</th>
<th>Node {item} s</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCE</td>
<td>B{1}, C{2}, E{3}, E{4}</td>
</tr>
<tr>
<td>ABFGH</td>
<td>B{1}, H{6}</td>
</tr>
<tr>
<td>ASIL</td>
<td>S{7}, S{8}, L{9}</td>
</tr>
</tbody>
</table>
Enhance Database Function

• Exploring the relationship between travelling and buying
• Find all rules that have “bread” as consequent with the specific paths
  – What the store should do to boost the sale of bread in the appropriate Web pages

Enhance Database (cont’d)

• Find all rules that have “milk” in the antecedent with the specific paths
  – What product may be affected if the store remove the Web page that sells milk
• Find all the rules relating items located on Web pages A and B with the specific paths
  – May help arranging Web pages by learning if the sale of items on Web pages A is related to the sale of items on Web pages B.
Example

- \(<ABCE : B\{1\}, E\{4\}> = 2\)
- \(<AB : B\{1\}> = 3\)
- We can derive \(<ABCE : B\{1\} \Rightarrow E\{4\}>\)
  - \(\text{sup}(<ABCE : B\{1\} \Rightarrow E\{4\}>) = 2\)
  - \(\text{confidence}(<ABCE : B\{1\} \Rightarrow E\{4\}>) = \frac{\text{sup}(<ABCE : B\{1\}, E\{4\}>)}{\text{sup}(<AB : B\{1\}>)} = \frac{2}{3} = 67\%\)

Data mining in the Web

- Data mining in the Web
  - Web access pattern collection
  - Web user pattern mining
  - \textit{Web data manipulation and query}
Web Data Manipulation and Query

- Agent-based approaches:
  - Intelligent agents
  - Data filtering and categorization
- Server-based approaches:
  - Multilevel databases
  - Web query systems

Agent-Based Approach

- Use intelligent agents to locate documents and services in the Web
  - WebCrawler, Alta Vista, MetaCrawler, ShopBot [Door96], Harvest [Brow94], FAQ-Finder [Hamm95], Information Manifold [Kirk95], OCCAM [Kwok96], Parasite [Sper97], ILA (Internet Learning Agent) [Perk95]
- Data filtering and categorization
  - HyPursuit [Weis96], BO (Bookmark Organizer) [Maar96]
Personalized Service

- With the knowledge of individual user behavior and preference, a well-designed Web site can deliver personalized service (using proper data preprocessing and filtering)
  - Recommendation system [agga98c]
  - Dynamic Web page under ASP and PHP is ready
  - Relevant works: WebWatcher [Arms95], PAINT [Oost94], Syskill&Webert [Pazz96], GroupLens [Resn94], Firefly [Shar95], others

Multilevel Databases

- At the higher levels, meta data or generalizations are extracted from lower levels, and organized in structured formats
- At the lowest level, semi-structured information are stored in various Web repositories
- Various relevant works
  - MLDB [Zaia95], [Khos96], [King96], ARANEUS system [Meri97]
Web Query Systems

- Web-based query systems
- Languages utilize
  - Standard database query languages such as SQL,
  - Structural information about Web documents
  - Natural language processing for the queries that are used in WWW searches
- Relevant works: TSIMMIS [Chaw94], Lorel [Quas95], UnQL [Bune96], WebLog [Laks96], W3QL [Kono98a] [Kono98b]

W3QL and W3QS

- W3QL
  - Able to query WWW page data and structural hypertext information
  - Extensible, able to integrate with UNIX utilities as well as user written program to analyze WWW page contents
- W3QS
  - A prototype for managing W3QL queries
  - Providing a library of PERL objects that can be used to implement new search algorithms
W3QL: Search a Site

Select
From n1,l1,(n2,l2),l3,n3  /* unbounded length path */
Where
n1 in {http://www.cs.technion.ac.il};
l1 in {/technion\ac\il/};
l2 in {/technion\ac\il/};
n3: PERLCOND ‘n3.format =~ /image/’; /* file analyzer */
n3 in {/technion\ac\il/}
Using ISEARCHd -d 5 -l 1000 /* search algorithm */

Query Example by W3QL

Select n2:
From n1,l1,n2
Where n1 in MyHomePage;
n2 in Technion
Using ISEARCHd
Query Results

<table>
<thead>
<tr>
<th>n1</th>
<th>n2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A HREF=&quot;<a href="http://www.cs.technion.ac.il/~konop/w3qs.html">http://www.cs.technion.ac.il/~konop/w3qs.html</a>&quot;</td>
<td>A HREF=&quot;<a href="http://www.technion.ac.il">http://www.technion.ac.il</a>&quot;</td>
</tr>
<tr>
<td>img src=&quot;Portrait.jpg&quot;</td>
<td><a href="http://www.cs.technion.ac.il/~konop/portrait.jpg">http://www.cs.technion.ac.il/~konop/portrait.jpg</a></td>
</tr>
<tr>
<td>IMG SRC=&quot;gogh.rest-work.jpg&quot;</td>
<td><a href="http://www.cs.technion.ac.il/~konop/publication.html">http://www.cs.technion.ac.il/~konop/publication.html</a></td>
</tr>
</tbody>
</table>

Remarks

- Data mining in the Web is an area of growing importance
  - In particular, the emerging of EC
  - More and more applications will benefit from the knowledge from data mining
- Web mining = Web data collection + traditional mining?
- Keep informed with new Web technology
  - Servelet, RMI and XML