Introduction

- Continuous growth in the size and the use of WWW
- Large and complicated web structures
- Huge financial interests in attracting customers, leading them to relevant content

Objectives:
- Increase # of visits on a server to increase ads expenditure
- Convincing users to buy products and services

Need for predicting user needs in order to improve the usability and user retention of a web site

Web Mining for Personalization
Tutorial Outline

- The Web personalization process
- Web mining techniques
  - Log structure and preprocessing
  - Web log data analysis
- Personalization systems and approaches
  - Semantic-based approaches
  - Link analysis-based approaches
- Future Directions & Challenges
Tutorial Outline

- The Web personalization process
- Web mining techniques
  - Log structure and preprocessing
  - Web log data analysis
- Personalization systems and approaches
  - Semantic-based approaches
  - Link analysis-based approaches
- Future Directions & Challenges
How?

Customization…

- Based on IP
How?

Customization...

- Based on personal preferences/characteristics
How?

Targeted ads/emails
- Based on content viewed/purchased

As someone who has shown interest in digital photography and video, you might be like to know about a great new device that lets you burn DVDs without a PC— the Sony DVDirect DVD Recorder.
How?

- Recommendations
  - Based on content & usage
  - “other people who bought/watched this item also bought/watched…”

- Also...
  - Index pages
  - Highlighting/Adding links
Web Personalization - Definition

“any action that adapts information or services provided by a Web site to the needs of a user/set of users, taking advantage of the knowledge gained from the user’s navigational behavior”

Web Personalization – Impact (I)

Revenues

- global investment in personalization technologies will grow from $500m in 2001 to $2.1bn in 2006
- North America: 67% of personalization revenues (lead in eCommerce and CRM technologies)
- Europe: 25% ($131m) of global investment

[Datamonitor]
Web Personalization – Impact (II)

Research

- More than 10 workshops on Web personalization
  - AAAI SWP, IJCAI ITWP, etc.

- Many relevant workshops on web usage mining/web information management
  - ACM WEBKDD, PKDD EWMF, ACM WIDM, etc.

- Related conferences
  - KDD, PKDD, WWW, ICDM, etc.
What we WILL discuss today

- Basic concepts and techniques for usage data preprocessing
- Basic web mining algorithms and their foundations
- Different web personalization approaches: content/collaborative filtering, recommendation algorithms, temporal/trend based issues
- Semantics & Web Personalization
- Link analysis & Web Personalization
What we will NOT discuss today

- User profiling
- Elaborate on data preprocessing
- Privacy Issues
- Personalization and E-commerce/Business Applications

Related Tutorials:

- Srivastava and Cooley (PKDD2001)
- Berendt, Mobasher, Spiliopoulou (KDD2001 & PKDD2002)
- T4 – Mining the Volatile Web (Friday)
Web Mining
The three cornerstones

WEB MINING

USAGE
- usage of web pages – stored in the web logs

CONTENT
- data presented to the end user – the web site’s content

STRUCTURE
- the way content is organized – the web site’s hyperlinks
Also…

- May combine with other sources of information:
  - e-commerce and product-oriented user events (e.g. ad or product click-throughs)
  - user profiles and/or user ratings
  - referrer logs, agent logs, and client-side cookies
Web Personalization

1. **Data Collection**
   - Web server
   - Log Files
   - Web site Structure

2. **Data Preprocessing**
   - Processed Web Data

3. **Data Analysis - Pattern Discovery**
   - Web Model

4. **Pattern Analysis**

5. **Personalization**
Tutorial Outline

- The Web personalization process
- Web mining techniques
  - Log structure and preprocessing
  - Web log data analysis
- Personalization systems and approaches
  - Semantic-based approaches
  - Link analysis-based approaches
- Future Directions & Challenges
Web Server Logs

- Common Log Format
  \textit{remotehost rfc931 authuser date “request” status bytes}

- Extended Log Format (W3C) [W3Clog]
  - Prefixes: \textit{c (client), s (server), r(remote), cs (client to server), sc, sr (server to remote server-proxies), rs}
  - Identifiers: \textit{date, time, ip, bytes, cached, status, comment, method, uri, uri-stem, uri-query}
### A Typical Web Log

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>IP Address</th>
<th>URL</th>
<th>Status Code</th>
<th>Referer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-01-28</td>
<td>14:31:33</td>
<td>66.77.73.87</td>
<td>GET /people/mhalk.html</td>
<td>404</td>
<td>FAST-WebCrawler/3.3+(<a href="mailto:crawler@fast.no">crawler@fast.no</a>;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FAST-WebCrawler/3.3+(<a href="mailto:crawler@fast.no">crawler@fast.no</a>;)</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>2002-01-28</td>
<td>14:36:26</td>
<td>192.93.2.2</td>
<td>GET /projects.html</td>
<td>200</td>
<td><a href="http://www.db-net.aueb.gr">www.db-net.aueb.gr</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686)</td>
<td></td>
<td><a href="http://www.google.com/search?hl=en&amp;q=DBGlobe+">http://www.google.com/search?hl=en&amp;q=DBGlobe+</a></td>
</tr>
<tr>
<td>2002-01-28</td>
<td>14:36:31</td>
<td>192.93.2.2</td>
<td>GET /index.php</td>
<td>200</td>
<td><a href="http://www.db-net.aueb.gr">www.db-net.aueb.gr</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686)</td>
<td></td>
<td><a href="http://www.db-net.aueb.gr/projects.html">http://www.db-net.aueb.gr/projects.html</a></td>
</tr>
<tr>
<td>2002-01-28</td>
<td>14:36:38</td>
<td>192.93.2.2</td>
<td>GET /jobs.htm</td>
<td>200</td>
<td><a href="http://www.db-net.aueb.gr">www.db-net.aueb.gr</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686)</td>
<td></td>
<td><a href="http://www.db-net.aueb.gr/index.php">http://www.db-net.aueb.gr/index.php</a></td>
</tr>
<tr>
<td>2002-01-28</td>
<td>14:36:38</td>
<td>192.93.2.2</td>
<td>GET /images/dbnetOnlyMicro.gif</td>
<td>200</td>
<td><a href="http://www.db-net.aueb.gr">www.db-net.aueb.gr</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686)</td>
<td></td>
<td><a href="http://www.db-net.aueb.gr/index.php">http://www.db-net.aueb.gr/index.php</a></td>
</tr>
</tbody>
</table>
Data Preprocessing

# Data Preprocessing

.aws log example:

<table>
<thead>
<tr>
<th>Date Time</th>
<th>IP Address</th>
<th>Method</th>
<th>URI Stem</th>
<th>Status</th>
<th>Server</th>
<th>User-Agent</th>
<th>Referer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002-01-28 14:36:26</td>
<td>192.93.2.2</td>
<td>GET</td>
<td>/projects.html</td>
<td>200</td>
<td><a href="http://www.db-net.aueb.gr">www.db-net.aueb.gr</a></td>
<td>Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686)</td>
<td><a href="http://www.google.com/search?hl=en&amp;q=DBGlobe+">http://www.google.com/search?hl=en&amp;q=DBGlobe+</a></td>
</tr>
<tr>
<td>2002-01-28 14:36:38</td>
<td>192.93.2.2</td>
<td>GET</td>
<td>/jobs.htm</td>
<td>200</td>
<td><a href="http://www.db-net.aueb.gr">www.db-net.aueb.gr</a></td>
<td>Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686)</td>
<td><a href="http://www.db-net.aueb.gr/index.php">http://www.db-net.aueb.gr/index.php</a></td>
</tr>
</tbody>
</table>
Tutorial Outline

- The Web personalization process
- Web mining techniques
  - Log structure and preprocessing
  - Web log data analysis
- Personalization systems and approaches
  - Semantic-based approaches
  - Link analysis-based approaches
- Future Directions & Challenges
Web Mining

- **Web usage mining (WUM):** The application of web log data analysis techniques to model users’ navigational behaviour.

  Use models for:
  - Decision making
  - Web personalization/recommendations

- **Integration of web mining techniques:** Combine web usage mining with content/structure mining techniques/data
Web log data
Analysis Techniques

- Statistical
- Content/Collaborative filtering
- Data mining
- Probabilistic
- Link analysis
Web log data
Analysis Techniques

- Statistical
- Content/Collaborative filtering
- Data mining
- Probabilistic
- Link analysis
Statistical Log Analysis

- Traffic Analysis:
  - Diagnostic statistics
  - Server statistics
  - Referrers statistics
  - User demographics
  - Client statistics
  - …

- Used for improving system’s performance, site modification, marketing decisions’ support
Statistical Log Analysis - Example

- Most Frequent paths traversed by users/Average length of path followed
  - Av. length of a path: 3 pages
- Entry/Exit Pages, single access pages
  - 65% of visitors left the site at page /home/misc/indifferentpage.htm
- # of successful/failed/redirected/cached hits
- Top referring sites, top keywords/phrases used
  - Top keyword used: “web mining”
- Most active countries/cities/organizations
Web log data
Analysis Techniques

- Statistical
- Content/Collaborative filtering
- Data mining
- Probabilistic
- Link analysis
Content-based Filtering

- Based on individual users’ preferences.
- Record users’ preference in specific items (e.g. purchasing/rating books/DVDs etc)
- Recommend items similar to items the user liked (purchased/rated) in the past
Collaborative Filtering

- Users rate/purchase objects
- Model ratings/purchases as vectors
  - item vector $<i_1, i_2, \ldots, i_n>$
  - e.g. user A vector (binary): $<1, 0, 0, 1, \ldots, 0>$
  - e.g. user B vector (weighted): $<0.6, 0, 0.1, \ldots, 0.9>$

- Users with similar ratings/purchases have analogous interests

- Amazon.com and other e-commerce sites use content-enhanced versions of collaborative filtering
Collaborative filtering - Example

<table>
<thead>
<tr>
<th>i1, i2, i3, i4, i5, i6</th>
<th>Cosine sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 1, 1, 0, 0, 1</td>
<td>1/3</td>
</tr>
<tr>
<td>0, 0, 1, 1, 1, 0</td>
<td>0</td>
</tr>
<tr>
<td>1, 0, 0, 0, 0, 0</td>
<td>0</td>
</tr>
<tr>
<td>1, 1, 0, 0, 0, 1</td>
<td>1/3</td>
</tr>
<tr>
<td>1, 0, 0, 0, 0, 1</td>
<td>1/4</td>
</tr>
</tbody>
</table>

Recommend items: i1, i3
Web log data
Analysis Techniques

- Statistical
- Content/Collaborative filtering
- Data mining
- Probabilistic
- Link analysis
Association Rules

- Find frequent patterns/associations/correlations among sets of items
- Find correlations between pages not directly connected
- Reveal associations between groups of users with specific interests
Association Rules - Example

- \( A \implies B \) (c, s)

\( A, B \): sets of items (usually web pages)
\( c \): confidence, \( s \): support
(apriori algorithm)

- e.g.:

  /events/ski.html, travel/ski_resorts.html \( \implies \)
  /equipment/ski_boots.html (85%, 3%)
Clustering

- Group together items with similar characteristics
  - user clusters (similar navigational behavior)
  - page clusters (groups of pages conceptually related)
Clustering - Example

Use a clustering algorithm and a similarity measure to group web documents
Sequential Pattern Discovery

- An extension of association rules mining
- Incorporate the notion of time sequence
- A web page or a set of pages accessed immediately after another set of pages

- Also: Sequence mining using Markov Models (next)
Early WUM Approaches

- Many WUM systems in past decade [YZ+96, CPY96, JF+97, SZ+97, WYB98]
  - OLAP techniques [ZXH98, HN+01]
  - Uncertainty & clustering [KJ+01]
  - Visualization [WebTool-MPT99, WebSIFT-SC+00]
  - Mining languages & sequential patterns [MiDAS-BB+99, MINT(WUM)-SFW99]

- Last 5 years: WUM for Web Personalization
Usage-based Approaches

- Adaptive Web Sites
  - Creation of Index Pages [PageGather-PE00a, IndexFinder-PE00b]

- Dynamic Recommendations
  - Aggregate Usage Profiles [WebPersonalizer-MD+00a, NP03]
  - Inductive Logic Programming [EGP02], Neural Networks [NP04]
  - Incremental Clustering of clickstream data [NC+03, SUGGEST-BS04]
Web log data
Analysis Techniques

- Statistical
- Content/Collaborative filtering
- Data mining
- Probabilistic
- Link analysis
Probabilistic Models

Model web sessions as a weighted Navigational Graph

- Nodes = pages of the web site
- Edges = links between pages
- Weights = # of transitions from “source” to “destination” page
- May have a root $R$ and repeating nodes (tree) or 2 special nodes $S$ (start) and $E$ (exit) (graph).
Navigational Graph

<table>
<thead>
<tr>
<th>User Session #</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a→b→c→d</td>
</tr>
<tr>
<td>2</td>
<td>a→b→e→d</td>
</tr>
<tr>
<td>3</td>
<td>a→c→d→f</td>
</tr>
<tr>
<td>4</td>
<td>b→c→b→g</td>
</tr>
<tr>
<td>5</td>
<td>b→c→f→a</td>
</tr>
</tbody>
</table>
Transitional Probabilities

- Create **Transition Probability Matrix** $TP$

  $TP_{i,j}$: The probability of transitioning from page(s) $i (p_i)$ to page $j (p_j)$ in one step

  \[
  TP_{i,j} = P(p_j / p_i) = \frac{w_{ij}}{\sum_{k \in Out(p_i)} w_{ik}}
  \]

- $w_{ij}$: The sum of all links from $p_i \rightarrow p_j$

- $w_i = \sum_{k \in In(p_i)} w_{ki}$

  the number of times a page was visited

- Similarly for longer “source” paths
Path Prediction

- e.g. User has visited $p_1, p_2, p_3$
- $NG: 2$ paths – $p_4, p_5$
  - $P(p_1, p_2, p_3, p_4) > P(p_1, p_2, p_3, p_5)$
  - Choose $p_4$

- BUT: $NG$ very big structure – expensive computations

- NEED a more “compact” model that:
  - Captures sequential dependence
  - Preserves statistical characteristics of (most popular) paths
  - Fits into main memory
Probabilistic Models

- Markov Chains (MC)
- Higher-order Markov Models (MM)
- Semi-Markov/Hybrid Models (HM)
- Tree-like Models (TM)
Markov Chains

- 1\textsuperscript{st}-order Markov Models
- Simple model – Captures sequential dependence
- Based on Markov Property: Next state to be visited is only a function of the current state and independent of the past

\[
P_{i,j} = P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i)
\]

- \(TP^m\): prob. of visiting a page in \(m\) steps
Markov Chains - Example
Higher-order Markov Models (I)

- Web navigation: NOT memory-less
- Relax Markov Property
- \( m \)-th-order Markov Model: Next state to be visited is a function of the past \( m-1 \) visits

\[
P_{i,j}^{(m)} = P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, ..., X_{n-m+1} = i_{n-m+1})
\]

- Use Chain Rule to make path predictions

\[
P(p_1p_2...p_k) = P(p_1) \prod_{i=2}^{k} P(p_i / p_{i-1}...p_{i-m})
\]
Higher-order Markov Models (II)

- More accurate for predicting navigational paths

- **BUT**: trade-off between *improved coverage* and *exponential increase in state-space complexity* as the order increases
  - e.g. 2\textsuperscript{nd} order MM: TP is $M^2 \times M$
MM-based Approaches

Reduce TP matrix size - Improve model accuracy

- Hypertext Probabilistic Grammar (HPG): 1st-order Markov model [BL00]
- Model-based user clustering (WebCANVAS) [CH+00]
- Node clustering using structural-statistical similarities [ZHH02]
- State cloning to represent 2nd-order transition probabilities [LL03]
Interpolated Markov Models

- Interpolate among models of various orders
  - Simple linear interpolation
    \[
P_{IMM}(X_n = i | X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \lambda_0 P(X_n = i) + \\
    \lambda_1 P(X_n = i | X_{n-1} = i_{n-1}) + \ldots + \lambda_n P(X_n = i | X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)
    \]
  - General linear interpolation (the weights depend on the history)
    \[
P_{IMM}(X_n = i | X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \lambda_0 P(X_n = i) + \\
    \lambda_1 (X_{n-1} = i_{n-1}) P(X_n = i | X_{n-1} = i_{n-1}) + \ldots + \\
    \lambda_n (X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) P(X_n = i | X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)
    \]
- Better prediction accuracy BUT require much more resources (preprocessing/training)
IMM-based Approaches

- Reduces state complexity – Improve prediction precision
- Use pruning/clustering techniques:
  - All-Kth-Markov models. Eliminate states expected to have low prediction accuracy [DK01]
  - Mixture model: Personalize global model and group users into clusters. Use of maximum entropy & 1st order MM [MPG03]
  - Relational Markov Models (1st and 2nd order). PROTEUS system [ADW02]
Tree Models

- Sequential Probabilistic Models
- Based on probabilities but model is tree-like
- Alternative to Markov models, usually offering more flexibility, scalability, easier to train/create
TM-based Approaches

Incorporate various “orders”

- Different tree-like Markov models, each one corresponding to an order [DJ02]
- Eliminate branches having non-popular pages as their root [CZ03]
- WAP (Web Access Pattern) trees. Focus on frequently changing web access patterns [ZB04]
- FS-tree (Frequent Sequences tree structure). Iterative mining [ERR04]
Web log data
Analysis Techniques

- Statistical
- Content/Collaborative filtering
- Data mining
- Probabilistic
- Link analysis
Link Analysis for Web Personalization

- The web is NOT just a collection of documents – its hyperlinks are important!
- A link from page A to page B may indicate:
  - A is related to B, or
  - A is recommending, citing or endorsing B
- Web Logs constitute sub(?) sets of the Web Graph - NG
- Potential in applying ranking algorithms to NGs
A and B are co-cited by C: are related or associated.

Strength of co-citation between A and B: number of times they are co-cited.
The Web log as a Markov Chain

- A Markov Chain that has two components:
  1) A network structure (each node/page is called a state).
  2) A transition probability of traversing a link given that the chain is in a state.

- Sum of outgoing probabilities for each state is equal to 1.

- A sequence of steps through the chain is called a random walk.
The Random Surfer

- Web Graph: Markov Chain
- Surfers randomly click on links
  - Probability of an outlink from page $A$ is $1/m$
  - $m = \#\text{outlinks from } A$
- The surfer occasionally is teleported to another web page, say $B$
  - Probability of visiting $B$ is $1/n$
  - $n = \#\text{web site(s) pages.}$
- If the surfer follows links for long enough, the PageRank of a web page is the probability that the surfer will visit that page.
PageRank (PR) - Definition

\[ PR(P) = dp + (1 - d) \left( \frac{PR(P_1)}{O(P_1)} + \frac{PR(P_2)}{O(P_2)} + \ldots + \frac{PR(P_n)}{O(P_n)} \right) \]

- \( P \): web page
- \( P_i \): web pages that link to \( P \)
- \( O(P_i) \): number of outlinks from \( P_i \)
- \( d \): dumping factor
- \( p = 1/N \): teleportation probability
- \( N \): size of web graph

Web Graph Example
Computation of PR Example

<table>
<thead>
<tr>
<th>Iteration</th>
<th>PR(A)</th>
<th>PR(B)</th>
<th>PR(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3333333333</td>
<td>0.3333333333</td>
<td>0.3333333333</td>
</tr>
<tr>
<td>1</td>
<td>0.3333333333</td>
<td>0.25</td>
<td>0.375</td>
</tr>
<tr>
<td>2</td>
<td>0.3541666667</td>
<td>0.2552083333</td>
<td>0.3828125</td>
</tr>
<tr>
<td>3</td>
<td>0.358072917</td>
<td>0.256184897</td>
<td>0.384277343</td>
</tr>
<tr>
<td>4</td>
<td>0.35880534</td>
<td>0.256368</td>
<td>0.384552003</td>
</tr>
<tr>
<td>5</td>
<td>0.358942667</td>
<td>0.256402333</td>
<td>0.3846035</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>12</td>
<td>0.35897436</td>
<td>0.256410257</td>
<td>0.384615383</td>
</tr>
</tbody>
</table>
Tutorial Outline

- The Web personalization process
- Web mining techniques
  - Log structure and preprocessing
  - Web log data analysis
- Personalization systems and approaches
  - Semantic-based approaches
  - Link analysis-based approaches
- Future Directions & Challenges
Semantic Web Personalization – Motivation (I)

- Apply web mining methods to Web log data (usage data)
- Use discovered (behavioral) patterns to personalize

www.sportal.com/events/ski.html,

www.sportal.com/travel/ski_resorts.html

www.sportal.com/equipment/ski_boots.html
Semantic Web Personalization – Motivation (II)

PROBLEMS:

- Not enough hits in the usage log
- URI content updates

NEW URI:
www.sportal.com/equipment/ski_boot_offers.html

- URIs with semantically relevant content ignored

www.sportal.com/weather/snowreport.html
Semantic Web Personalization – Motivation (III)

- What about visits to semantically similar pages?

  www.sportal.com/sports/winter_sports/ski.html,
  www.sportal.com/travel/winter/hotels.html
Semantic Web Mining (SWM)

- Fast-emerging research area
- How can **Semantic Web** and **Web Mining** benefit from each other? [BHS02]
  - Connective link: **Ontologies**
  - Apply web mining techniques to semi-automatically create ontologies for building Semantic Web sites
  - **Exploit** **semantic structures such as ontologies to improve the results of Web Mining**

- Semantic Web Personalization (SWP)
SWM/SWP Approaches

Use semantic modeling

- Preference Mining: model user preferences as strict partial orders “A is better than B”. Strong semantic expressiveness of results (usage) [HEK03]

- Probabilistic Latent Semantic Analysis: latent factors that “explain” underlying relationships among pageviews (usage & content) [JZM04]
SWM/SWP Approaches

Use ontologies/concept hierarchies

- “Service based” concept hierarchies: Analyze search behavior of users. Also use for grouping web pages [BS00, B02]
- Ontologies for automatically characterizing usage and content profiles. [DM02]
- Semi-Markov process over concept hierarchy of web site’s topics [AG03]
SWM/SWP Approaches

Create conceptual logs

- SEAL [OB+03]
- WebML/WebRatio [ML+04]
- SEWeP [EVV03, EL+04]
Semantic Web Mining & Personalization Systems:

SEAL

SEAL Framework

SEmantic Web PortAL

Use:
- Ontology (portal’s backbone)
- RDF/RDF Schema
- Web pages & content metadata
- Semantic Log File
Application: KA2 Portal
Concept extraction

- Ontology underlying semantic web portal
- Static pages: Inherent RDF annotations
  - Use to map URLs to ontology terms
- Dynamic pages: Analyze query strings
Semantic Log File

Log File & feature vector:

<table>
<thead>
<tr>
<th>UserID</th>
<th>Time Int.</th>
<th>Person</th>
<th>Publication</th>
<th>name</th>
<th>author</th>
</tr>
</thead>
<tbody>
<tr>
<td>4711</td>
<td>t_{4711,1}</td>
<td>1</td>
<td>1</td>
<td>…</td>
<td>1</td>
</tr>
<tr>
<td>4711</td>
<td>t_{4711,2}</td>
<td>1</td>
<td>0</td>
<td>…</td>
<td>1</td>
</tr>
<tr>
<td>0815</td>
<td>t_{0815,1}</td>
<td>0</td>
<td>1</td>
<td>…</td>
<td>0</td>
</tr>
<tr>
<td>…</td>
<td>t_{anon}</td>
<td>1</td>
<td>1</td>
<td>…</td>
<td>1</td>
</tr>
</tbody>
</table>

Ontology concepts:
Semantic Web Mining

- Exploit concept hierarchy (taxonomy part):
  - Use generalizations/specializations to expand feature vector
- Generalize web usage mining queries
- Find underlying associations on mining results
Semantic Web Mining & Personalization Systems:

WebML / WebRatio framework

WebML/WebRatio framework

- WebML language: XML representation of Web pages
- Include content & structure metadata
- Create WebML conceptual logs (XML)
WebML Schema Example

```xml
01 <PAGE id="page3" name="Research Area">
02  <CONTENTUNITS>
03    <DATAUNIT id="dau84" name="Research Area"
04      entity="ent4" entity_name="Research Area">
05      <DISPLAYATTRIBUTE attribute="att51" name="Area Title"></DISPLAYATTRIBUTE>
06      <DISPLAYATTRIBUTE attribute="att57" name="Area Description"></DISPLAYATTRIBUTE>
07      <SELECTOR>
08        <SELECTORCONDITION attributes="att58" att_name="OID"
09          id="cond90" sel_name="Area Selection"
10          predicate="eq" value="Selected_Area_OID"></SELECTOR>
11    </DATAUNIT>
12  </CONTENTUNITS>
13  <INDEXUNIT id="inu99" name="Research Fields"
14    entity="ent19" entity_name="Res_Area_Field">
15    <SORTATTRIBUTE attribute="att60" name="Field Title"
16      order="ascending"></SORTATTRIBUTE>
17    <DISPLAYATTRIBUTE attribute="att60" name="Field Title"></DISPLAYATTRIBUTE>
18    <SELECTOR>
19      <SELECTORCONDITION relationship="rel7" rel_name="Area2Field"
20        id="cond40" sel_name="Fields_Selection"
21        predicate="in"></SELECTOR>
22  </INDEXUNIT>
23  .......
24 </PAGE>
```
Conceptual Log Analysis

- **MINE RULE:** constraint-based mining language
- Use only conceptual logs to get answers:
  - Most frequent crawling paths
  - Anomalies detection
  - High traffic users
  - Etc…
Semantic Web Mining & Personalization Systems:

SEWeP

[EVV03] M. Eirinaki, M. Vazirgiannis, SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, SIGKDD’03
[EL+04] M. Eirinaki, C. Lampos, S. Paulakis, M. Vazirgiannis, Web Personalization Integrating Content Semantics and Navigational Patterns, ACM WIDM 04
SEWeP System

Expand the recommendation set using Web site’s content semantics

www.sportal.com/accessories/ski_boots.html
www.sportal.com/weather/snowreport.html
www.sportal.com/accessories/ski_bootOffers.html

ski, snow, winter
SEWeP Personalization

- Utilize:
  - Web site’s usage
  - Web page semantics
  - Domain-specific Taxonomy

- Map extracted keywords to taxonomy categories (concepts)

- Create C-Logs (concept logs) & semantic document clusters
Similarity Mapping

Semantic similarity

Similarity measure

&

WORDNET

keywords

concept hierarchy
C-Logs Snapshot

#Fields: c-ip cs-uri-stem Title Categories Keywords

- 213.249.0.155 /projects.htm projects of db-net database data system project research medicine database data management systems novel projects research local teams position medical private

- 213.249.0.155 /people.htm people of db-net student research phd erasmus person professor student research phd associate erasmus people professor
Semantic Document Clustering

- Web pages characterized with taxonomy terms

  .../accessories/ski_boots.html
  .../weather/snowreport.html
  .../accessories/ski_boot_offers.html

- Discover document clusters based on similarity between taxonomy terms
- Use these clusters to expand recommendation set (with documents from the same cluster)
Category-based Recommendations

- Express user behavior using concepts
- Use *semantic* relationships between concepts as expressed by their topology in taxonomy
- Pattern matching to user’s navigational behavior no longer exact

E.g. “cinema, music” “movie, singer”
Recombination Engine

User visit:
/events/ski.html, /travel/ski_resorts.html

Original Recommendations:
 => /accessories/ski_boots.html

Semantic Recommendations:
 => /weather/snowreport.html
 => /accessories/ski_boot_offers.html

Category-based Recommendations:
sports, ski, winter, travel
 => snow, ski accessories
# Semantic Web Mining/Personalization Systems

<table>
<thead>
<tr>
<th></th>
<th>Ontology - Taxonomy</th>
<th>Concept logs</th>
<th>Mining Language</th>
<th>Content semantics</th>
<th>SWM</th>
<th>SWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEWeP</td>
<td>x</td>
<td>x</td>
<td></td>
<td>Automatic extraction</td>
<td>(x)</td>
<td>x</td>
</tr>
<tr>
<td>SEAL</td>
<td>x</td>
<td>x</td>
<td></td>
<td>Embedded</td>
<td>x</td>
<td>(x)</td>
</tr>
<tr>
<td>WebML / WebRatio</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>Embedded</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
Tutorial Outline

- The Web personalization process
- Web mining techniques
  - Log structure and preprocessing
  - Web log data analysis
- Personalization systems and approaches
  - Semantic-based approaches
  - Link analysis-based approaches
- Future Directions & Challenges
Link Analysis for Web Personalization

- The web is NOT just a collection of documents – hyperlinks are important!

- Page ranking resulting from the $NG$ should impact the personalization process
Personalized PageRank

\[ PR(P) = dp + (1 - d) \left( \frac{PR(P_1)}{O(P_1)} + \frac{PR(P_2)}{O(P_2)} + \ldots + \frac{PR(P_n)}{O(P_n)} \right) \]

- \( p \): teleportation probability
  - Uniform in PR
  - Can bias to prefer pages
- Known as *personalization* probability vector
Topic-sensitive PageRank

\[ PR(P) = dv + (1 - d) \left( \frac{PR(P_1)}{O(P_1)} + \frac{PR(P_2)}{O(P_2)} + \ldots + \frac{PR(P_n)}{O(P_n)} \right) \]

- Web search context
- Predefined personalization vectors \( v \)
- **Bias** the jump based on user query
  - \( v \) prefers the topics user is interested in.

Temporal aspects in personalization

- WWW evolves at high pace
  - Rankings must be frequently recomputed,
  - Do not always reflect current authorities

- Temporal Link-analysis techniques potential for personalization. Key features:
  - Evolution
  - Temporal/trend aspects
Temporal aspects (I)

- Concept of temporal interest defined by
  - A time window \([t1,t2]\)
- Graph \(G_{ti}(V,E)\) contains all nodes and edges that exist at some point in the interval \([t1,t2]\), that is whose timestamps fulfill:
  \[TS_{Deletion} > t1 \land TS_{Creation} < t2\]
Temporal aspects (II)

- **Freshness** $f$ measures the relevance of a timestamp $ts$ with respect to a temporal interest.

\[
f(ts) = \begin{cases} 
1 & \text{if } TS_{\text{Origin}} \leq ts \leq TS_{\text{End}} \\
\frac{1}{(TS_{\text{Origin}} - ts) + 1} & \text{if } t_1 \leq ts < TS_{\text{Origin}} \\
\frac{1}{(ts - TS_{\text{End}}) + 1} & \text{if } TS_{\text{End}} < ts \leq t_2 \\
\text{otherwise} & 
\end{cases}
\]

- **Freshness of node** $x$: $f(x) = f(TS_{\text{Lastmod}}(x))$
- **Freshness of edge** $x,y$: $f(x,y) = f(TS_{\text{Lastmod}}(x,y))$
Temporal aspects (III)

- **Activity** $a$ measures the frequency of change expressed by a set of timestamps $TS$ with respect to a temporal interest.

$$a(TS) = \begin{cases} 
  \text{if } TS \neq \emptyset : & \sum_{t1}^{t2} \{f(ts) | ts \in TS\} \\
  \text{otherwise : } & e
\end{cases}$$

- **Activity of node** $x$: $a(x) = a(TS_{Modifications}(x))$

- **Activity of edge** $x, y$: $a(x, y) = a(TS_{Modifications}(x, y))$
T-Rank - Design

- **Objective** is a ranking of nodes according to the authority with *respect to the temporal interest*

- **Modified PageRank** on graph $G_{ti}(V,E)$
  - Transition probabilities $t(x,y)$ depend on
    - Freshness of the nodes, edges
    - Freshness of the incoming edges
  - Random jump probabilities $s(y)$ depend on
    - Freshness/Activity of nodes
    - Freshness/Activity of incoming edges

Usage-based PageRank (UPR) - Definition

\[
UPR^n(p_i) = d \frac{w_i}{\sum_{p_j \in WS} w_j} + (1 - d) \sum_{p_j \in In(p_i)} \left( UPR^{n-1}(p_j) \times \frac{w_{j \to i}}{\sum_{p_k \in Out(p_j)} w_{j \to k}} \right)
\]

- Web personalization context
- Personalization vector based on usage data
- \(w_i\): #visits to page \(p_i\)
- \(w_{j \to i}\): #visits from \(p_j\) to \(p_i\)
UPR for Web Path Prediction

- Compute UPR for Web site’s pages
  - Bias to prefer pages visited more often
- Hybrid ranking (usage + structure)
- Use as *prior probabilities* to create ranked recommendation set
  - e.g. Using Markov models (chain rule):

\[
P(p_1p_2...p_k) = \text{UPR}(p_1) \times \prod_{i=2}^{k} P(p_i / p_{i-1}...p_{i-m})
\]

Localized UPR (l-UPR)

- **UPR** provides “global” ranking of Web site’s pages
- **l-UPR**: Bias UPR focusing on current visitor’s path
  - Use a Navigational Graph (NG) synopsis (e.g. MM)
  - Create subgraph (prNG) by expanding current visit
  - Apply UPR on subgraph
  - *Local* personalized ranking
  - Recommend pages with high *l-UPR*

I-UPR - Example

prNG (1st order Markov Model, depth = 2) for path a → b

prNG (1st order Markov Model, depth = 2) for path b → c
Tutorial Outline

- The Web personalization process
- Web mining techniques
  - Log structure and preprocessing
  - Web log data analysis
- Personalization systems and approaches
  - Semantic-based approaches
  - Link analysis-based approaches
- Future Directions & Challenge
Future Directions & Challenges

- Web grows at a tremendous pace
  - 25% new links and 8% new pages per week [NCO04]
  - 15% of pages weekly updated [FMNW03]
- Web personalization: area with significant industrial and research impact
Future Directions & Challenges

- Integration of web and usage graphs
  - Higher quality recommendations
  - More realistic web page rankings!

- Context for personalization (space/time/device issues to modify personalized content)
  - Personalization on wireless devices/interactive tv [Datamonitor]

- Integration of content semantics & structural knowledge in the personalization process
Thank you!
References

References

[CH+00] I.Cadez, D.Heckerman, C.Meek, P. Smyth, S. White, Visualization of Navigation Patterns on a Web Site Using Model Based Clustering, in Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston MA, 2000, pp. 280-284


References

- [EVK05] M. Eirinaki, M. Vazirgiannis, D. Kapogiannis, Web Path Recommendations Based on Page Ranking and Markov Models, in Proceedings of the 7th International Workshop on Web Information and Data Management (ACM WIDM 05), November 2005
- [EVV03] M. Eirinaki, M. Vazirgiannis, SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, in Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD’03), Washington DC, August 2003
- [HN+01] Zhexue Huang, Joe Ng, David W. Cheung, Michael K. Ng, Wai-Ki Ching, A Cube Model for Web Access Sessions and Cluster Analysis, in Proceedings of the Mining Log Data Across All Customer TouchPoints Workshop (WEBKDD’01), August 2001
References

References


[WCA] Web characterization terminology & definitions. Available at http://www.w3.org/1999/05/WCA-terms/


[ZXH98] Osmar R. Zaiane, Man Xin, Jiawei Han, *Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs*, in Proceedings of Advances in Digital Libraries Conference (ADL'98), Santa Barbara, CA, April, 1998