Building Community Web Directories With Probabilistic Latent Semantic Analysis

Dimitrios Pierrakos\textsuperscript{1,2}, Georgios Paliouras\textsuperscript{1}

\textsuperscript{1}Institute of Informatics & Telecommunications, NCSR “Demokritos”, Greece
\textsuperscript{2}Department of Informatics & Telecommunications, University of Athens, Greece

Outline

- WWW “hassles” - Proposed Solutions
- Community Web Directories
- Experimental Results
- Conclusions & Future work
WWW “hassles”

- Information Overload

- The abundance problem:
  99% of the online information is of no interest to 99% of people

- Information retrieval only through limited query interfaces to various search engines

Proposed Solutions

- Web Directories

- Web Personalization

Hello, Dimitris Pierrakos. We think you will like these items.
Web Directories

- E.g. Yahoo!, Open Directory Project (ODP)
- Organize the Web Content into thematic directories
- Allows Web users to locate information that relates to their interests, through a hierarchical navigation process
- Used as a starting page for navigating the Web
- Problems:
  - Manually constructed, hence limited topic coverage
  - Difficult to navigate, due to their size and complexity

Web Personalization

- Adaptability of Web-based information systems to the needs and interests of individuals or groups of users
- A personalized Web site recognizes its users, collects information about their preferences and adapts its services, in order to match the users’ needs
- Problems:
  - Acquisition of accurate and operational models for the users
  - Models evolve across time
User Modeling

The technology that supports Web Personalization

Web Usage Mining for Web Personalization

- User Modeling:
  Create models, that can be used to adapt the system to the user's requirements

- Web Usage Mining:
  Data Mining for Web User Modeling

- Mining User Communities:
  Web Usage Mining for modeling groups of users

- Construction of User Communities can facilitate Web Personalization
Outline

- WWW “hassles” - Proposed Solutions
  - Community Web Directories
  - Experimental Results
  - Conclusions & Future work

Community Web Directories

- Combination of Web Directories and Web Personalization
- Method:
  - Analyzing usage data collected by the proxy servers of an Internet Service Provider (ISP)
  - Construction of user community models with the aid of Web Usage Mining
  - Construction of usable Web directories that correspond to the interests of user communities
ISP Usage Data

- Large volumes
- Semantic diversity
- Record the navigational behavior of the user throughout the Web, rather than within a particular Web site

Data Collection & Preprocessing

- Data Cleaning from noise (e.g. images, HTTP error codes, etc.)
- Identify access sessions: Sequences of page accesses under a specific time interval restriction
- How?
  - Grouping the logs by date and IP address
  - Selecting a time-frame which two records from the same IP address can be considered to belong in the same access session
  - Grouping the pages accessed by the same IP within the time-frame to form a session
Initial Web directory

- Document clustering based on terms frequencies in Web pages
- Use a hierarchical agglomerative approach to create Web Directories
- Nodes represent clusters of Web pages that form thematic categories
- Map the Web pages to the categories
- Use categories as characteristics of user behavior
- Results:
  - Reduction of the dimensionality of the problem
  - Semantic description of user behavior

Initial Web directory: Example
Our earlier work used clustering to construct Community Web Directories.

- Clustering: Based on “observable” behavior of users
- Latent Factor model: A number of latent factors “rule” user behavior
Clustering vs. Latent Factor Model

- Factor $z$ - Jobs
- Factor $z'$ - B2B

Computer/Companies

Probabilistic Latent Semantic Analysis (PLSA)

- Method for discovering Latent Factors in co-occurrence data
- Associates an un-observed class variable

$\tilde{z} \in Z = \{\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_{k}\}$

with each observation in data
PLSA - Community Web Directories

- User sessions $U = \{u_1, u_2, \ldots, u_i\}$
- Web directory categories $C = \{c_1, c_2, \ldots, c_j\}$
- Each pair $(u_i, c_j)$ is associated with the existence of a latent factor $z_{ik}$

Probabilistic Model:
- $P(u_i)$: a priori probability of a user session $u_i$
- $P(z_k | u_i)$: the conditional probability of latent factor $z_k$ given user session $u_i$
- $P(c_j | z_k)$: the conditional probability of accessing category $c_j$ given the latent factor $z_k$
- Through Bayes Rule calculations:

$$P(u_i, c_j) = P(z_k) \sum_k P(u_i | z_k) P(c_j | z_k)$$

- The above probabilities are the unknown parameters of the model and can be calculated using the Expectation-Maximization Algorithm.
Community Web Directories

- Discover important categories for the $k$ factors
- Latent Factor Assignment Probability (LFAP) Threshold: $P(c_i | z_k) \geq LFAP$
- Build the Community Web Directory for each factor $z_k$
Outline

- WWW “hassles” - Proposed Solutions
- Web Community Directories
  - Experimental Results
- Conclusions & Future work

Experimental Results

- ISP Proxy Log Web Usage Data
- 781,069 records
- Agglomerative Clustering
- Map pages ➔ Categories
- 60’ time threshold for access sessions
- Results:
  - 2,253 access sessions
  - 998 distinct categories
Experimental Results

- 10-fold Cross-Validation:
  - Train the model 10 times, each time leaving out one of the subsets from training
  - Use the omitted subset to evaluate the model.
  - Results are the average of 10 runs for each experiment.
- Assign sessions to Web Directories using $P(u_i|z_i)$
- Build the final Community Web Directory by selecting and joining the three most prevalent Web Directories for each user session
- Result: 
  - Session-specific Community Web Directory
- Use “target” categories:
  - Hide each category in each user session and see if the user can get to it using the session-specific Community Web Directory
Model Evaluation

- How well the model can predict what the user is looking for: Model Coverage
- What the user gains by using a community directory:

\[ \text{ClickPath} = \sum_{i=1}^{\text{pathdepth}} i^* \text{(branchfactor)} \]

Experimental Results

#Factors: 20

Coverage & User Gain
Experimental Results

<table>
<thead>
<tr>
<th>#5 Factors</th>
<th>#10 Factors</th>
<th>#15 Factors</th>
<th>#20 Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>0.63</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>User Gain</td>
<td>0.47</td>
<td>0.50</td>
<td>0.55</td>
</tr>
</tbody>
</table>

LFAP = 0.1

Outline

- WWW “hassles” - Proposed Solutions
- Web Community Directories
- Experimental Results
- Conclusions & Future work
Conclusions

- Community Web Directories is a new approach to Web personalization.
- Community Web Directories can be employed by Internet Service Providers, Web Portals, etc.
- Latent factor modeling identifies latent user characteristics.
- It derives high-quality community directories, providing significant gain to users.

Future Work

- Different methods of constructing the initial thematic hierarchy could be examined or an already available thematic category can be employed.
- Additional evaluation is required, in order to test the robustness of the algorithm to a changing environment and the usability of the resulting community directories.
Building Community Web Directories With Probabilistic Latent Semantic Analysis

Dimitrios Pierrakos\textsuperscript{1,2}, Georgios Paliouras\textsuperscript{1}

\textsuperscript{1}Institute of Informatics & Telecomms, NCSR “Demokritos”, Greece
\textsuperscript{2}Department of Informatics & Telecommunications, University of Athens, Greece