User Modeling Combining Access Logs, Page Content and Semantics

Blaž Fortuna
Dunja Mladenić
Marko Grobelnik

Artificial Intelligence Laboratory
Jožef Stefan Institute
Outline

- System Overview
- Data sources
- Defining segments
- User modeling
- Experiments
- Conclusions
System Overview

User Data

Access Logs

Articles

Index

User Modeling

Segment Profile

Segment Definition

Advertisers

Analysts
Access Logs

- User interactions with the website
- Each page-view described with:
  - User ID
  - Date and Time
  - Location (from IP address)
  - Requested page
  - Referring page
  - Search query (from Referring page)
  - Browser, Operating System, Device (from User agent)
- Users tracked using cookies
  - Tag with unique ID at the first visit
**User ID cookie:** 1234567890

**IP:** 123.123.123.123 *(Beijing, China)*

**Requested URL:**

**Referring URL:**

**Date and time:** 2009-08-25 08:12:34

**User agent:** Mozilla/5.0 (Windows; U; Windows NT 5.1; en)
AppleWebKit/526.9 (KHTML, like Gecko) Version/4.0dpi Safari/526.8
(Safari, Windows, PC)
Articles

- Content and Semantics about requested pages

- Each page described with:
  - Content
  - Annotations
    - Named Entities (e.g. Obama, Mount Rushmore, Afghanistan, Vietnam)
    - Topics (e.g. politics, opinion, sports)
  - Content meta-data (e.g. author, publish date, editorial desk)
  - Page meta-data (e.g. article, home-page, section-front)
User Data

- Provided only for registered users
  - ~20% unique users in our case
  - Can generalize to all using machine learning

- Each registered users described with:
  - Gender
  - Year of birth
  - Household income

- Noisy
User Segment

- User segment:
  
  Subset of website visitors sharing some common characteristics

- Example:
  
  - [Gender = Male]
  - [Age ≥ 40]
  - [Referring domain = facebook.com]
  - [Requested page topic = Travel]
  - ...

...
Defining Segments

- Must be simple enough so it can be used by domain experts

- Our solution
  - Index all users using inverted index
  - Segment definition equals faceted search query over users
  - Ad-hoc segment definitions

Indexed fields:

- Domain
- Sub-domain
- Page URL
- Page Meta Tags
- Page Title
- Page Content
- Named Entities
- Referring Search Term
- Referring Domain
- Referring URL
- Country (from IP)
- State (from IP)
- City (from IP)
- Date
- Day of the Week
- Hour of the day
- User Agent
- Income
- Age
- Gender
Example
User Modeling

- Feature space
  - Extracted from subset of fields
  - Using vector space model
  - Vector elements for each field are normalized

- Training set
  - One visit = one vector
  - One user = a centroid of all his/her visits
  - Users from the segment form positive class
  - Sample of other users form negative class

- Classification algorithm
  - Support Vector Machine
  - Good dealing with high dimensional data
  - Linear kernel
  - Stochastic gradient descent
    - Good for sampling
Segment visualization

- Using SVM for feature selection
- Visualize a segment by displaying keywords significant for correct classification
- Useful information for the website editors

<table>
<thead>
<tr>
<th>Gender</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td>≥ $100,000</td>
</tr>
<tr>
<td>Meta Data</td>
<td>Category Style</td>
</tr>
</tbody>
</table>

Keywords:
BOOK CANCER CHILDREN CHOP DESIGNED DR EAT FAMILY FOODS HAIR HOME HOUSE KENNEDY MS RESEARCH SCHOOLS STUDENTS STUDY WOMEN
Experimental setting

- Real-world dataset from a major news publishing website
  - 5 million daily users, 1 million registered
- Tested prediction of three demographic dimensions:
  - Gender, Age, Income
- Three user groups based on the number of visits:
  - \( \geq 2 \), \( \geq 10 \), \( \geq 50 \)
- Evaluation:
  - Break Even Point (BEP)
  - 10-fold cross validation

<table>
<thead>
<tr>
<th>Category</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>250,000</td>
</tr>
<tr>
<td>Female</td>
<td>250,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>21-30</td>
<td>100,000</td>
</tr>
<tr>
<td>31-40</td>
<td>100,000</td>
</tr>
<tr>
<td>41-50</td>
<td>100,000</td>
</tr>
<tr>
<td>51-60</td>
<td>100,000</td>
</tr>
<tr>
<td>61-80</td>
<td>100,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-24k</td>
<td>50,000</td>
</tr>
<tr>
<td>25k-49k</td>
<td>50,000</td>
</tr>
<tr>
<td>50k-74k</td>
<td>50,000</td>
</tr>
<tr>
<td>75k-99k</td>
<td>50,000</td>
</tr>
<tr>
<td>100k-149k</td>
<td>50,000</td>
</tr>
<tr>
<td>150k-254k</td>
<td>50,000</td>
</tr>
</tbody>
</table>
Combining Features

- **Context** – features that can be obtained from access logs, such as time, referring page, location and device.

- **Content features:**
  - **Text Features** – keywords extracted from the articles
  - **Named Entities** – automatically extracted named entities
  - **All Metadata** – assigned to the article by the authors and editors
    - byline; topics; main keywords; people, organization and countries mentioned in the article; publish date.

- **All Content** – combination of text features, named entities and metadata features.

- **All Features** – combination of all above features.
Gender (≥10 visits)

- Male: 50.0%, 55.0%, 60.0%, 65.0%, 70.0%, 75.0%, 80.0%
- Female: 50.0%, 55.0%, 60.0%, 65.0%, 70.0%, 75.0%, 80.0%

- Context
- Text Features
- Named Entities
- All Meta Data
- All Content
- All Features

Legend:
- Green: Male
- Light Green: Female
Age (≥10 visits)

- Context: 20.00%
- Text Features: 22.00%
- Named Entities: 24.00%
- All Meta Data: 26.00%
- All Content: 28.00%
- All Features: 30.00%

Age groups: 21-30, 41-50, 61-80
Income (≥10 visits)

Text Features
Named Entities
All Meta Data

0-24  50-74  150-254
Conclusions

- Modeling user segments
  - User friendly way to define complex segments

- Combining several data sources
  - Usage logs, content and semantics

- Tomorrow (related work):
  - **SemSearch** – “Learning to Rank for Semantic Search”
    - Using Wikipedia usage data for ranking in RDF datasets
  - **LDOW** – “Automatically Annotating Text with Linked Open Data”
Thank you

- Questions?