Data Mining for Web Personalization

Patrick Dudas

Outline

• Personalization
• Data mining
  • Examples
• Web mining
• MapReduce
• Data Preprocessing
• Knowledge Discovery
• Evaluation
• Information High
Personalization

- Goal of data mining approach is for “automatic personalization”
- Automatic Personalization:
  - Content-based
  - Collaborative
  - Rule-based

Rule-based (Brief overview)

- Create decision rules
  - Implicitly/Explicitly
- Highly domain dependent
  - Rules nontransferable
- Profiles are based on user input
  - Biased
  - Static
  - Degrade over time
Content-based (Brief overview)

- Profile based on users past experiences and their interest (ratings)
  - Think Amazon, Pandora, eBay
- Vector similarities based on cosine similarity
- Bayesian classification
- Remember: Ratings = Profile = Recommendation

Collaborative (Brief overview)

- Creating groups of users based on ratings
  - Nearest neighbor approach
- Once grouped, recommendation based on the other neighbors are presented
- More users or items = more dimensions of data
  - Dynamic or real-time not applicable
Data Mining

- Data rich descriptions
- Large volumes of data
  - reliable models
- Automated data collection
- Evaluate results/make decisions
- Integration with existing data sources

Examples of Large Datasets

- [http://aws.amazon.com/datasets](http://aws.amazon.com/datasets)
- Featured data sets:
  - Illumina - Jay Flatley (CEO of Illumina) Human Genome Data
    Setcience 315(5814): 972.
    • 350 GB
  - YRI Trio Dataset
    • 700GB
  - Sloan Digital Sky Survey DR6 Subset
    • 160 GB
  - Genome, survey data, Google Books n-gram corpuses, traffic
    statistics, OpenStreetMap dataset, Wikipedia traffic…
Data Mining Web Personalization

- Recommendations based on Web objects:
  - Items
  - Pages
  - Documents
    - Navigation by links

- Web mining
  - Pros: Personalization (duh.), real-time, more enriched datasets
  - Cons: Privacy issues, building complex systems that misrepresent the individual

Extend the Data Mining Paradigm

1) Data Preparation and Transformation
2) Pattern Discovery
3) Recommendation
Data Preparation and Transformation

- Web logs
  - Date/time usage
  - Site information
  - Resource requested (image, video, etc.)
- Site files/meta-data
- The power of the cookie

Data Preparation and Transformation (cont.)

- Pageview:
  - User actions (where they clicked and the path)
  - User events (what they are trying to accomplish)
- Session:
  - Sequence of page views
MapReduce

- Google design
- Hoodop implemented
- C++, C#, Erlang, Java, Ocaml, Perl, Python, Ruby, F#, R...

Example
Usage Data Pre-Processing

Pattern Discovery

- We have data! Now what?
  - Cluster
  - Classification
  - Association Rule Discovery
  - Sequential pattern Discovery
  - Markov Models
  - Latent Variable Model
Clustering

- Partitioning
  - Split your data into groups
  - K-means
- Hierarchical
  - Divisive (top-down)
    - Start with everything, find groups
  - Agglomerative (bottom-up)
    - Start with a cluster and add additional information
- Model-based
  - Building a model for the data (best fit)

K-means

\[ d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \]
User-Based Clustering

- Start with the user profile
- Partition into k-groups of profiles
  - Based on similarity

Association Discovery

- Support
  - min(support)
- Confidence
  - min(confidence)
Evaluation (Personalization Model)

- Challenges:
  - Recommendation algorithms may require unique set of evaluation metrics
  - Personalization actions may be different
    - Domain
    - Intended application
    - Data gathered
  - Check for overfitting data
- Training set
- ROC Curve

ROC Curve

- $TPR = TP / (TP + FN)$
- $FPR = FP / (FP + TN)$

<table>
<thead>
<tr>
<th></th>
<th>&quot;Yes&quot;</th>
<th>&quot;No&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>.95</td>
<td>.05</td>
</tr>
<tr>
<td>N</td>
<td>.20</td>
<td>.80</td>
</tr>
</tbody>
</table>
Information High

- Information is addictive
- Information can be misleading
- Information ethics
- Information is power, sometimes too powerful

Personal Suggestions

- Develop a hypothesis
- Figure out what data is needed
- Make informed decisions
  - Don’t trust just your judgment
  - Experts are experts for a reason!
  - Develop a way to validate based on experience
- Then get more data if needed
Sources


Thank you!

- Questions?