Domain-specific user preference prediction based on multiple user activities

Author: YUNFEI LONG, Qin Lu, Yue Xiao, MingLei Li, Chu-Ren Huang.

Dept. of Computing, The Hong Kong Polytechnic University
Outline

1. Introduction
2. Dataset Construction
3. Our Proposed Method
4. Experiments
5. Conclusion
Introduction

- **What is User profile?** A visual display of personal data associated with a specific user.

- **Why acquiring user profiles?**
  - Personalized recommendation
  - Personalized opinion/emotion
  - Prediction of stances
Current Problems

- How to acquire user profiles from the web?
  - Explicit: *Structured profile* components in web pages.
  - Implicit: *User activities* (Posted text, Social network and Interested topics).

- The Challenges:
  - Structured information is sparsely available: *use of unstructured data*.
  - Hard to explore multiple components of user activities at the same time: *Proposed an integrated framework*.
  - Lack of user activity and user profile data: *Build dataset for benchmarking/experiments*. 
Related Works

- Previous Methods for User profile prediction:
  - Linear classification learning based: Feature + classifier (Rao and Yarowsky. 2010)
    - Features: BoW, POS, excitement, social linguistic (agreement, abbreviation, and punctuation… )
    - Classifiers: SVM, Logistic regression, Naïve Bayes, etc.
  - Need labor intensive feature engineering.
    - Features especially social linguistic related features need professional designing.
  - Hard to incorporate non-text features.
  - Mostly in user social demographic detection: Gender, Age, Race…(Rao and Yarowsky. 2010)
Our Objectives

- Build a user profile data include three parts of user activities: user posted comments, user social network and user interested topics.
- Build an integrated model to learn user preference from three part of activities.
- Solution based on two premises
  - Homophily theory: similar individuals have similar preferences.
  - Embedding theory: similar users are represented by similar vectors if they are making similar comments, having similar followers, or sharing similar interested topics.
Corpus Construction

- **Data Source:**
  - All discussing threads from March 2012 to April 2016.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>17011</td>
</tr>
<tr>
<td>Comments</td>
<td>423758</td>
</tr>
<tr>
<td>Connected users</td>
<td>76447</td>
</tr>
<tr>
<td>Topics/threads</td>
<td>38455</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Min</th>
<th>Max</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>1</td>
<td>1747</td>
<td>25</td>
</tr>
<tr>
<td>Friends</td>
<td>1</td>
<td>1500</td>
<td>69</td>
</tr>
<tr>
<td>Topics</td>
<td>1</td>
<td>742</td>
<td>17</td>
</tr>
</tbody>
</table>

Statistic of collected Hupu corpus

- Reflect the unbalanced activities problem in social media data.

User activity information
Corpus Construction

- **User preferred teams:**
  - Basketball, user can select one of CBA’s 20 team as he/she’s favorite team.
  - 17,011 users have favorite team, not uniformly distributed.
    - Popular teams like Guangdong Southern Tigers and Beijing Ducks have 4,221 and 3,159 loyalists.
    - The least popular teams like Beijing Fly Dragons and Jiangsu Kings only have 31 and 42 loyalists.
    - Reflects the unbalanced labeling problems in social media data.
  - More reliable as golden answer
Proposed model: Model framework

- **Task**: User preferred team prediction.

- **Model framework**:

  ![Diagram of the model framework]

  - Soft max classifier
  - Concatenate all three subparts
  - Deep walk
  - Doc2 vec
  - User comments
  - User friend
  - User interested topics
User representation

- **User representation by three parts of user activities:**
  \[ u_i \propto \{ G_{W_i}, G_{S_i}, G_{I_i} \} \]
  - User comments, user social network and user interested topic are represented as \( G_{W_i}, G_{S_i}, G_{I_i} \) respectively.

- **Integrated User representation:**
  - Let \( e_W, e_S, e_I \) represent Comments(words), Social Networks and interested topics

\[ e_U = e_W \oplus e_S \oplus e_I \]

- **Comments representation:**
  - Input: user generated comments.
  - Document embedding problem
  - Embedding algorithm: Document to Vector. (Doc2vec, Mikolov 2014)
User representation

- **Social network representation:**
  - Input: User, Friends in user social network.
  - User and User’s friend build a network, we can use network embedding to obtain this representation.
  - Embedding algorithm: DeepWalk (Perozzi 2014)

- **Interested topic representation:**
  - Input: Topic threads in its list
  - User’s interested topic can be treated as ‘words’, except no word order.
  - Embedding algorithm: Document to Vector. (Doc2vec, Mikolov 2014)
Performance Evaluation

- Compare to other user presentation models:
  - Task: prediction user’s favorite team, a 20 class classification task.
  - Classifier: Soft-max classifier, no parameter tuning.
  - The dimension size of all experiments are set to 300 except BOW.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-words</td>
<td>0.2472</td>
<td>0.2695</td>
<td>0.2579</td>
</tr>
<tr>
<td>Latent Dirichlet Allocation</td>
<td>0.1745</td>
<td>0.2742</td>
<td>0.2133</td>
</tr>
<tr>
<td>Average word embedding</td>
<td>0.2241</td>
<td>0.2914</td>
<td>0.2538</td>
</tr>
<tr>
<td>Document-to-vector (Doc2vec)</td>
<td>0.3503</td>
<td>0.3689</td>
<td>0.3594</td>
</tr>
<tr>
<td>Singular Value Decomposition</td>
<td>0.3709</td>
<td>0.3176</td>
<td>0.3422</td>
</tr>
<tr>
<td>Probabilistic Matrix Factorization</td>
<td>0.3776</td>
<td>0.3564</td>
<td>0.3667</td>
</tr>
<tr>
<td>Integrated User Representation model</td>
<td><strong>0.4182</strong></td>
<td><strong>0.3521</strong></td>
<td><strong>0.3822</strong></td>
</tr>
</tbody>
</table>
**Experiment: Impact of user activity components**

- Examine the effect of each user activity: Comment (C), Network (N) and Interested topics (I).

<table>
<thead>
<tr>
<th>Feature Combination</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment (C)</td>
<td>0.4014</td>
<td>0.3487</td>
<td>0.3707</td>
<td>0.0007</td>
</tr>
<tr>
<td>Network (N)</td>
<td>0.1996</td>
<td>0.2520</td>
<td>0.2227</td>
<td>&lt;=10e-15</td>
</tr>
<tr>
<td>Interested Topics (I)</td>
<td>0.1875</td>
<td>0.2516</td>
<td>0.2149</td>
<td>&lt;=10e-15</td>
</tr>
<tr>
<td>N+I</td>
<td>0.2880</td>
<td>0.2601</td>
<td>0.2738</td>
<td>&lt;=10e-15</td>
</tr>
<tr>
<td>C+I</td>
<td>0.4011</td>
<td>0.3507</td>
<td>0.3741</td>
<td>0.002</td>
</tr>
<tr>
<td>C+N</td>
<td><strong>0.4066</strong></td>
<td>0.3502</td>
<td>0.3763</td>
<td>0.032</td>
</tr>
<tr>
<td>C+N+I</td>
<td>0.4052</td>
<td><strong>0.3632</strong></td>
<td><strong>0.3815</strong></td>
<td>N/A</td>
</tr>
</tbody>
</table>

Performance of different user activities
Performance Analysis

- Both user comments, user social network and user interested topic contain user preference information.

- Social networks and topics are not sufficient for the preference prediction task. ----May be due to the data sparseness issue in this dataset.

- Precision in the all integrated version is slightly lower than using only user comments and social network data. ---When recall is improved, more noise may also be introduced to have adverse effect on precision.
Conclusion

- A user profile corpus contain multiple activities. public available
  - To be made publically available after final check.

- A novel **integrated model** to learn user profiles from multiple user activities.
  - Different embedding methods for different component
  - Best performance to all baseline methods

 Future Works:
  - Deal with data sparseness and imbalance problem.
Thank you!
Selected Reference: