THE APPLICATION OF MACHINE LEARNING METHODS FOR ANALYSIS OF TEXT FORUMS FOR CREATING LEARNING OBJECTS

Grozin V.A., Dobrenko N.V., Gusarova N. F., ITMO University, Saint Petersburg, Russia
Ning Tao, Changchun University of Science and Technology
Introduction

• Learning object (LO) is any digital resource that can be reused to support learning.
• LOs concerning high-dynamic branches of science and technologies become outdated and trivial rapidly

=> Automated updating

Great source of knowledge is forums, but:
• They have a lot of irrelevant information
• Forum language can be foreign for teacher

Also, teacher has own individual learning objectives
Goal

Investigate task of automatically identification of information potentially useful for preparing educational materials (learning objects) within text forums.
Set-up

Algorithm of marking up a forum:
1. Choose a forum and narrow topic within as a set of keywords
2. Allocate thread which names contain a topic of interest (keyword)
3. Fetch all posts from those threads
4. Mark up “Usefulness” of those posts (0..2)
## Collected dataset

<table>
<thead>
<tr>
<th>#</th>
<th>Forum</th>
<th>Language</th>
<th>Topic</th>
<th>Threads/posts</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gamedev.ru</td>
<td>Russian</td>
<td>Unity</td>
<td>10/410</td>
<td>unity</td>
</tr>
<tr>
<td>2</td>
<td>hifi-forum.de</td>
<td>German</td>
<td>Windows vs Linux</td>
<td>13/173</td>
<td>windows, linux</td>
</tr>
<tr>
<td>3</td>
<td>forum.modelsworld.ru</td>
<td>Russian</td>
<td>Ship modeling</td>
<td>3/150</td>
<td>ship, model</td>
</tr>
<tr>
<td>4</td>
<td>5500.forumactif.org</td>
<td>French</td>
<td>Ship modeling</td>
<td>3/150</td>
<td>ship, model</td>
</tr>
<tr>
<td>5</td>
<td>bbs.csdn.net</td>
<td>Chinese</td>
<td>cocos2d-x</td>
<td>11/120</td>
<td>cocos</td>
</tr>
<tr>
<td>6</td>
<td>bbs.chinaunix.net</td>
<td>Chinese</td>
<td>Linux for beginners</td>
<td>11/103</td>
<td>linux</td>
</tr>
</tbody>
</table>
Features

• Textual features
• Thread-based features
• Social features (calculated using social graph)
Models and baselines

Collected dataset is divided into training (60%) and test (40%) sets

Models:
• Linear model
• Gradient boosting model (tree assembly)

Baselines:
• Selection of the first messages of each thread
• Usage of stemming, lemmatization and extensive expert-made semantic keywords list
Evaluation

Normalized Cumulative Gain:

\[
NCG = \frac{\sum_{i=1}^{N} rel_i}{CG_{\text{max}}^N}
\]

- N is number of selected posts
- \( rel(i) \) is quality of i-th selected post
- \( CG_{\text{max}}^N \) is maximum possible such sum for specified N
Feature selection

• Features can be selected using relative influence metric (for GBM) and significance (linear regression).

• Best features were found for each forum; there were recurring “best features”:
  – Text length
  – Sentiment value (calculated using sentiment keywords)
  – Positions in thread
Conclusion

• Dataset was collected and marked up
• Textual, social and thread-based features were extracted
• Linear and non-linear models were constructed
• Quality of information retrieval using constructed models were evaluated using NCG metric and compared with two baselines
• Most important features were extracted
Thanks!
Related works

Question answering systems are hard to use:
• Multiple languages
• Typical question is “What’s new about X”

Traditional text summarization approaches are also hard to use:
• Informal language
• Topic drifting

Information retrieval approach
Social features

- If user A mentions user B in his post, there is a connection between A and B
- We construct graph using those link
- Then, we can extract graph features:
  - Betweenness
  - In degree
  - Out degree