ECML PKDD Discovery Challenge 2008
Spam Detection and Tag Recommendations in Social Bookmarking Systems

Andreas Hotho, Dominik Benz, Beate Krause, Robert Jäschke
Knowledge & Data Engineering Group, University of Kassel

Wikis, Blogs, Bookmarking Tools
Mining the Web 2.0 Workshop

Bettina Berendt - K.U. Leuven
Natalie Glance - Google
Andreas Hotho - University of Kassel
Agenda

ECML PKDD Discovery Challenge

Wikis, Blogs, Bookmarking Tools
- Mining the Web 2.0

Program
ECML PKDD Discovery Challenge 2008

- **Website:** [http://www.kde.cs.uni-kassel.de/ws/rsdc08/](http://www.kde.cs.uni-kassel.de/ws/rsdc08/)
- **Dataset:**
  - Social bookmarking data from BibSonomy
    [http://www.bibsonomy.org](http://www.bibsonomy.org)
  - Training data released on May 5th, 2008 - complete snapshot
  - Test data released on July 30th, 2008 - 1.5 months snapshots
  - 48h time to compute results on test data
- **Submissions:**
  - 150 registered mailing list users (= access to training data)
  - 18 result submissions (13 spam detection + 5 tag recommendation)
  - 13 paper submissions - 11 accepted
Tag Recommendation Task

• **Support user during tagging process**
• **Recommend tags on the posting page**
• **Goal:** learn a model which effectively predicts the keywords a user has in mind and will use when describing a web page
## Results

<table>
<thead>
<tr>
<th>Sub. ID</th>
<th>F1M</th>
<th>Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>72209</td>
<td>0.19325</td>
<td><strong>RSDC'08: Tag Recommendations using Bookmark Content</strong> by M. Tatu, M. Srikanth and T. D'Silva</td>
</tr>
<tr>
<td>89760</td>
<td>0.18674</td>
<td><strong>Tag Recommendation for Folksonomies Oriented towards Individual Users</strong> by M. Lipczak</td>
</tr>
<tr>
<td>27845</td>
<td>0.02840</td>
<td><strong>Multilabel Text Classification for Automated Tag Suggestion</strong> by I. Katakis, G. Tsoumakas and I. Vlahavas</td>
</tr>
<tr>
<td>27876</td>
<td>0.02203</td>
<td></td>
</tr>
<tr>
<td>68481</td>
<td>0.01406</td>
<td></td>
</tr>
</tbody>
</table>
Tag Recommendation Task

- all train tags known
- 1st (f1m: 0.19325)
- 2nd (f1m: 0.18674)
- 3rd (f1m: 0.02840)
- 4th (f1m: 0.02203)
- 5th (f1m: 0.01406)
- most popular tags

Number of recommended tags vs. F1-Measure chart.
Tag Recommendation Task

)! (42406 posts)

all train tags known
1st
2nd
3rd
4th
5th
most popular tags

Recall

Number of recommended tags

UR (1245 posts)

all train tags known
1st
2nd
3rd
4th
5th
most popular tags

Recall

Number of recommended tags

U_ (14925 posts)

all train tags known
1st
2nd
3rd
4th
5th
most popular tags

Recall

Number of recommended tags

_R (966 posts)

all train tags known
1st
2nd
3rd
4th
5th
most popular tags by user

Recall

Number of recommended tags

most popular tags by resource

Recall

Number of recommended tags
Fighting Spam

http://www.flickr.com/photos/gov/442222657/
Spam Detection Task

- Growing popularity attracts spam
- Two goals:
  - Attract people
  - Increase PageRank
- Counter measures (e.g., Captchas) are not sufficient
- 25,000 manually labeled spammers in training data (vs. 2,000 non-spammers)
- Goal: learn a model which predicts whether a user is a spammer or not
Spam Detection Task

Results

<table>
<thead>
<tr>
<th>Sub. ID</th>
<th>AUC</th>
<th>Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>39014</td>
<td>0.97961</td>
<td>A novel supervised learning algorithm and its use for Spam Detection in Social Bookmarking Systems by A. Gkanogiannis and T. Kalamboukis</td>
</tr>
<tr>
<td>83234</td>
<td>0.97032</td>
<td>Rank for spam detection - ECML Discovery Challenge by P. Gramme and J.-F. Chevalier</td>
</tr>
<tr>
<td>15076</td>
<td>0.93899</td>
<td>Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking by C. Kim and K.-B. Hwang</td>
</tr>
<tr>
<td>97510</td>
<td>0.93640</td>
<td></td>
</tr>
<tr>
<td>44293</td>
<td>0.93259</td>
<td></td>
</tr>
<tr>
<td>55409</td>
<td>0.91365</td>
<td></td>
</tr>
<tr>
<td>69806</td>
<td>0.88366</td>
<td></td>
</tr>
<tr>
<td>75540</td>
<td>0.87847</td>
<td></td>
</tr>
<tr>
<td>28752</td>
<td>0.84684</td>
<td></td>
</tr>
<tr>
<td>21710</td>
<td>0.84684</td>
<td></td>
</tr>
<tr>
<td>85695</td>
<td>0.70553</td>
<td></td>
</tr>
<tr>
<td>70358</td>
<td>0.47069</td>
<td></td>
</tr>
<tr>
<td>56347</td>
<td>0.35898</td>
<td></td>
</tr>
</tbody>
</table>
Spam Detection Task

The diagram shows various ROC curves representing different spam detection tasks, each with a corresponding AUC score.

- 1st (AUC = 0.97961)
- 2nd (AUC = 0.97032)
- 3rd (AUC = 0.93899)
- 4th (AUC = 0.93640)
- 5th (AUC = 0.93259)
- 6th (AUC = 0.91365)
- 7th (AUC = 0.88366)
- 8th (AUC = 0.87847)
- 9th (AUC = 0.84684)
- 10th (AUC = 0.70553)
- 11th (AUC = 0.47069)
- 12th (AUC = 0.35898)
Spam Detection Task

![Graph showing ROC curves for Spam Detection Task]

1st (AUC = 0.97961)
2nd (AUC = 0.97032)
Spam Detection Task

![ROC Curve for Spam Detection Task]

- 3rd (AUC = 0.93899)
- 4th (AUC = 0.93640)
- 5th (AUC = 0.93259)
Spam Detection Task

spammers in BibSonomy
Spam Detection Task

„good“ users in BibSonomy

Map of the Internet provided by http://xkcd.com/195
Agenda

ECML PKDD Discovery Challenge

Wikis, Blogs, Bookmarking Tools
- Mining the Web 2.0

Program
The workshop focuses on research in analyzing wikis, blogs and tagging systems.

Looking for contributions which:

- apply state-of-the-art data mining and machine learning methods on Web 2.0 data,
- discuss aspects on the intersection of Web 2.0 and Knowledge Discovery,
- can identify the power of advanced data mining operating on Web 2.0 data.

The contributions address the three major topics of the workshop, tagging, wikis and blogs.
Many thanks to the PC!

- Sarabjot Singh Anand, University of Warwick, UK
- Mathias Bauer, mineway, Germany
- Janez Brank, Jožef Stefan Institute, Slovenia
- Michelangelo Ceci, University of Bari, Italy
- Ed H. Chi, PARC, USA
- Brian Davison, Lehigh University, USA
- Marco de Gemmis, University of Bari, Italy
- Miha Grcar, Jožef Stefan Institute, Slovenia
- Marko Grobelnik, Jožef Stefan Institute, Slovenia
- Pasquale Lops, University of Bari, Italy
- Ernestina Menasalvas, Universidad Politécnica de Madrid, Spain
- Dunja Mladežič, Jožef Stefan Institute, Slovenia
- Ion Muslea, SRI International, USA
- Giovanni Semeraro, University of Bari, Italy
- Ian Soboroff, National Institute of Standards and Technology, USA
- Myra Spiliopoulou, Otto-von-Guericke-Universitaet Magdeburg, Germany
- Gerd Stumme, University of Kassel, Germany
- Maarten van Someren, Universiteit van Amsterdam, The Netherlands
- Michael Wurst, University of Dortmund, Germany
ECML PKDD Discovery Challenge

Wikis, Blogs, Bookmarking Tools
- Mining the Web 2.0

Program
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 9:00 - 10:10 | A novel supervised learning algorithm and its use for Spam Detection in Social Bookmarking Systems (30 min)  
A. Gkanogiannis and T. Kalamboukis  
Rank for spam detection - ECML Discovery Challenge (15 min)  
P. Gramme and J.-F. Chevalier  
Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking (15 min)  
C. Kim and K.-B. Hwang                                                                 |
| 10:10 - 10:40| Coffee break                                                         |
**Program**

**Legend**

- **Discovery Challenge**: Spam Detection Task
- **Discovery Challenge**: Tag Recommendation Task
- Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop

**Time**

**Network Structures & Folksonomies**

**Predicting Tag Spam Examining Cooccurrences, Network Structures and URL Components** (15 min)
N. Neubauer and K. Obermayer

**Using Co-occurrence of Tags and Resources to Identify Spammers** (15 min)
R. Krestel and L. Chen

10:40 - 12:30

**Identifying Ideological Perspectives of Web Videos using Patterns Emerging from Folksonomies** (30 min)
Wei-Hao Lin and Alex Hauptmann

**Topical Structure Discovery in Folksonomies** (30 min)
Ilija Subasic and Bettina Berendt

**Wikipedia As the Premiere Source for Targeted Hypernym Discovery** (20 min)
Tomas Kliegr, Vojtech Svahek, Krishna Chandramouli, Jan Nemrava and Ebroul Izquierdo

12:30 - 14:00

Lunch
Program

Legend

Discovery Challenge: Spam Detection Task
Discovery Challenge: Tag Recommendation Task
Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop

Time

Recommendation/Prediction

RSDC'08: Tag Recommendations using Bookmark Content (30 min)
M. Tatu, M. Srikanth and T. D'Silva

14:00 - 15:30
Tag Recommendation for Folksonomies Oriented towards Individual Users (15 min)
M. Lipczak

Multilabel Text Classification for Automated Tag Suggestion (15 min)
I. Katakis, G. Tsoumakas and I. Vlahavas

15:30 - 16:00
BaggTaming - Learning from Wild and Tame Data (30 min)
Toshihiro Kamishima, Masahiro Hamasaki and Shotaro Akaho

Coffee break
**Program**

**Legend**

- **Discovery Challenge**: Spam Detection Task
- **Discovery Challenge**: Tag Recommendation Task
- **Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop**

**Time**

**Blog Analysis & Spam**

Clustering blog entries based on the hybrid document model enhanced by the extended anchor texts and co-referencing links (20 min)
Hiroshi Ishikawa, Masashi Tsuchida and Hajime Takekawa

Using Language Models for Spam Detection in Social Bookmarking (15 min)
T. Bogers and A. van den Bosch

Using Semantic Features to Detect Spamming in Social Bookmarking Systems (15 min)
A. Madkour, T. Hefni, A. Hefny and K. S. Refaat

Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems (15 min)
A. Kyriakopoulou and T. Kalambokis

Discussion

17:30 - opening of the conference