# 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*

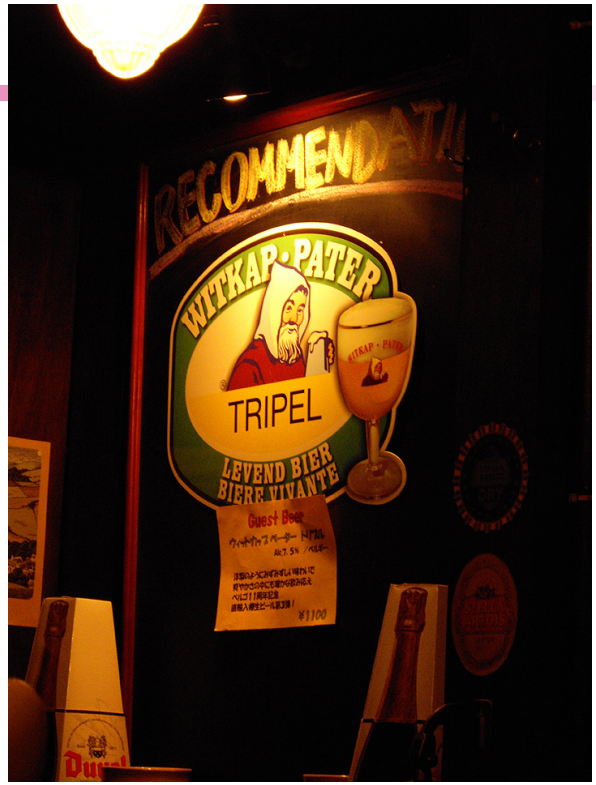
---

# 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

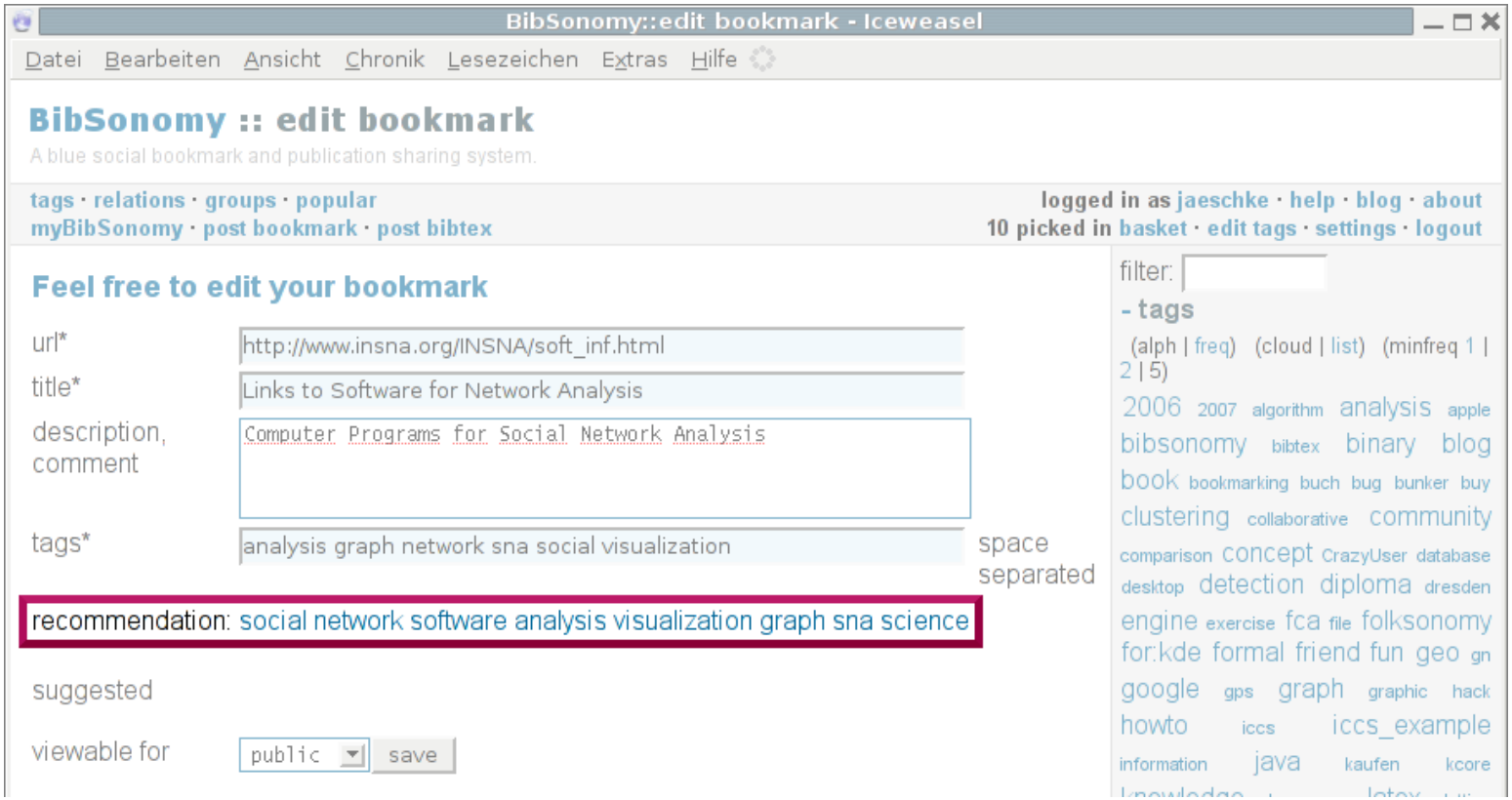
Wikis, Blogs, Bookmarking Tools
– Mining the Web 2.0

Program

# ECML PKDD Discovery Challenge 2008

- Website: http://www.kde.cs.uni-kassel.de/ws/rsdc08/
- Dataset:
  - Social bookmarking data from BibSonomy 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



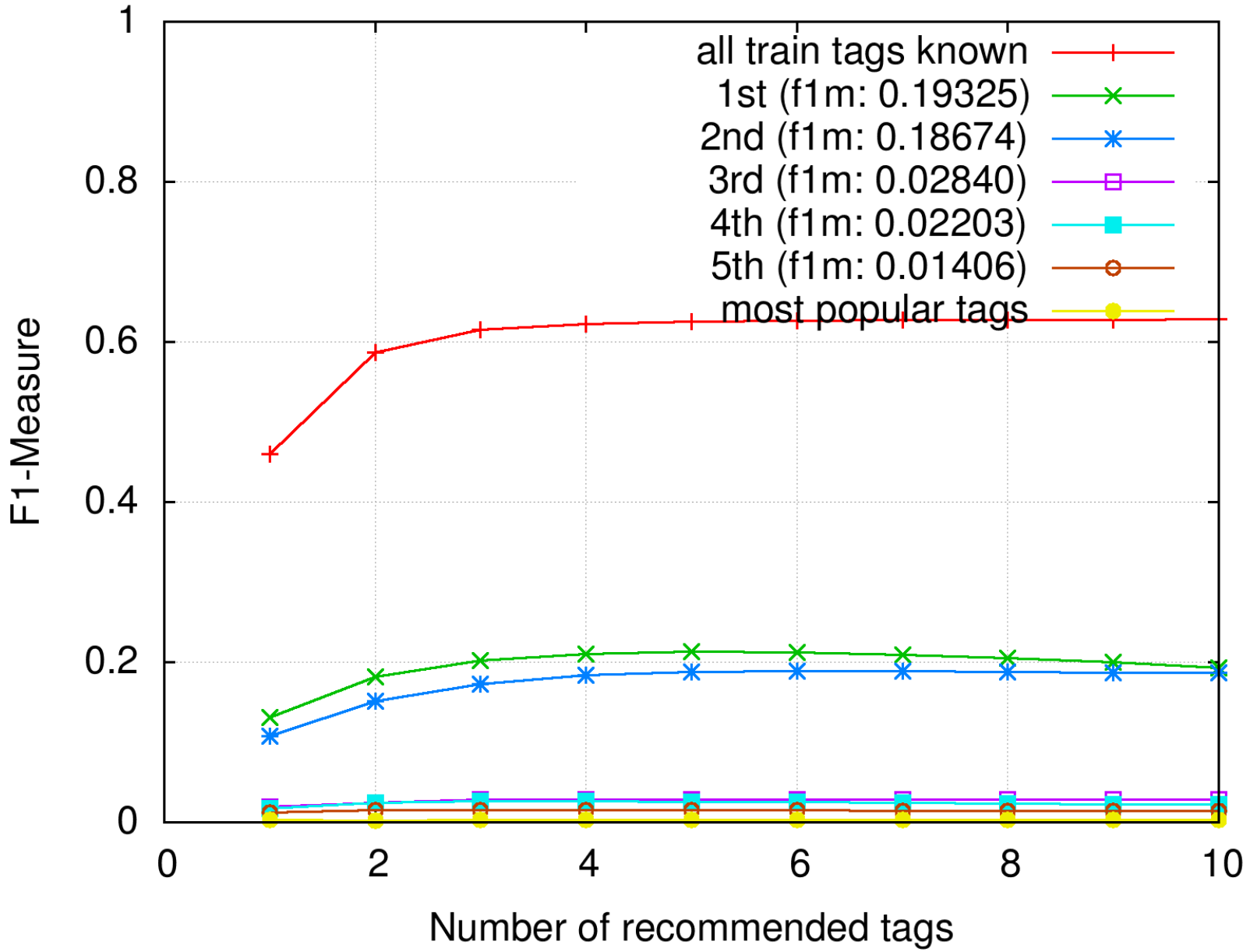

- Support user during tagging process
- Recommend tags on the posting page
- **Goal:** learn a model which effectively predicts the keywords a user has in mind and will use when describing a web page

# Tag Recommendation Task

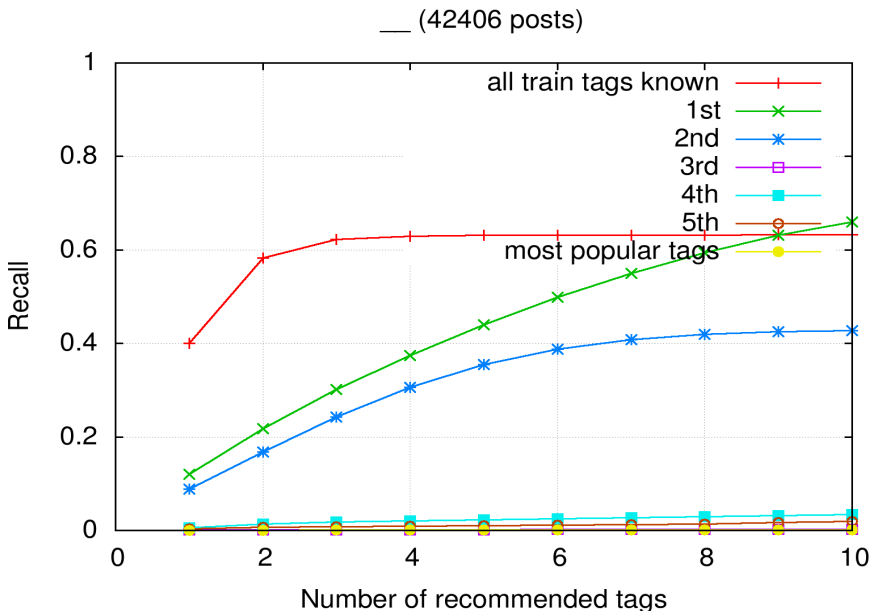
## Results

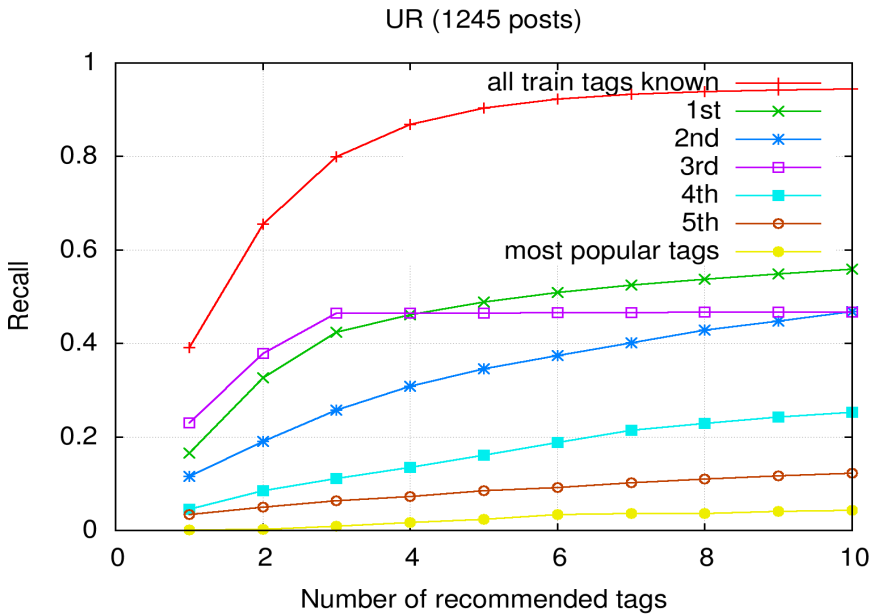
| Sub. ID | F1M | Team |
|---|---|---|
| 72209 | 0.19325 | **RSDC'08: Tag Recommendations using Bookmark Content** by M. Tatu, M. Srikanth and T. D'Silva |
| 89760 | 0.18674 | **Tag Recommendation for Folksonomies Oriented towards Individual Users** by M. Lipczak |
| 27845 | 0.02840 | **Multilabel Text Classification for Automated Tag Suggestion** by I. Katakis, G. Tsoumakas and I. Vlahavas |
| 27876 | 0.02203 | |
| 68481 | 0.01406 | |

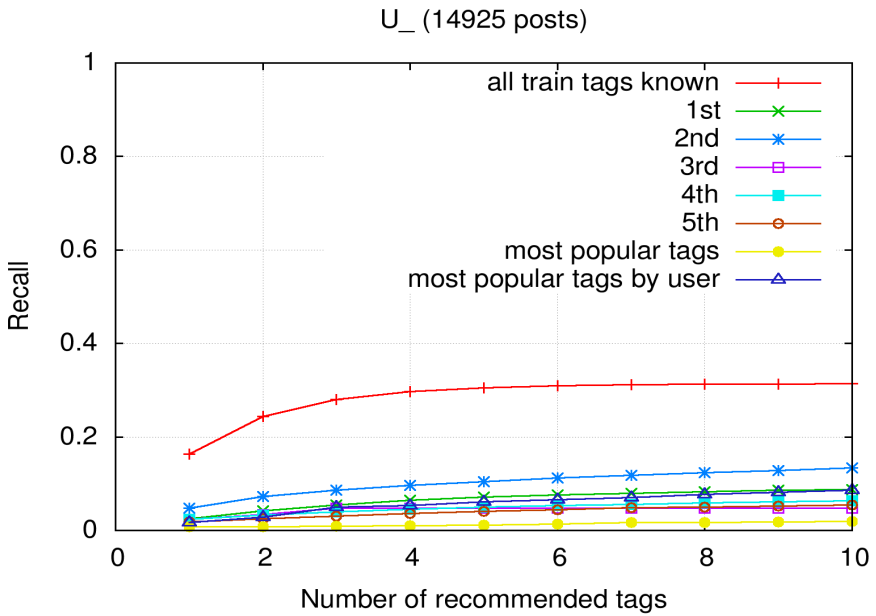
# Tag Recommendation Task

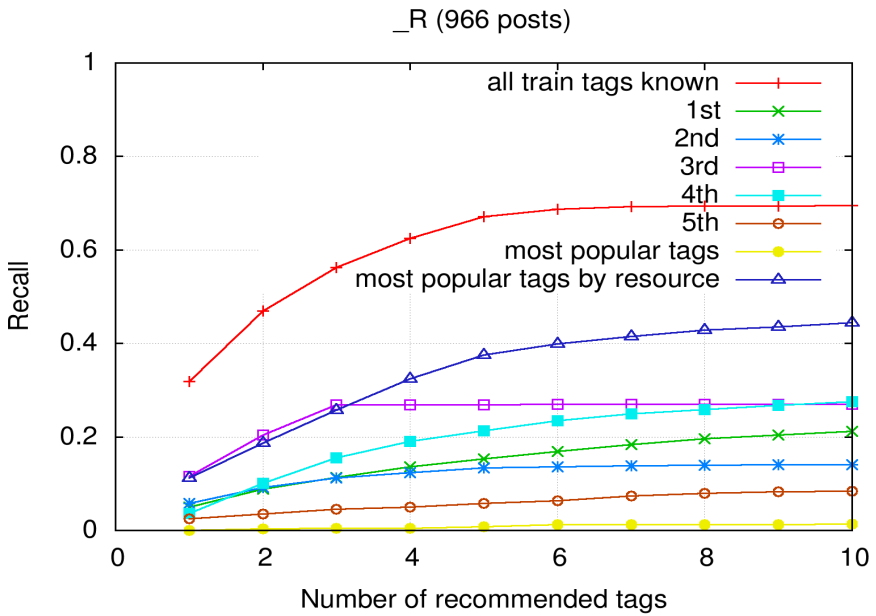


# Tag Recommendation Task



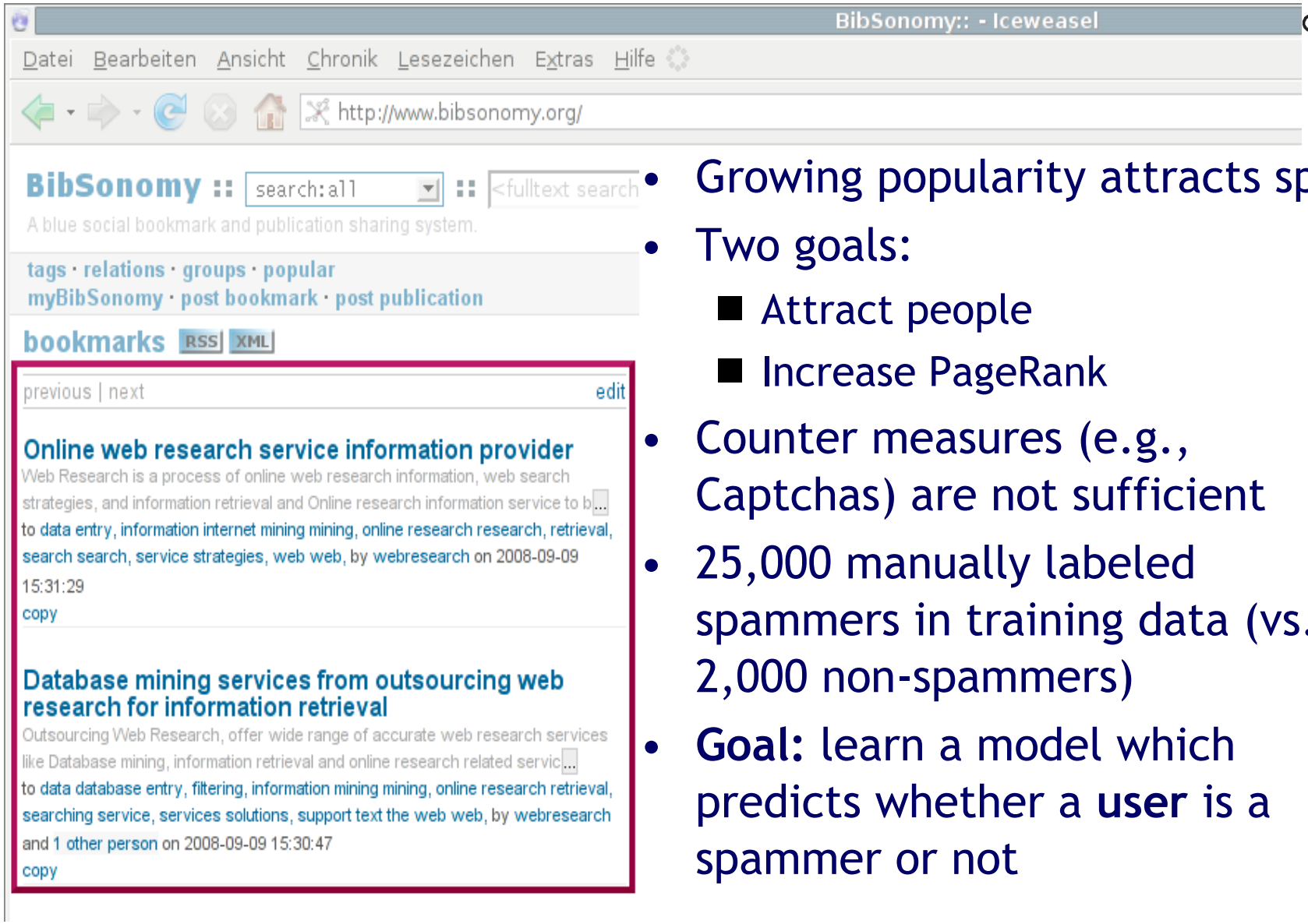



# Fighting Spam


http://www.flickr.com/photos/gov/442222

# Spam Detection Task



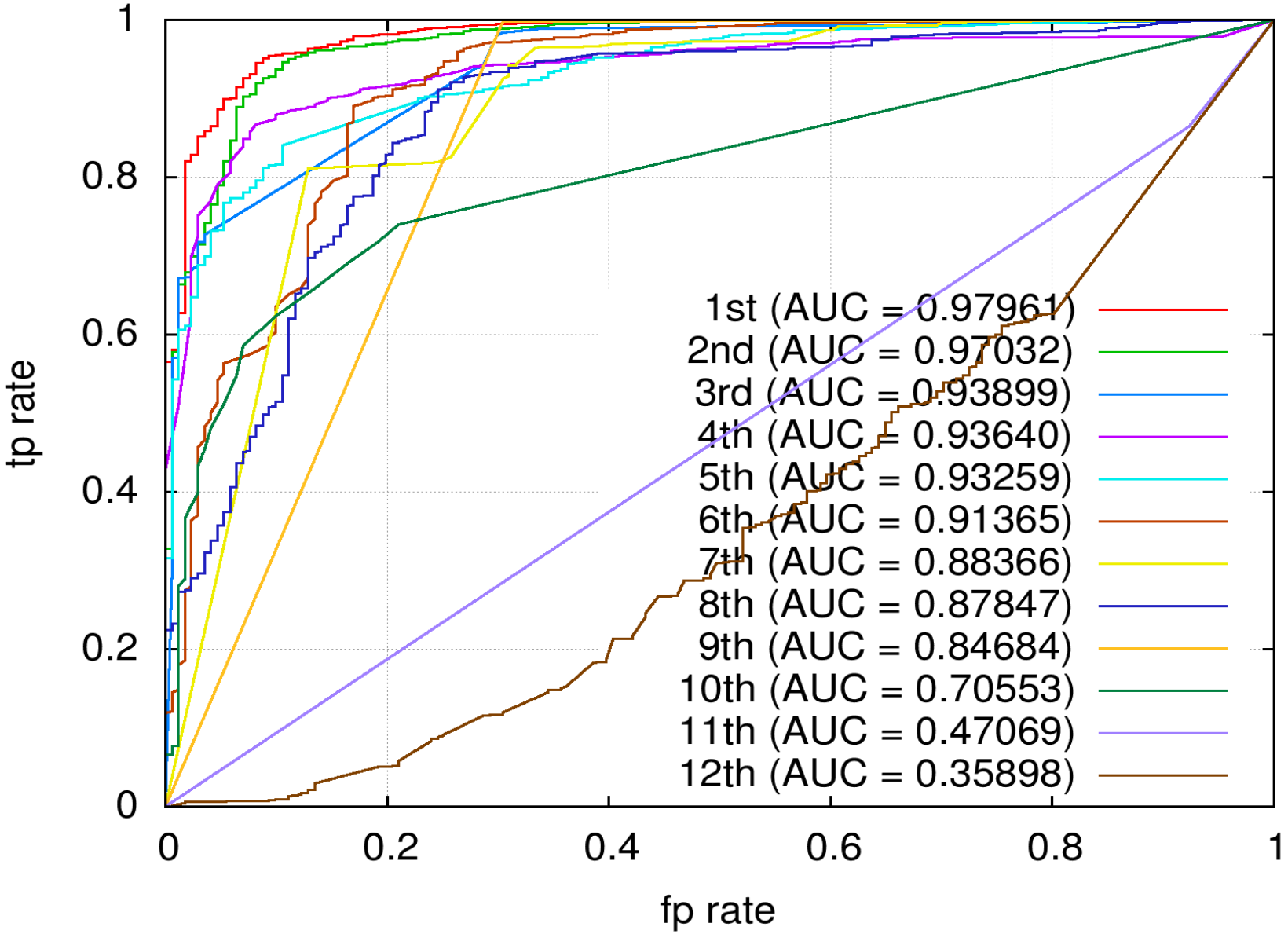

- Growing popularity attracts spam
- Two goals:
  - Attract people
  - Increase PageRank
- Counter measures (e.g., Captchas) are not sufficient
- 25,000 manually labeled spammers in training data (vs. 2,000 non-spammers)
- **Goal:** learn a model which predicts whether a **user** is a spammer or not

# Spam Detection Task

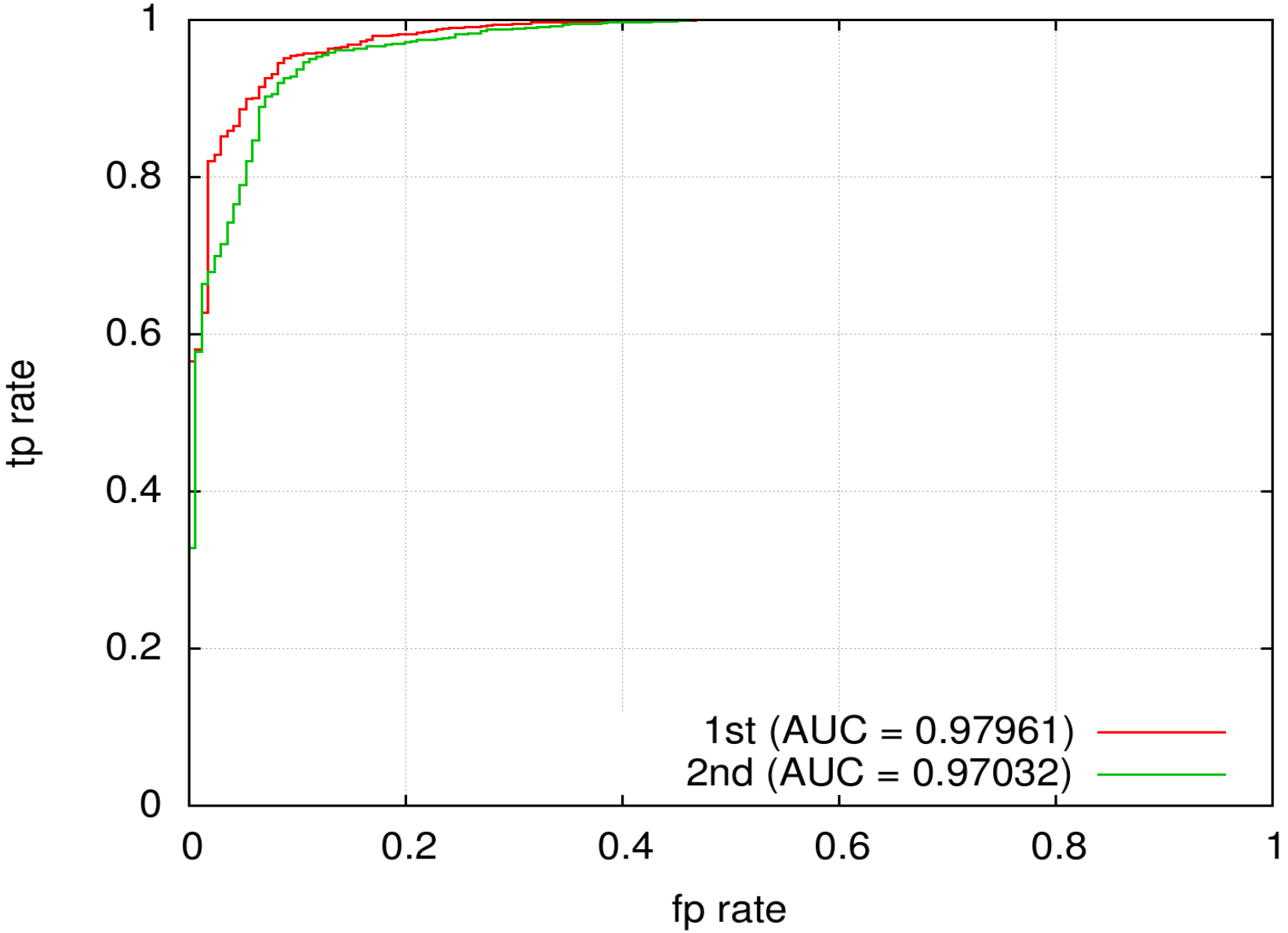
## Results

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

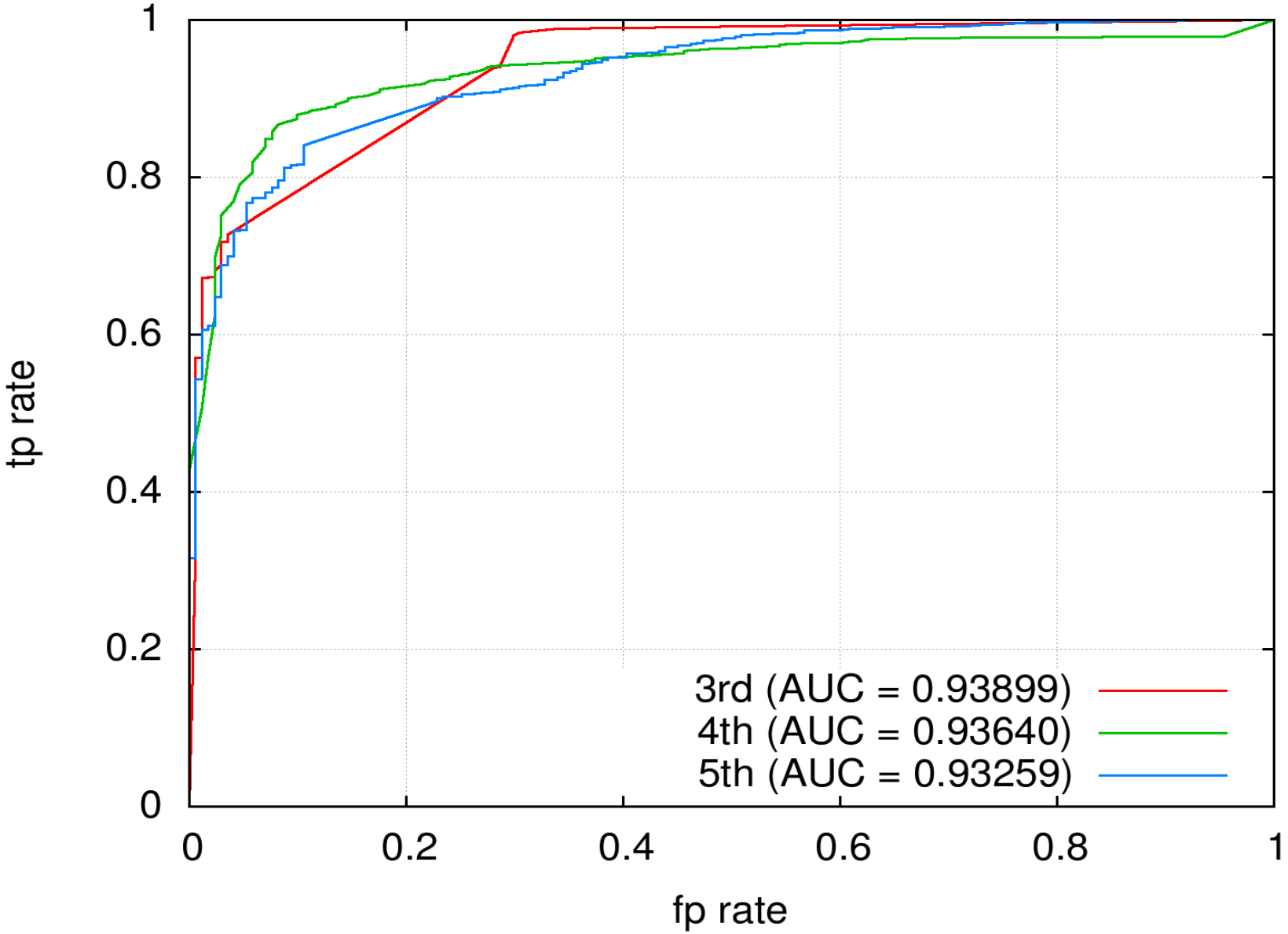
# Spam Detection Task



# Spam Detection Task



# Spam Detection Task



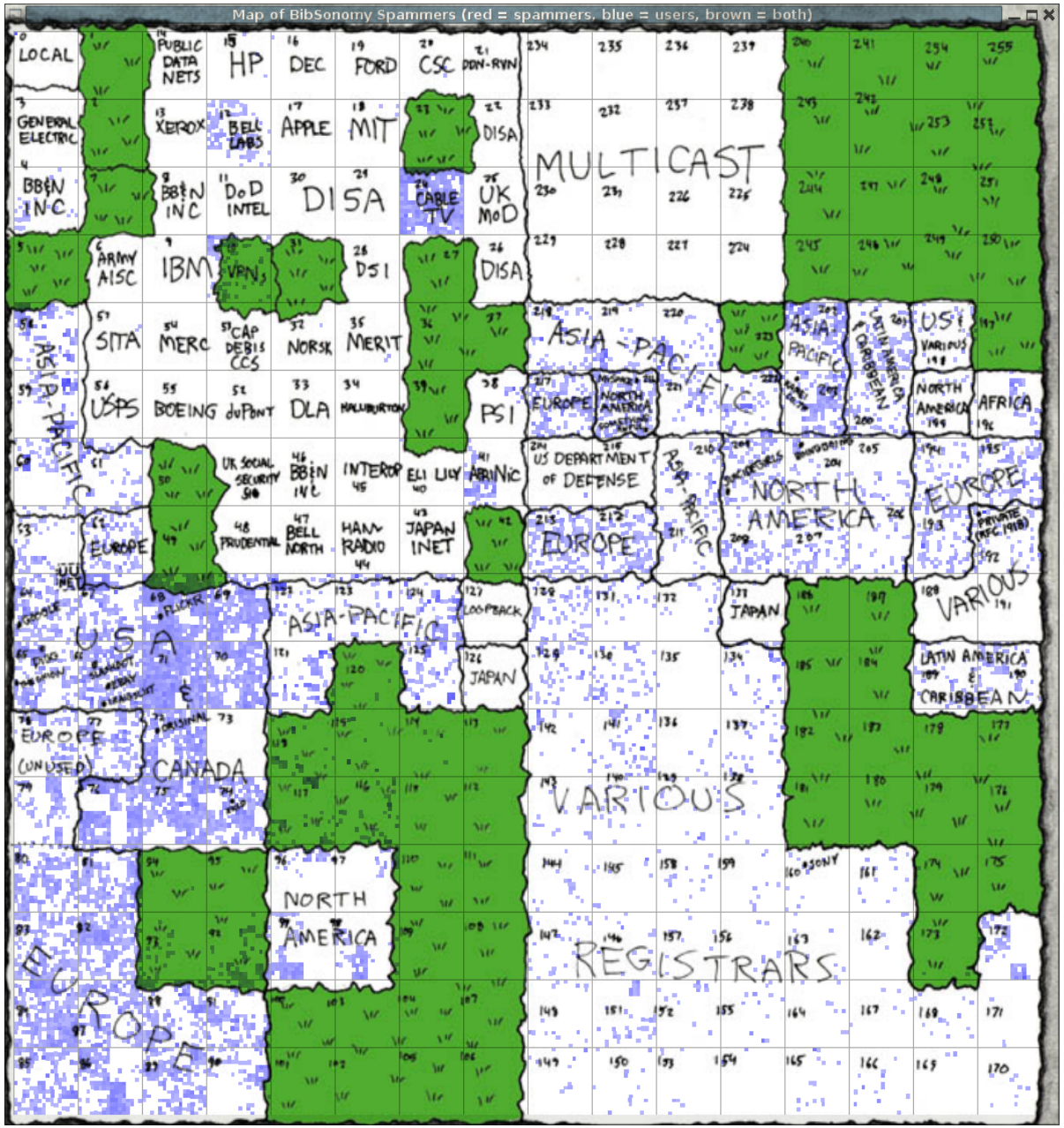
# Spam Detection Task

spammers in BibSonomy


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

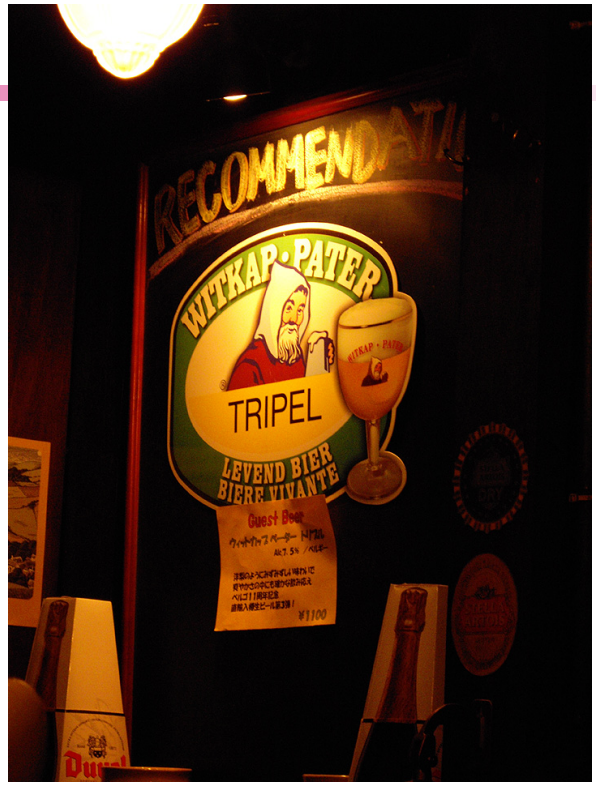
# Spam Detection Task

„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

# Wikis, Blogs, Bookmarking Tools – Mining the Web 2.0 Workshop

Website: http://www.kde.cs.uni-kassel.de/ws/wbbtmine2008

- 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!

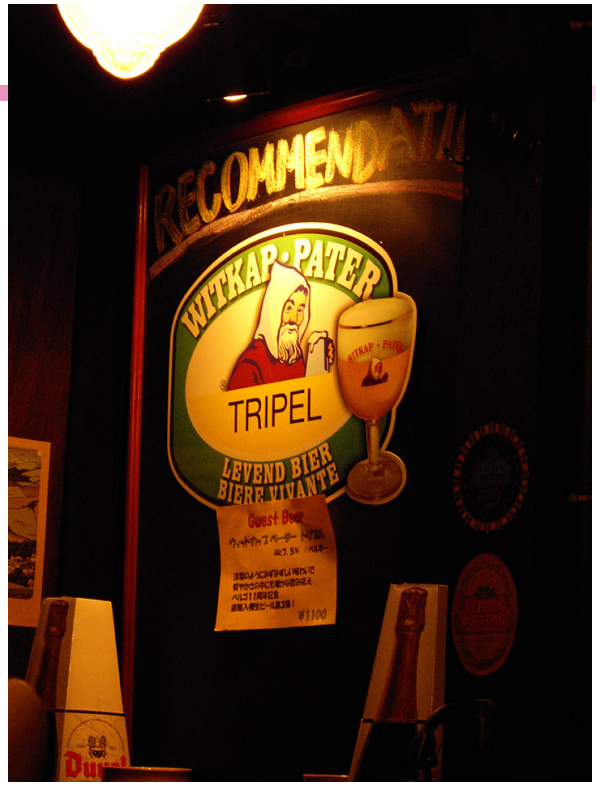
- Sarabjot Singh Anand, University of Warwick, UK
- Mathias Bauer, mineway, Germany
- Janez Brank, Jozef Stefan Institute, Slovenia
- Michelangelo Ceci, University of Bari, Italy
- Ed H. Chi, PARC, USA
- Brian Davison, Lehigh University, USA
- Marco de Gemmis, University of Bari, Italy
- Miha Grcar, Jozef Stefan Institute, Slovenia
- Marko Grobelnik, Jozef Stefan Institute, Slovenia
- Pasquale Lops, University of Bari, Italy
- Ernestina Menasalvas, Universidad Politecnica de Madrid, Spain
- Dunja Mladenic, Jozef 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

# Agenda

ECML PKDD Discovery Challenge

Wikis, Blogs, Bookmarking Tools
– Mining the Web 2.0

Program




# Program

| Legend | |
|---|---|
| | Discovery Challenge: Spam Detection Task<br>Discovery Challenge: Tag Recommendation Task<br>Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop |
| **Time** | |
| | **Spam** |
| 9:00 - 10:10 | A novel supervised learning algorithm and its use for Spam Detection in Social Bookmarking Systems (30 min)<br>A. Gkanogiannis and T. Kalamboukis<br><br>Rank for spam detection - ECML Discovery Challenge (15 min)<br>P. Gramme and J.-F. Chevalier<br><br>Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking (15 min)<br>C. Kim and K.-B. Hwang |
| 10:10 - 10:40 | Coffee break |

# Program

| Legend | |
|---|---|
| | Discovery Challenge: Spam Detection Task<br>Discovery Challenge: Tag Recommendation Task<br>Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop |
| **Time** | |
| 10:40 - 12:30 | **Network Structures & Folksonomies**<br><br>Predicting Tag Spam Examining Cooccurrences, Network Structures and URL Components (15 min)<br>N. Neubauer and K. Obermayer<br><br>Using Co-occurence of Tags and Resources to Identify Spammers (15 min)<br>R. Krestel and L. Chen<br><br>Identifying Ideological Perspectives of Web Videos using Patterns Emerging from Folksonomies (30 min)<br>Wei-Hao Lin and Alex Hauptmann<br><br>Topical Structure Discovery in Folksonomies (30 min)<br>Ilija Subasic and Bettina Berendt<br><br>Wikipedia As the Premiere Source for Targeted Hypernym Discovery (20 min)<br>Tomas Kliegr, Vojtech Svatek, Krishna Chandramouli, Jan Nemrava and Ebroul Izquierdo |
| 12:30 - 14:00 | Lunch |

# Program

| Legend | |
|---|---|
| | Discovery Challenge: Spam Detection Task<br>Discovery Challenge: Tag Recommendation Task<br>Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop |
| **Time** | |
| | **Recommendation/Prediction** |
| 14:00 - 15:30 | RSDC'08: Tag Recommendations using Bookmark Content (30 min)<br>M. Tatu, M. Srikanth and T. D'Silva<br><br>Tag Recommendation for Folksonomies Oriented towards Individual Users (15 min)<br>M. Lipczak<br><br>Multilabel Text Classification for Automated Tag Suggestion (15 min)<br>I. Katakis, G. Tsoumakas and I. Vlahavas<br><br>BaggTaming - Learning from Wild and Tame Data (30 min)<br>Toshihiro Kamishima, Masahiro Hamasaki and Shotaro Akaho |
| 15:30 - 16:00 | Coffee break |

# Program

| | |
|---|---|
| **Legend** | |
| | Discovery Challenge: Spam Detection Task<br>Discovery Challenge: Tag Recommendation Task<br>Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop |
| **Time** | |
| | **Blog Analysis & Spam** |
| 16:00 - 17:15 | Clustering blog entries based on the hybrid document model enhanced by the extended anchor texts and co-referencing links (20 min)<br>Hiroshi Ishikawa, Masashi Tsuchida and Hajime Takekawa<br><br>Using Language Models for Spam Detection in Social Bookmarking (15 min)<br>T. Bogers and A. van den Bosch<br><br>Using Semantic Features to Detect Spamming in Social Bookmarking Systems (15 min)<br>A. Madkour, T. Hefni, A. Hefny and K. S. Refaat<br><br>Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems (15 min)<br>A. Kyriakopoulou and T. Kalamboukis<br><br>Discussion |
| 17:30 - | opening of the conference |