



# 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*

---

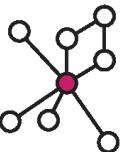
## Wikis, Blogs, Bookmarking Tools

Mining the Web 2.0 Workshop

*Bettina Berendt - K.U. Leuven*

*Natalie Glance - Google*

*Andreas Hotho - University of Kassel*



## ECML PKDD Discovery Challenge

Wikis, Blogs, Bookmarking Tools  
- Mining the Web 2.0

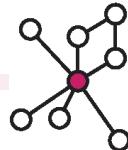
## Program





- 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



BibSonomy::edit bookmark - Iceweasel

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

**BibSonomy :: edit bookmark**

A blue social bookmark and publication sharing system.

tags · relations · groups · popular  
myBibSonomy · post bookmark · post bibtex

logged in as jaeschke · help · blog · about  
10 picked in basket · edit tags · settings · logout

**Feel free to edit your bookmark**

url\* http://www.insna.org/INSNA/soft\_inf.html

title\* Links to Software for Network Analysis

description, comment Computer Programs for Social Network Analysis

tags\* analysis graph network sna social visualization

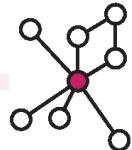
recommendation: social network software analysis visualization graph sna science

suggested

viewable for public save

filter: - tags  
(alph | freq) (cloud | list) (minfreq 1 | 2 | 5)  
2006 2007 algorithm analysis apple  
bibsonomy bibtex binary blog  
book bookmarking buch bug bunker buy  
clustering collaborative community  
comparison concept CrazyUser database  
desktop detection diploma dresden  
engine exercise fca file folksonomy  
for:kde formal friend fun geo gn  
google gps graph graphic hack  
howto iccs iccs\_example  
information java kaufen kcore  
knowledge latex ...

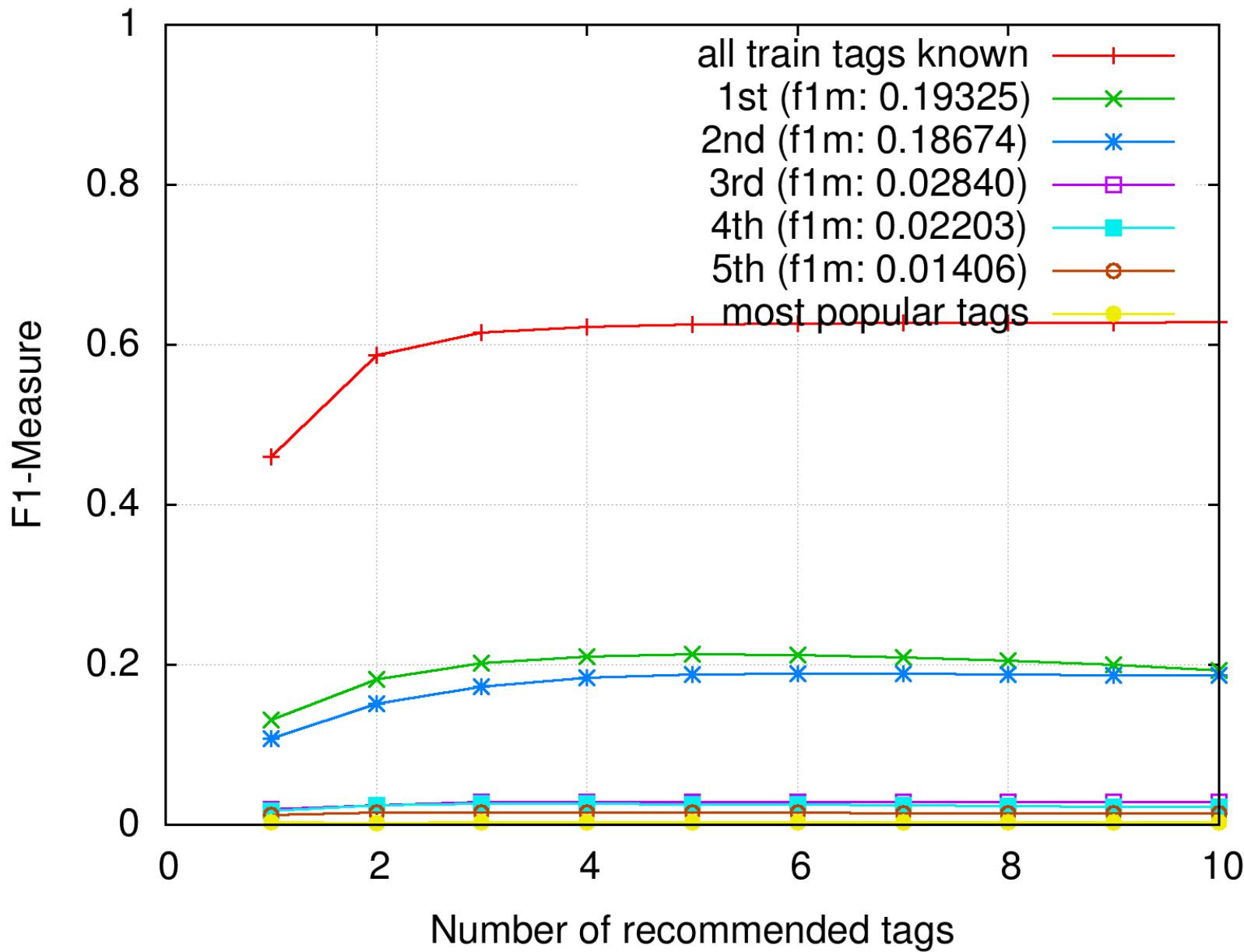
- Support user during tagging process
- Recommend tags on the posting page
- **Goal:** learn a model which effectively predicts the keywords a user has in mind and will use when describing a web page



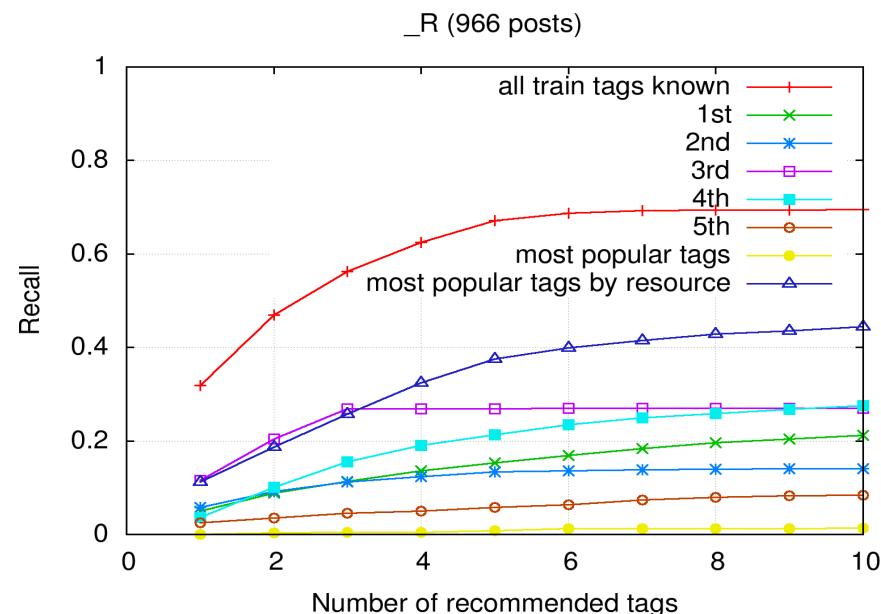
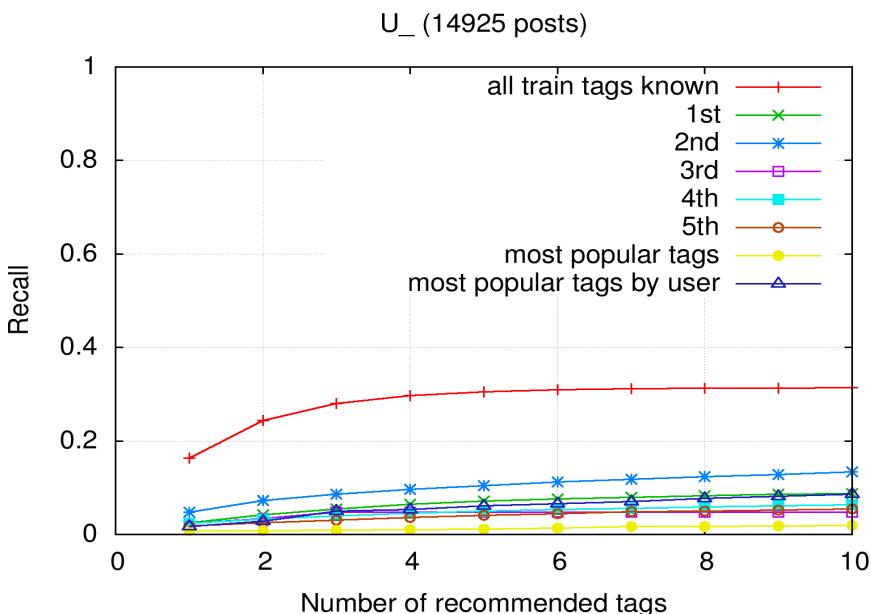
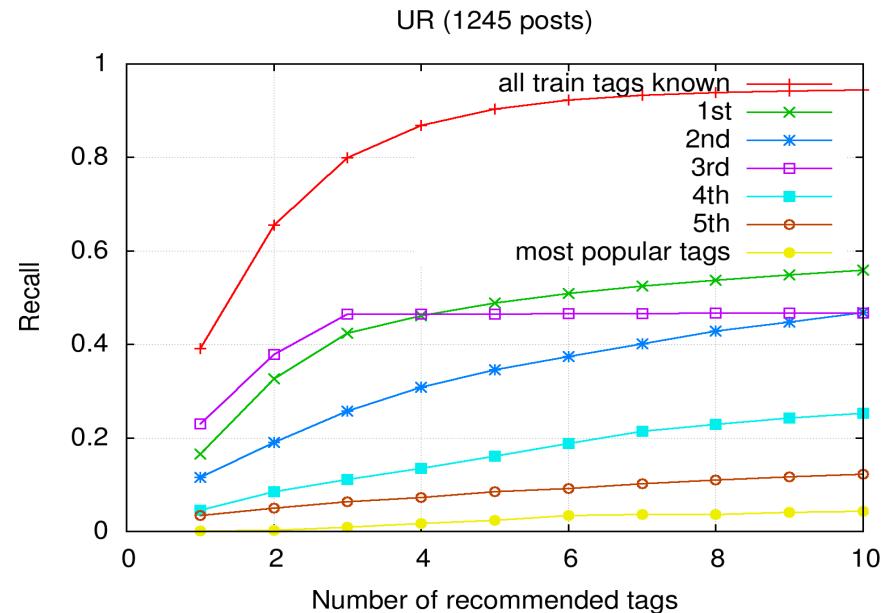
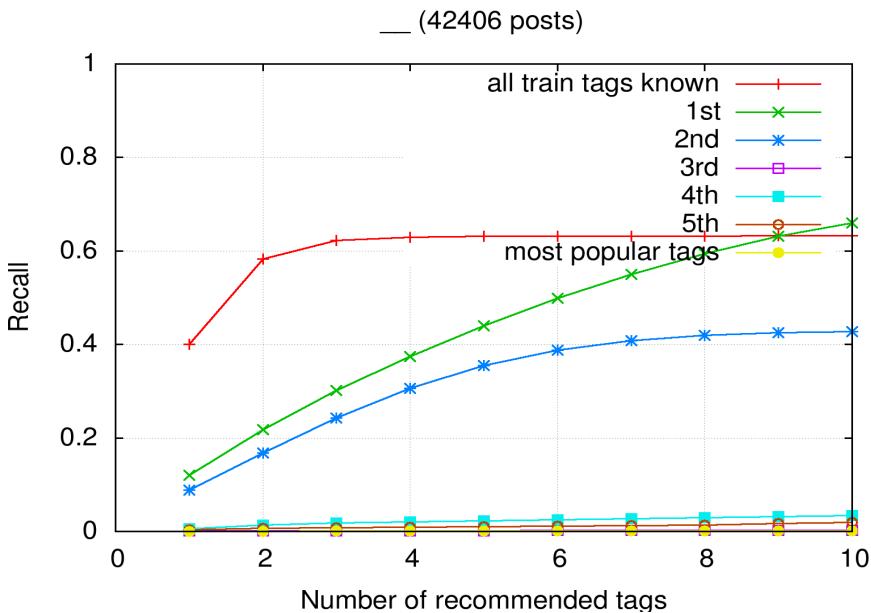
## Results

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

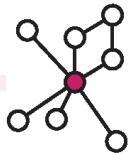
# Tag Recommendation Task



# Tag Recommendation Task



# Fighting Spam



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

# Spam Detection Task



The screenshot shows a web browser window for 'BibSonomy:: - Iceweasel'. The address bar displays 'http://www.bibsonomy.org/'. The page content is a social bookmarking site with the following visible elements:

- BibSonomy ::** search:all <fulltext search
- A blue social bookmark and publication sharing system.
- tags · relations · groups · popular
- myBibSonomy · post bookmark · post publication
- bookmarks** RSS XML
- Online web research service information provider**  
Web Research is a process of online web research information, web search strategies, and information retrieval and Online research information service to b...  
to data entry, information internet mining mining, online research research, retrieval, search search, service strategies, web web, by webresearch on 2008-09-09  
15:31:29  
copy
- Database mining services from outsourcing web research for information retrieval**  
Outsourcing Web Research, offer wide range of accurate web research services like Database mining, information retrieval and online research related service...  
to data database entry, filtering, information mining mining, online research retrieval, searching service, services solutions, support text the web web, by webresearch and 1 other person on 2008-09-09 15:30:47  
copy

**• Growing popularity attracts spam**

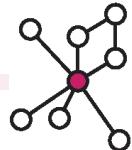
**• Two goals:**

- 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)**

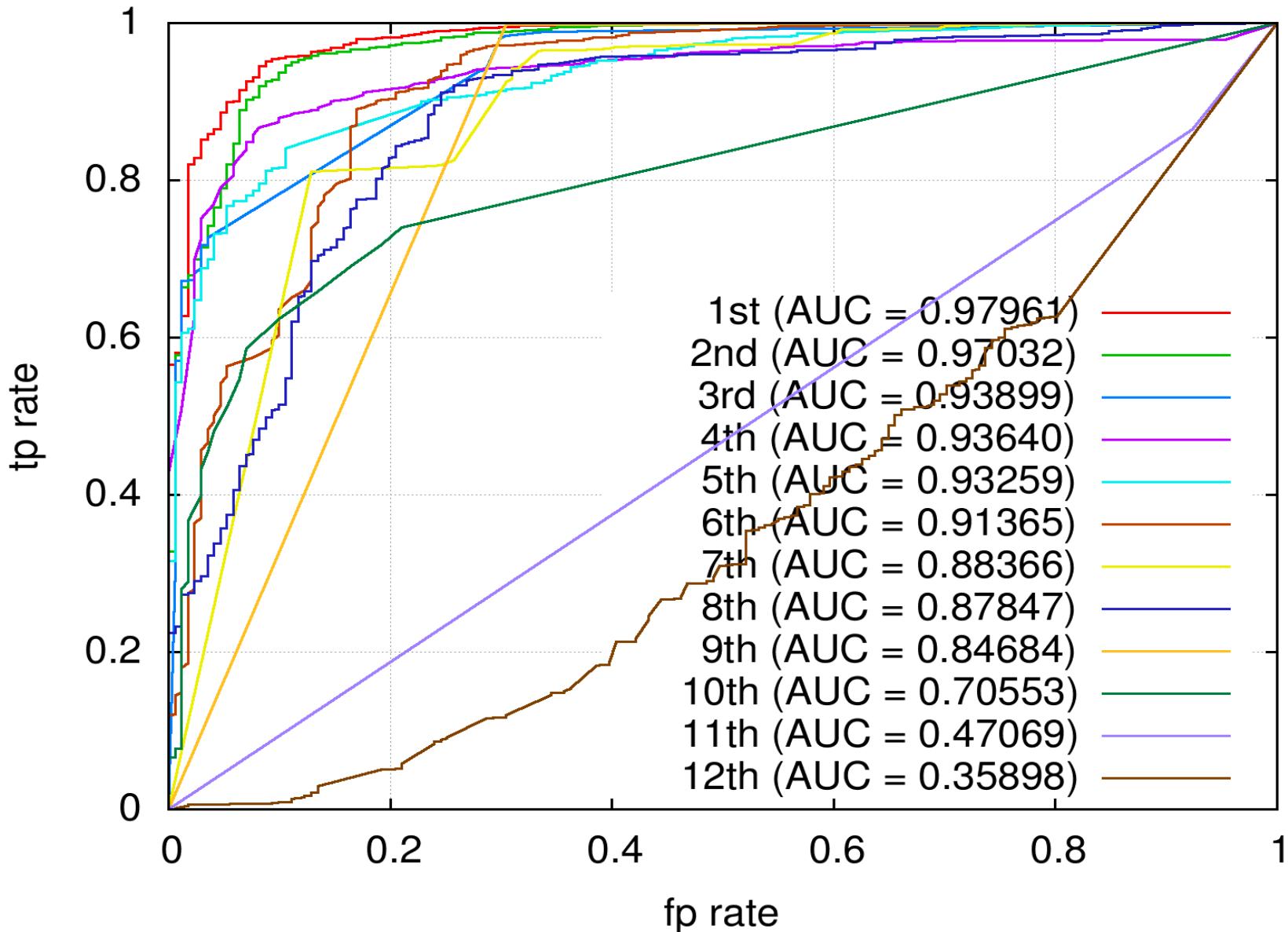
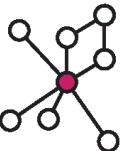
**• Goal: learn a model which predicts whether a user is a spammer or not**



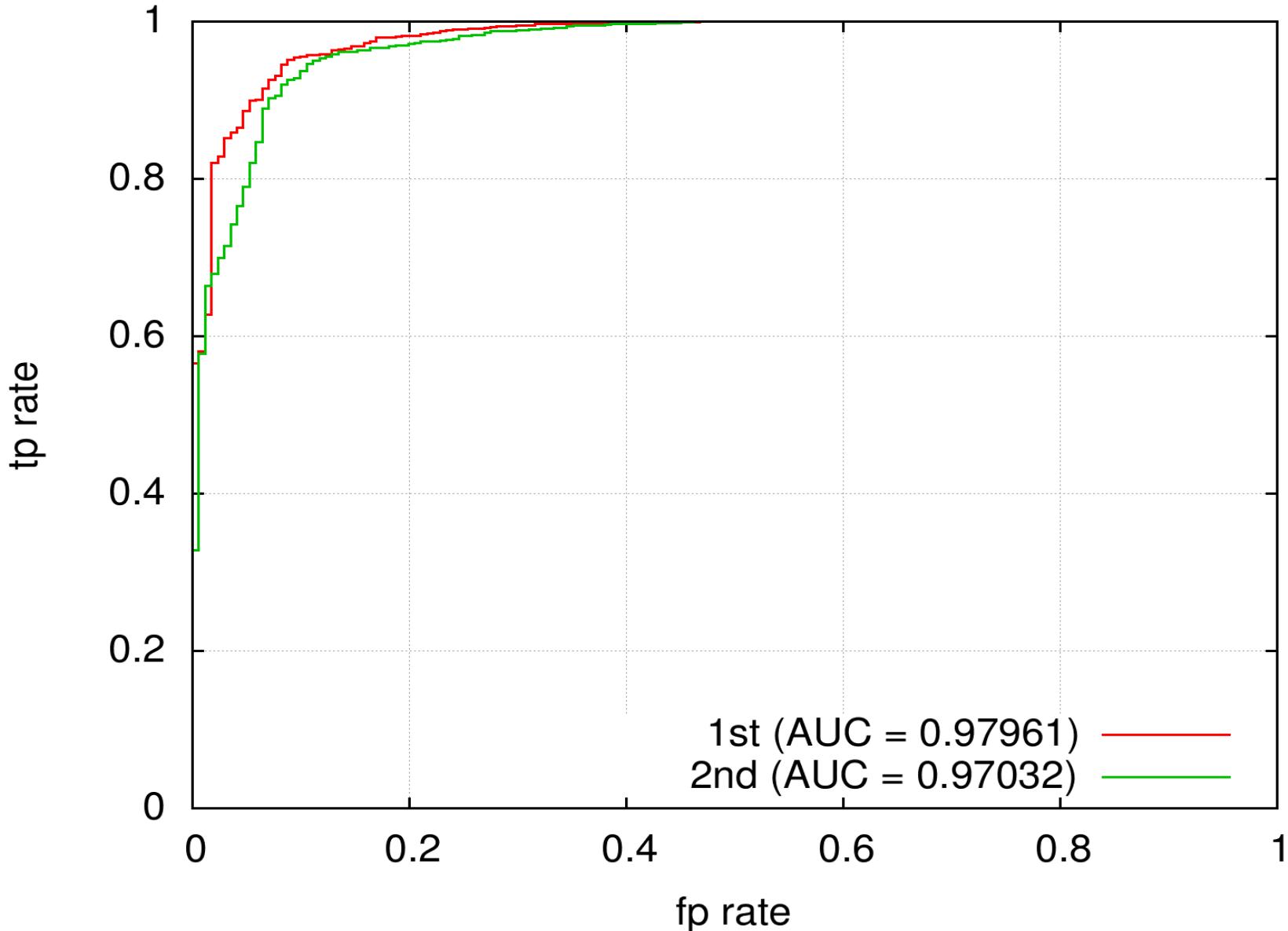
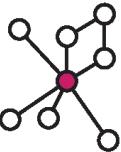
## Results

Sub. ID	AUC	Team
39014	0.97961	<b>A novel supervised learning algorithm and its use for Spam Detection in Social Bookmarking Systems</b> by A. Gkanogiannis and T. Kalamboukis
83234	0.97032	<b>Rank for spam detection - ECML Discovery Challenge</b> by P. Gramme and J.-F. Chevalier
15076	0.93899	<b>Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking</b> 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	

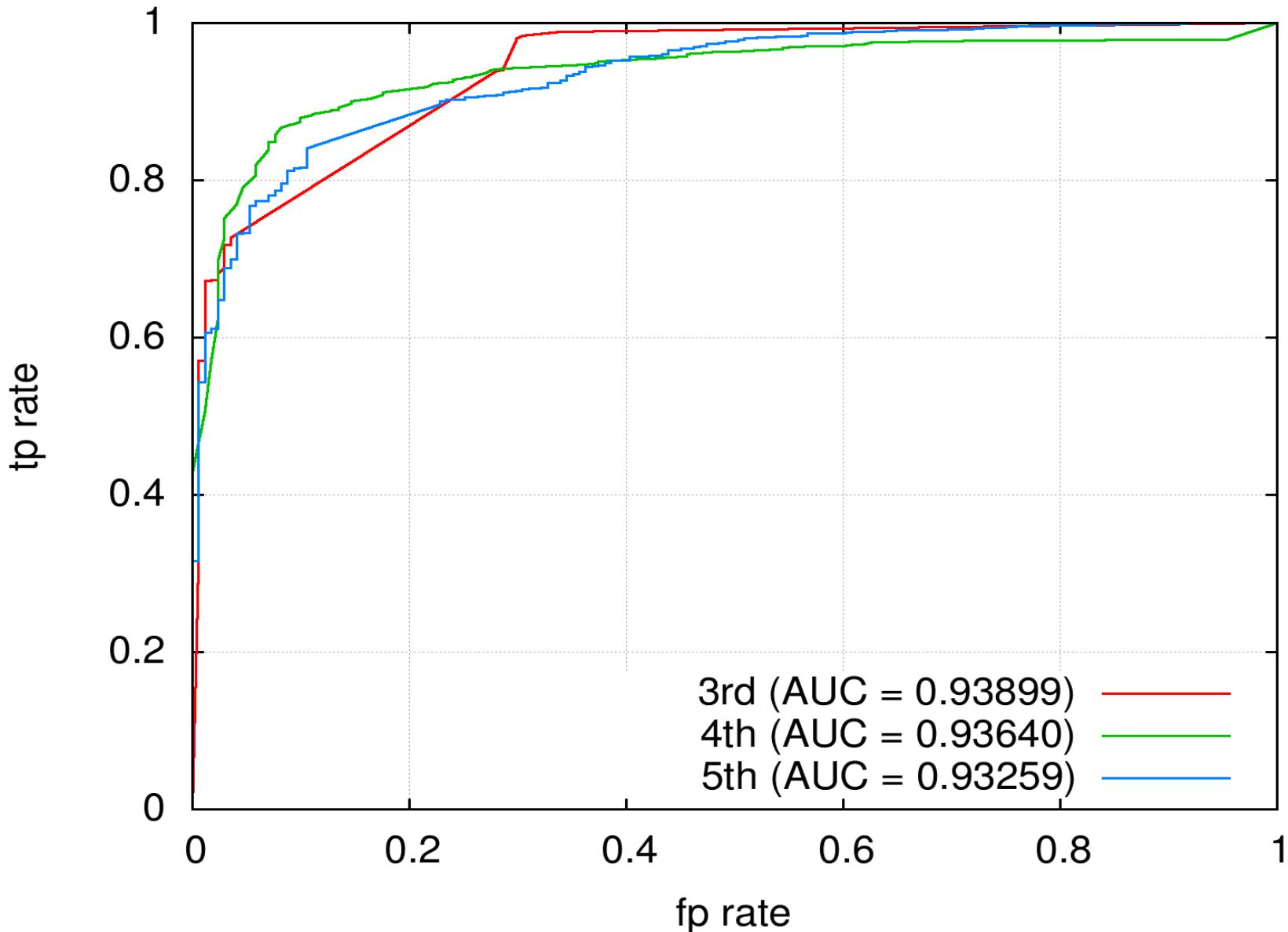
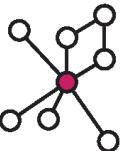
# Spam Detection Task



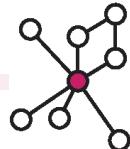
# Spam Detection Task



# Spam Detection Task



## Spam Detection Task

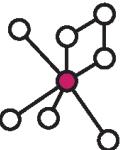


# spammers in BibSonomy

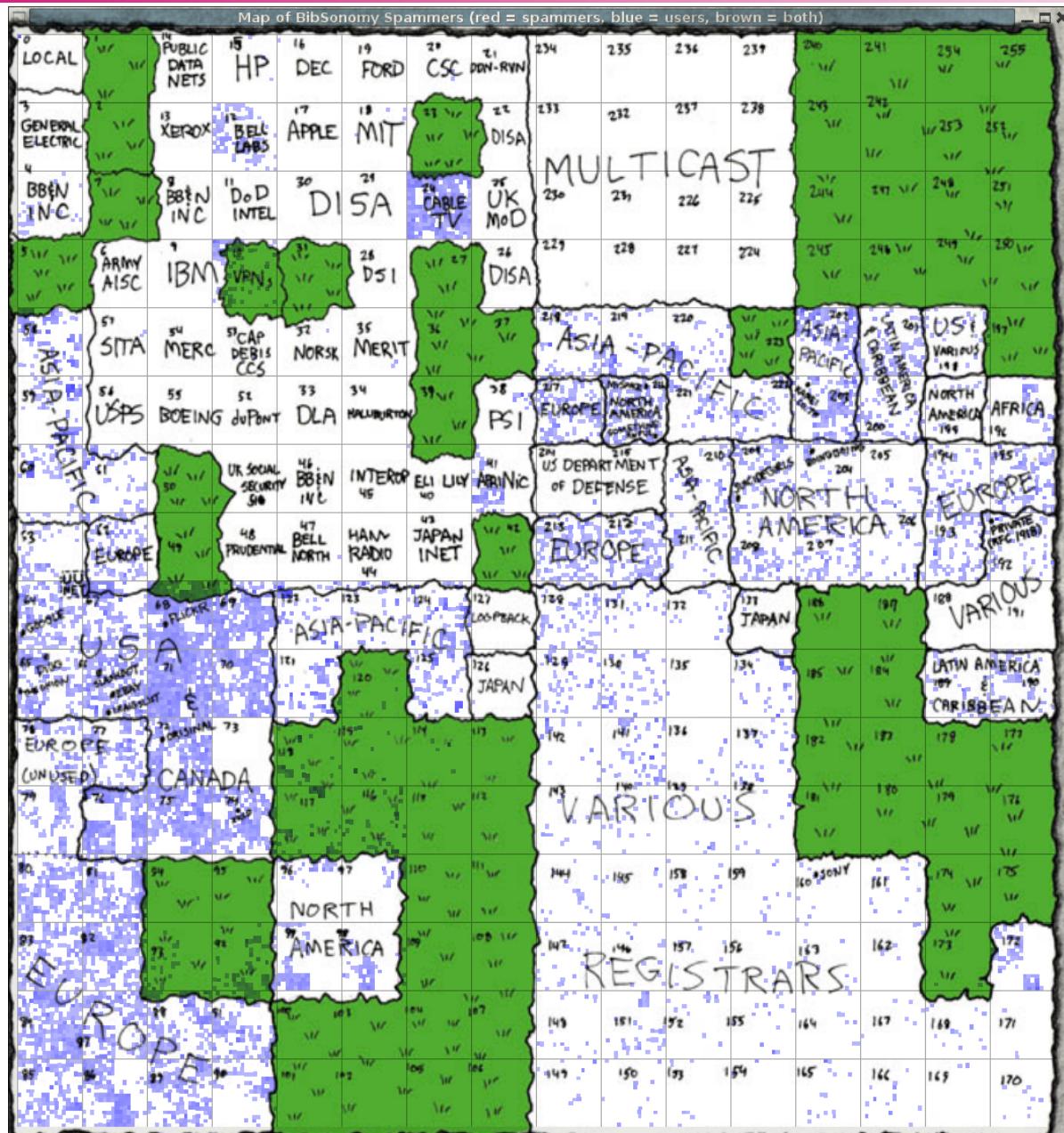


# Map of the Internet provided by <http://xkcd.com/195>

# 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





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!

- Sarbjot 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





## Legend

Discovery Challenge: Spam Detection Task

Discovery Challenge: Tag Recommendation Task

Wikis, Blogs, Bookmarking Tools - Mining the Web 2.0 Workshop

## Time

### Spam

A novel supervised learning algorithm and its use for Spam Detection in Social Bookmarking Systems (30 min)

A. Gkanogiannis and T. Kalamboukis

9:00 -  
10:10

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



## 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 Svatek, Krishna Chandramouli, Jan Nemrava and Ebroul Izquierdo

12:30 -

Lunch

14:00



## 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

BaggTaming - Learning from Wild and Tame Data (30 min)

Toshihiro Kamishima, Masahiro Hamasaki and Shotaro Akaho

15:30 -  
16:00

Coffee break



## 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 Spaming 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. Kalamboukis

Discussion

17:30 - opening of the conference