AIRWeb 2008

Fourth International Workshop on
Adversarial Information Retrieval on the Web

Web Spam Challenge 2008

Carlos Castillo
Yahoo!

Kumar Chellapilla
Microsoft

Ludovic Denoyer
Univ. Paris 6

Beijing 2008
One World, One Web

Web Spam Olympics?

Web Spam ~~Olympics?~~ Challenge

New web spam dataset

Old

- WEBSPAM-UK2006
  - 77M pages
  - 11,402 hosts
  - 7,473 labeled
  - 26% spam

New

- **WEBSPAM-UK2007**
  - 100M pages
  - 114,529 hosts
  - 6,479 labeled
  - 6% spam

`http://www.yr-bcn.es/webspam/datasets/`

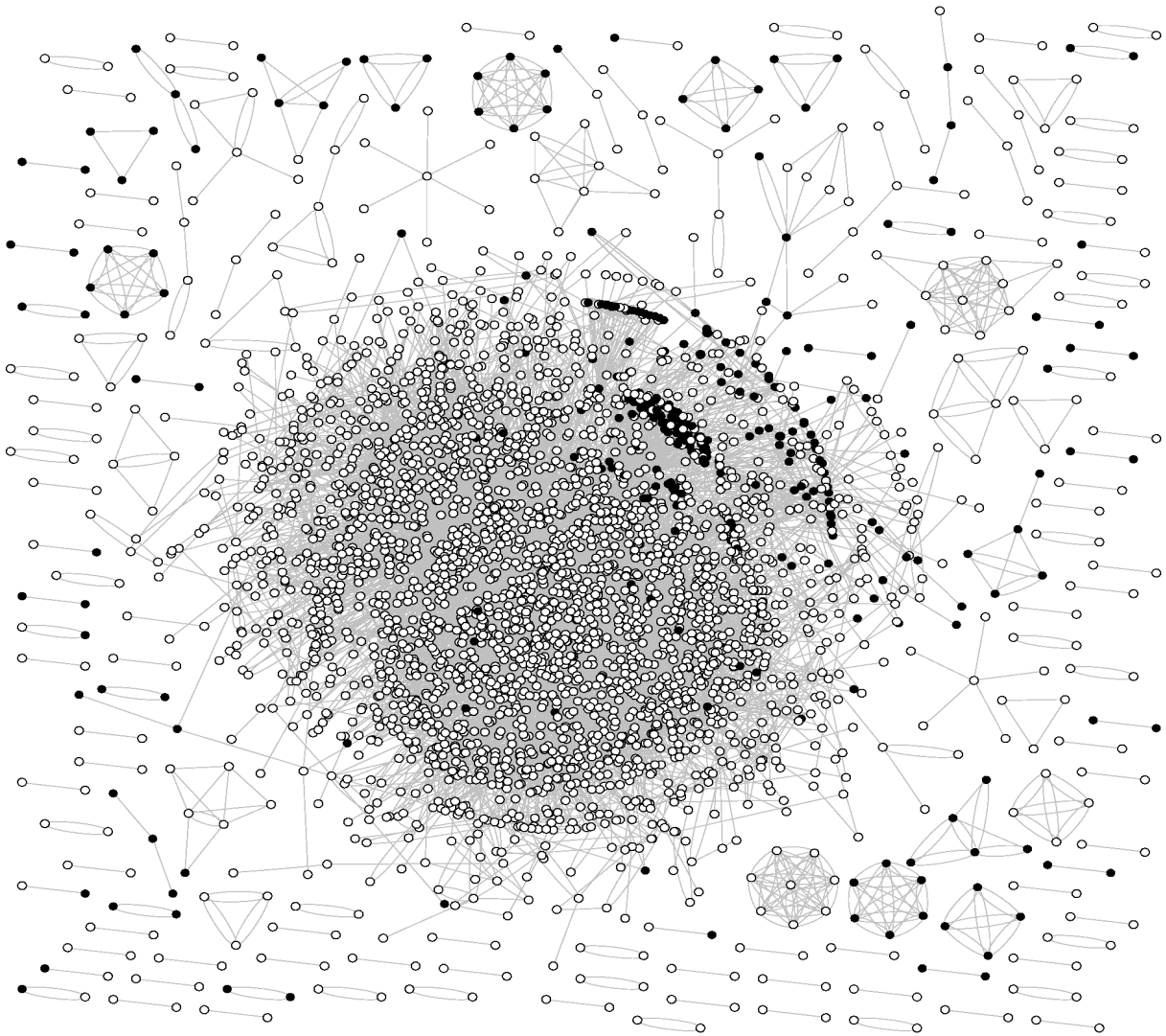
A significant labeling effort

- Thiago Alves, Luca Becchetti, Klaus Berberich, Paolo Boldi, Ilaria Bordino, David Buffoni, Guido Caldarelli, Armando Carvalho, Carlos Castillo, James Caverlee, Carlo Crociani, Na Dai, Brian D. Davison, Matteo Di Gioia, Pascal Filoche, Antonio Gulli, Zoltan Gyongyi, Marcin Hryculak, Thomas Lavergne, Nelly Litvak, Mario Paniccia, Josiane Xavier Parreira, XiaoGuang Qi, Simon Racz, Steve Ross Webb, Maddalena Selis, Fabrizio Silvestri, Elena Smirnova, Marcin Sydow, Sylvie Tricot, Tanguy Urvoy, Yana Volkovich, Jian Wang, Baoning Wu, Bin Zhou
- Two phases of labeling (label + revise)

Timeline

- 2006 May – 11K hosts .UK crawl by Uni Milan
- 2006 Oct – Labeling of data
- 2006 Nov – WEBSPAM-UK2006 data available
- 2007 May – 110K hosts .UK crawl by Uni Milan
- **2007 Apr – Track I competition (IR+ML)**
- 2007 Jun – Training data available
- **2007 Jul – Track II competition (ML)**
- 2007 Dec – Labeling of data
- 2008 Jan – WEBSPAM-UK2007 data available

Old WEBSPAM-UK2006 (used in tracks I and II of 2007)



- 11,402 hosts
- 7,373 labeled
- 26% spam

Track I - April 2007 - 11k hosts

- Metrics: F-Measure and AUC
- Results by AUC:
  - Cormack 0.96
  - Benczur et al. 0.93
  - Filoche et al. 0.93
  - Geng et al. 0.93
  - Abou et al. 0.91

Track I - April 2007 - 11k hosts

- Metrics: F-Measure and AUC
- Results by AUC:
  - Cormack 0.96
  - Benczur et al. 0.93
  - Filoche et al. 0.93
  - Geng et al. 0.93
  - Abou et al. 0.91

4 teams tied in first place according to the evaluation rules

Track I - April 2007 - 11k hosts

- Metrics: F-Measure and AUC
- Results by AUC:
  - Cormack 0.96
  - Benczur et al. 0.93
  - Filoche et al. 0.93
  - Geng et al. 0.93
  - Abou et al. 0.91

4 teams tied in first place according to the rules

Fun factor = 0.0

"Next time we should pick a single winner!"

Track II – July 2007 - Small sample

- ML-only competition
- Results by AUC:
  - Abernethy et al. 0.952
  - Tang et al.0.951
  - Filoche et al. 0.927
  - Csalogany et al. 0.877
  - Tian et al. 0.863

Track II - July 2007 - Large sample

- Results by AUC:
  - Witschel 0.998
  - Filoche 0.991
  - Tang 0.989
  - Csalogany 0.973
- High scores due to sampling by simulated crawling and train/test split at page level (labels are at host level)

WEBSPAM-UK2007

Old

- WEBSPAM-UK2006
  - 77M pages
  - 11,402 hosts
  - 7,473 labeled
  - 26% spam

New

- **WEBSPAM-UK2007**
  - 100M pages
  - 114,529 hosts
  - 6,479 labeled
  - 6% spam

`http://www.yr-bcn.es/webspam/datasets/`

Track III (today) - 110K hosts

- This year's AUC:
  - From 0.73 to 0.85
- Previous year's AUC:

  From 0.91 to 0.96

Harder problem:

- Larger class imbalance
- Better train/test split
- Lower labeling coverage

Rule for photo finish

- Results by AUC:
  - X 0.85 ← 1st place
  - X 0.82
  - X 0.81
  - …

"The team with the highest AUC score will be ranked first. If two consecutively ranked submissions differ by less than 1 percentage point (0.01) in their AUC score a tie will be declared for that rank."

http://webspam.lip6.fr/

This year's participants (1/2)

- Evgeny Skvortsov @ *SFU University, Canada*
  - Constraint programming
- Dávid Siklósi and Andras Benczúr @ *MTA SZTAKI, Hungary*
  - Latent dirichlet allocation, compression, stacked graphical learning
- Guanggang Geng, Xiaobo Jin, and Chunheng Wang @ *CASIA, China*
  - Hostgraph link features, ERUS

This year's participants (2/2)

- Konstantin Bauman, Alexey Brodskiy, Sergey Kacher, Elmira Kalimulina, Ruslan Kovalev, Mikhail Lebedev, Dmitry Orlov, Pavel Sushin, Pavel Zryumov, Dmitry Leshchiner and Ilya Muchnik @ *Data Analysis School Moscow, Russia*
  - Content features, compression, ensemble
- Yuchun Tang, Yuanchen He and Sven Krasser @ *Secure Computing Corp, USA*
  - Random forest
- Jacob Abernethy and Olivier Chapelle @ *UC Berkeley, Yahoo, USA*
  - Learning from content + links