Web Spam Detection via Commercial Intent Analysis

István Bíró,
Computer and Automation Institute,
Hungarian Academy of Sciences

Joint work with András A. Benczúr, Károly Csalogány and Tamás Sarlós

May 8, 2007

Contents

Brief recap of spam

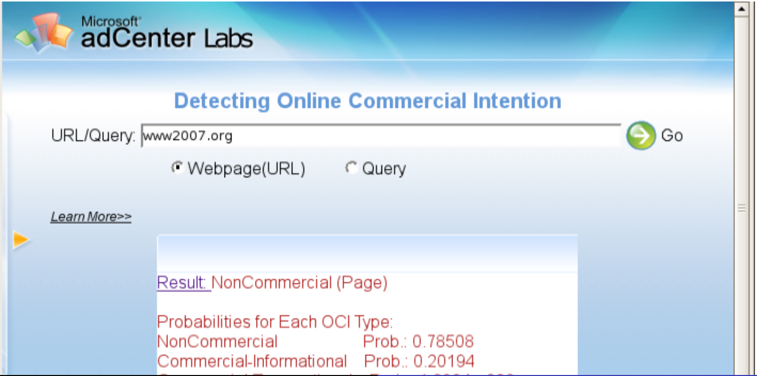
- High revenue for top search engine ratings
- Manipulations, "Search Engine Optimization"
  - content spam – focus of the talk
  - link spam
- Previous content based features: templatic nature of machine generated pages
  - keywords, popular words
  - distribution, entropy, compressibility
- Our Starting Point:
  - Spammers want financial gain [Gyöngyi et al.,2005]
  - Capture the semantics of spam content

Commercial features

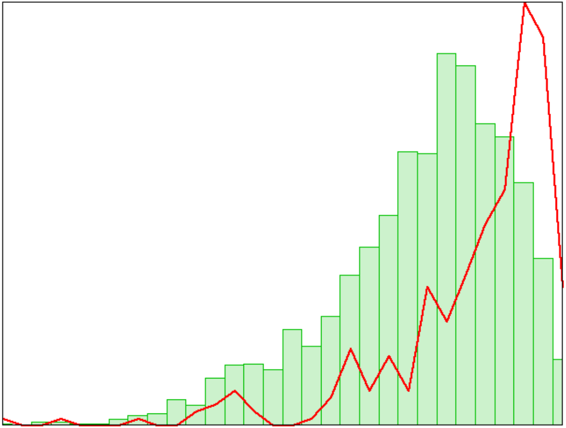
- Online Commercial Intention (OCI) value
- The Yahoo! Mindset
- Google AdWords
- Google AdSense
- Spammer search engine success

Microsoft OCI

- http://adlab.msn.com/OCI/oci.aspx
- commercial-informational, c.-transactional or non-comm.
- SVM utilizing textual content and HTML tags
- Scores obtained for 4995 hosts out of 5622



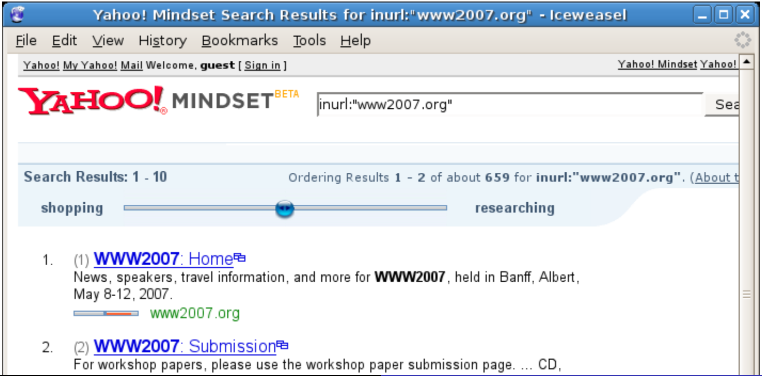
Microsoft OCI



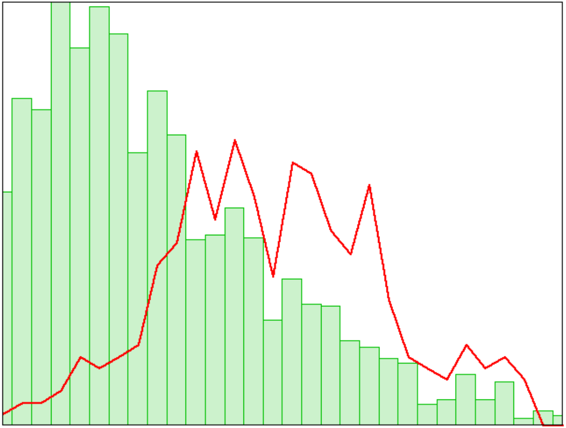
Distribution of commercial-informational score across labeled spam and nonspam sites

Yahoo! Mindset

- http://mindset.research.yahoo.com
- Range from -2 (commercial) to 2 (informational)
- Linear SVM classifier
- Scores obtained for 3170 hosts out of 5622



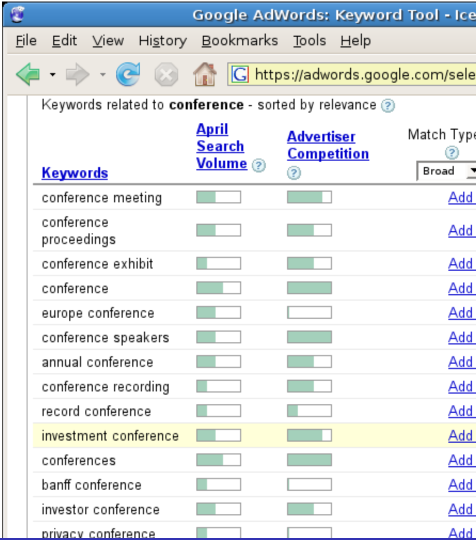
Yahoo! Mindset



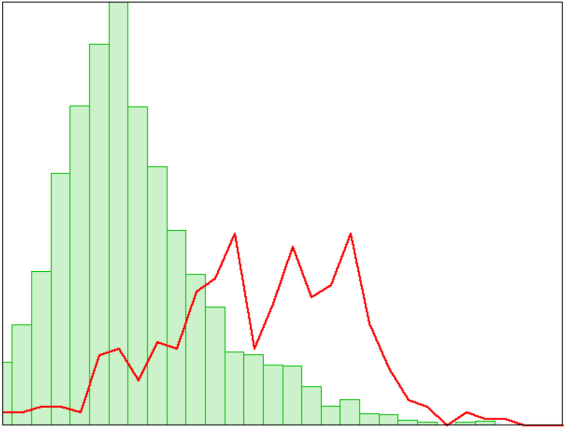
Distribution of Mindset score across labeled spam and nonspam sites

Google Adwords

- http://adwords.google.com
- Adwords Keyword Tool from Google API
  - Search volume, Estimated cost per click (CPC) and ad position etc
  - Advertiser competition: rel. amount of advertisers bidding on that keyword



Google Adwords



Distribution of avg. advertiser competition across labeled spam and nonspam sites

Google AdSense

- http://www.google.com/adsense
- Extracted features:
  - Total number of Google ads over the host
  - Fraction of pages containing at least one ad
  - Average number of Google ads over pages containing ads

Spammer search engine success

- Computed the top 1000 results for the queries composed of keywords with the highest competition score using an IR system.
- Giving $\frac{1}{i^2}$ penalty score to the ith page in ranking
- Features formed by adding up the penalty scores

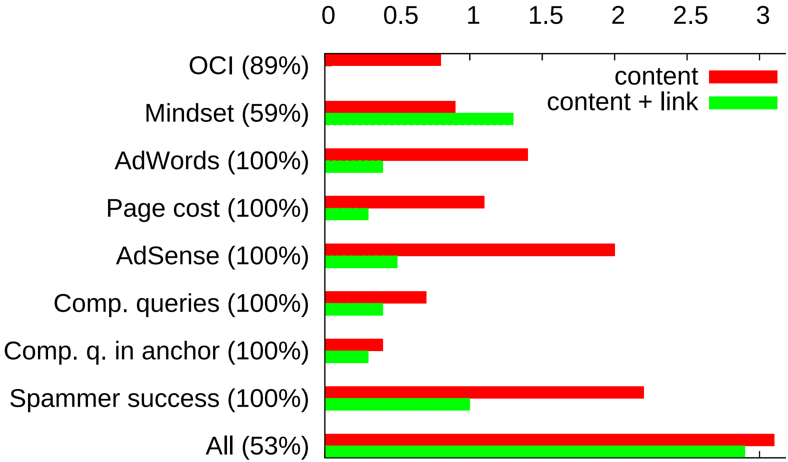
Outline

- Introduction
- Commercial Intent Features
- Evaluation
- Results

Dataset and FrameWork

- WEBSPAM-UK2006 dataset (Domain or Two Humans)
- adding the obtained features to the publicly available
  - content features
  - content + link features
- Weka implementation of C4.5
- Baseline and our results were computed on the hosts that have all features (2922)
- Crossvalidation with the same settings as [Castillo et al., 2006]
- Using Hungarian Academy of Sciences Search Engine
  - tf.idf based ranking combined with 25% HostRank scores
  - increased weights for query words within URL, anchor text, title and additional HTML elements.

F-measure Improvements of Feature Sets



Thank you!

- István Bíró, ibiro@ilab.sztaki.hu
- http://www.ilab.sztaki.hu/websearch