

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

Introduction

Commercial Intent Features

Evaluation

Results

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

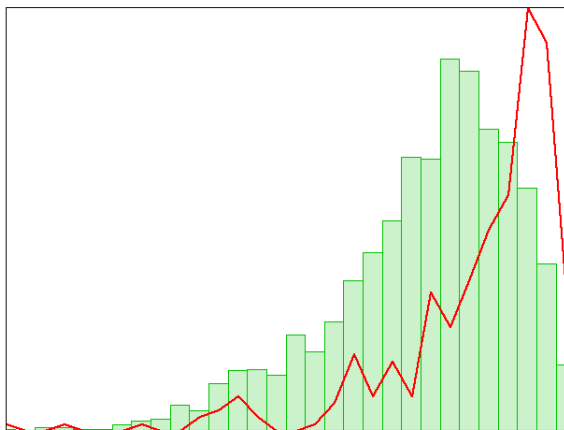
Microsoft OCI

- ▶ <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

The screenshot shows the Microsoft adCenter Labs interface. At the top left is the Microsoft logo and the text 'adCenter Labs'. The main heading is 'Detecting Online Commercial Intention'. Below this is a search bar with the text 'URL/Query: www2007.org' and a green 'Go' button. There are two radio buttons below the search bar: 'Webpage(URL)' (selected) and 'Query'. A link '[Learn More>>](#)' is visible on the left. The results section shows 'Result: NonCommercial (Page)' in red text. Below this, it lists 'Probabilities for Each OCI Type:' with a table:

NonCommercial	Prob.: 0.78508
Commercial-Informational	Prob.: 0.20194

Microsoft OCI



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

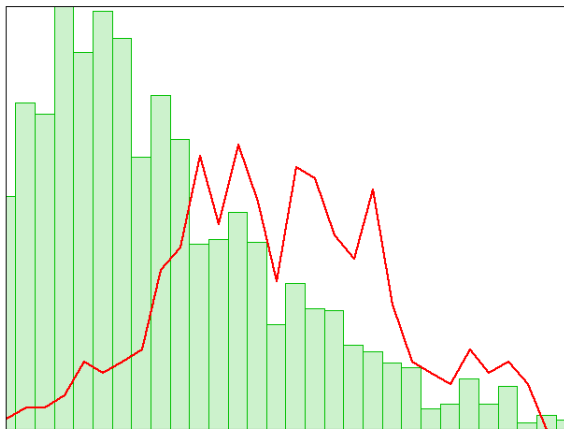
Yahoo! Mindset

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

The screenshot shows a web browser window titled "Yahoo! Mindset Search Results for inurl:'www2007.org' - Iceweasel". The browser's address bar contains the search query "inurl:'www2007.org'". The page header includes the "YAHOO! MINDSET BETA" logo and a search bar. Below the search bar, the results are displayed as "Search Results: 1 - 10" for "inurl:'www2007.org'", with a total of 659 results. A slider interface is visible, ranging from "shopping" to "researching". The first two search results are:

- (1) [WWW2007: Home](#)
News, speakers, travel information, and more for **WWW2007**, held in Banff, Albert, May 8-12, 2007.
[www2007.org](#)
- (2) [WWW2007: Submission](#)
For workshop papers, please use the workshop paper submission page. ... CD,

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: Keyword Tool - Ice

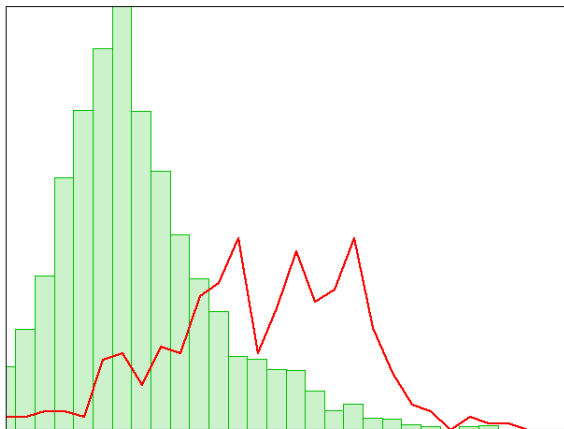
File Edit View History Bookmarks Tools Help

https://adwords.google.com/sele

Keywords related to **conference** - sorted by relevance ?

<u>Keywords</u>	<u>April Search Volume</u> ?	<u>Advertiser Competition</u> ?	Match Type ?
conference meeting	<input type="checkbox"/>	<input type="checkbox"/>	Add
conference proceedings	<input type="checkbox"/>	<input type="checkbox"/>	Add
conference exhibit	<input type="checkbox"/>	<input type="checkbox"/>	Add
conference	<input type="checkbox"/>	<input type="checkbox"/>	Add
europa conference	<input type="checkbox"/>	<input type="checkbox"/>	Add
conference speakers	<input type="checkbox"/>	<input type="checkbox"/>	Add
annual conference	<input type="checkbox"/>	<input type="checkbox"/>	Add
conference recording	<input type="checkbox"/>	<input type="checkbox"/>	Add
record conference	<input type="checkbox"/>	<input type="checkbox"/>	Add
investment conference	<input type="checkbox"/>	<input type="checkbox"/>	Add
conferences	<input type="checkbox"/>	<input type="checkbox"/>	Add
banff conference	<input type="checkbox"/>	<input type="checkbox"/>	Add
investor conference	<input type="checkbox"/>	<input type="checkbox"/>	Add
privacy conference	<input type="checkbox"/>	<input type="checkbox"/>	Add

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 i th page in ranking
- ▶ Features formed by adding up the penalty scores

Outline

Introduction

Commercial Intent Features

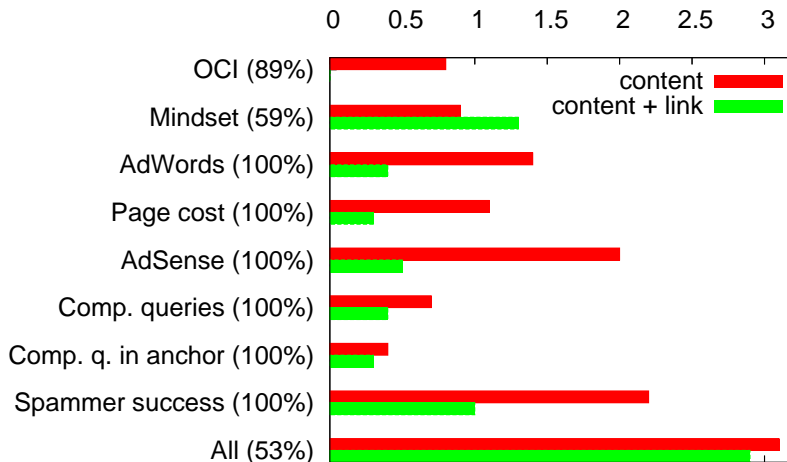
Evaluation

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>