AIRWeb'07

Third International Workshop on
Adversarial Information Retrieval on the Web

Web Spam Challenge Track I

http://webspam.lip6.fr/

Supported by the EU PASCAL Network of Excellente Challenge Program

PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

Web Spam Challenge Overview

- Shared dataset
  - Page links and contents (77M in 11,000 hosts)
  - 500 GB full version, 11 GB summary version (comp.)
- Training labels (~5,500 hosts)
  - And sets of pre-computed features (optional)
- Testing labels (~2,300 hosts)
  - Labeled by participants

Dataset: WEBSPAM-UK2006

1)Crawling of base data

2)Design of guidelines and interface

3)Labeling by volunteers

Crawling of base data

1)Crawling of base data

- Laboratory of Algorithms on the Web, University of Milan: P. Boldi, M. Santini and S. Vigna, using UbiCrawler
- May 2006, .UK domain

2)Design of guidelines and interface

3)Labeling by volunteers

Guidelines and interface

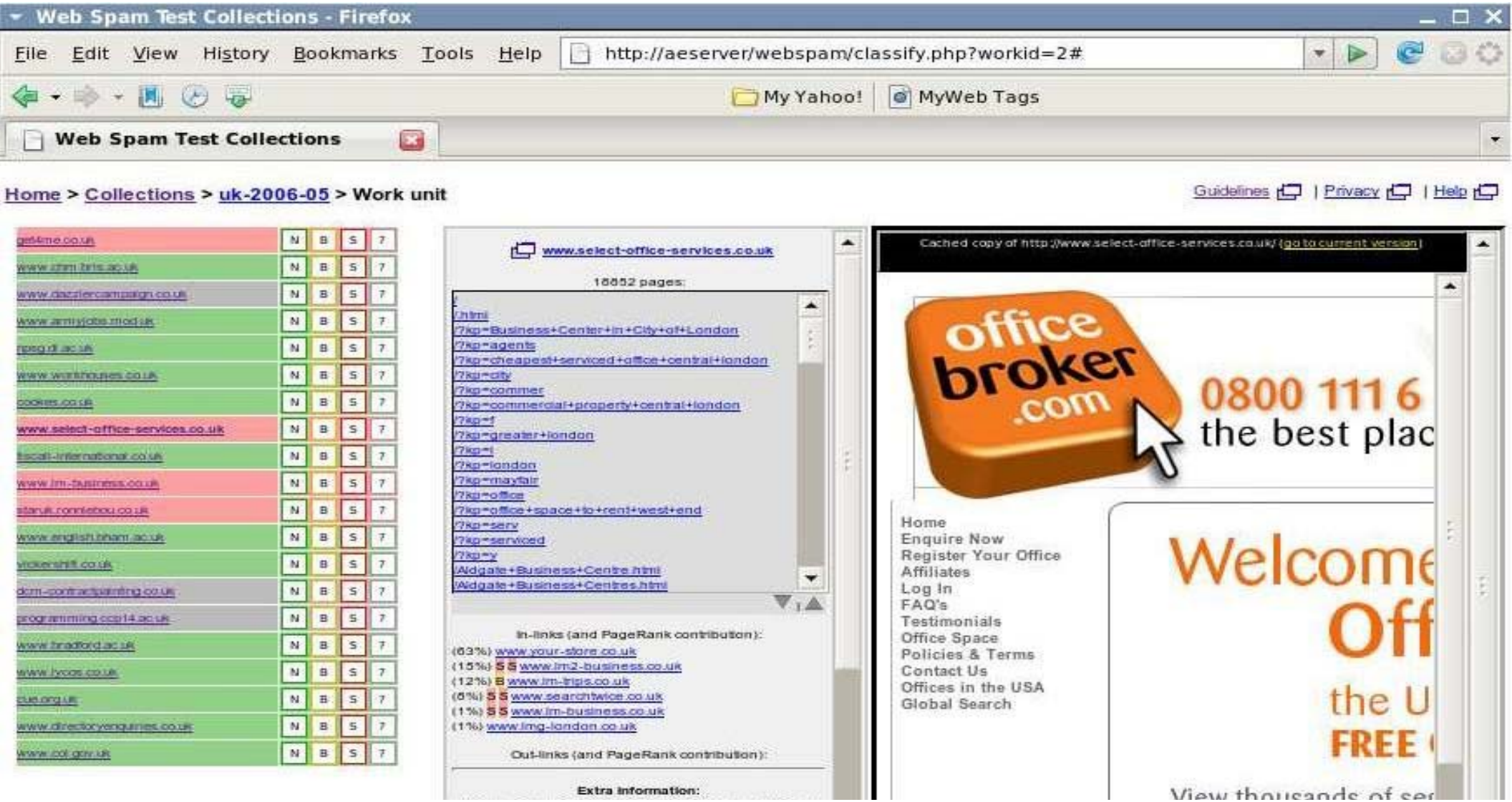
1)Crawling of base data

2)Design of guidelines and interface

- Available at http://www.yr-bcn.es/webspam/
- Examples and guidelines

3)Labeling by volunteers

Interface


Web Spam Test Collections - Firefox
File Edit View History Bookmarks Tools Help
http://aeserver/webspam/classify.php?workid=2#
My Yahoo!
MyWeb Tags
Web Spam Test Collections
Home > Collections > uk-2006-05 > Work unit
Guidelines | Privacy | Help
www.select-office-services.co.uk
16652 pages:
In-links (and PageRank contribution):
Out-links (and PageRank contribution):
Extra information:
Cached copy of http://www.select-office-services.co.uk/ (go to current version)
office broker .com
0800 111 6
the best plac
Home
Enquire Now
Register Your Office
Affiliates
Log In
FAQ's
Testimonials
Office Space
Policies & Terms
Contact Us
Offices in the USA
Global Search
Welcome
Off
the U
FREE

Labeling by volunteers

1)Crawling of base data

2)Design of guidelines and interface

3)Labeling by volunteers

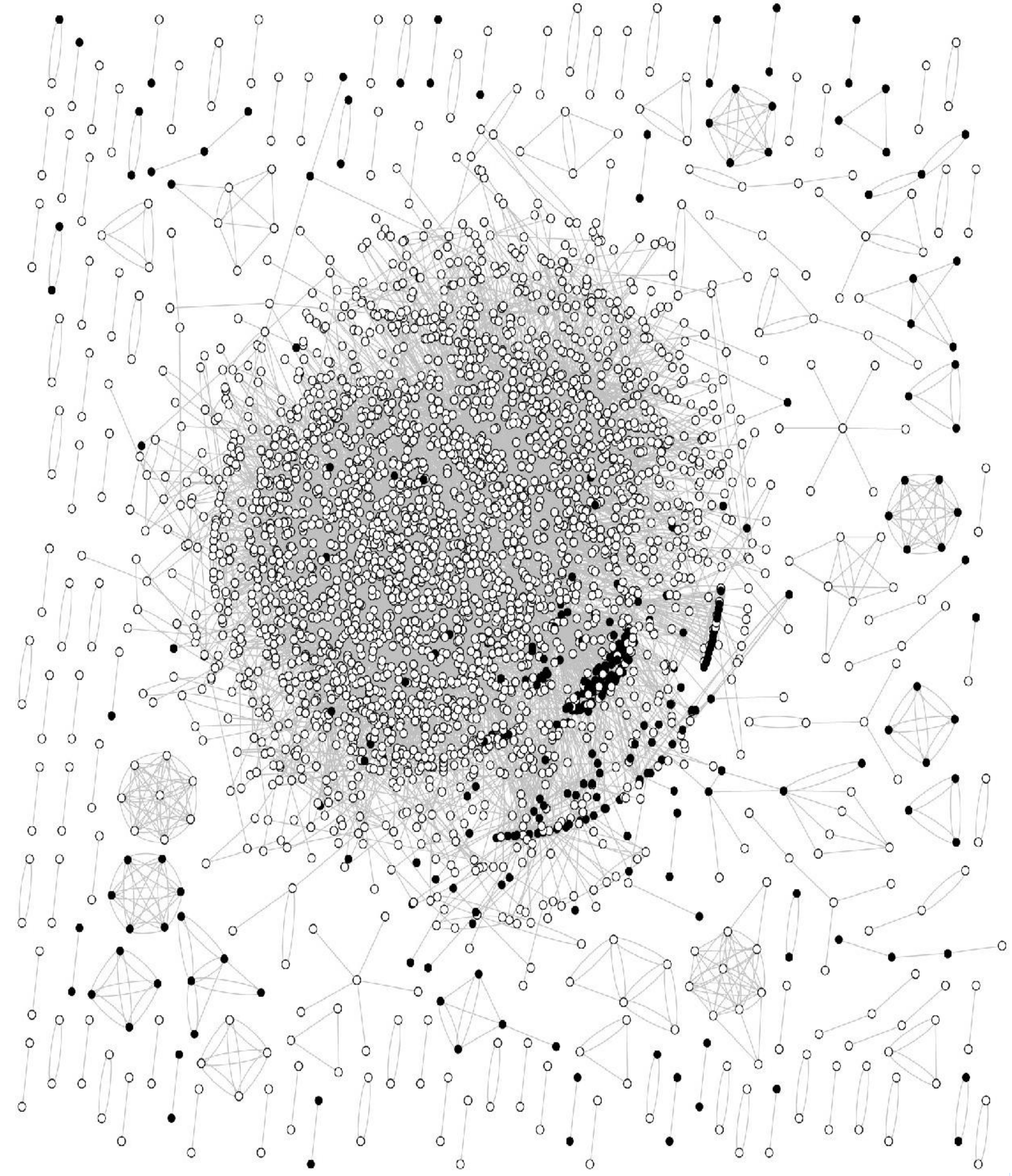
- Host-level labeling as Normal/Borderline/Spam
- 20 volunteers, paired at random for each host
- In general their labels agree

Sample labels

2547	infoserve.qub.ac.uk	AUTO_domain:N		
2548	infrahs.hud.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2549	intoanypaydayloans.datamuseprima.co.uk	josiane:S	brian:S	
2550	insight.zdnet.co.uk	AUTO_dmoz:N	wenzhong:N	
2552	instruments.abcaz.co.uk	antonio:N	chato:B	
2554	insurance.onlineholidays.org.uk	xiaoguang:N	tamas:S	chato:S
2557	intarch.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2558	interact.brighton.ac.uk	AUTO_domain:N		
2559	internal.bath.ac.uk	AUTO_domain:N		
2560	internal.cs.ncl.ac.uk	AUTO_domain:N		
2561	internal.iop.kcl.ac.uk	AUTO_domain:N		
2564	internetmegastores.co.uk	thomas:S	antonio:N	chato:B
2565	interrnt.nhm.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2566	intranet.bton.ac.uk:259	antonio:?	chato:N	
2567	intranet.ses.ic.ac.uk	AUTO_domain:N		
2568	intranet.londonmet.ac.uk	AUTO_domain:N		
2569	intranet.open.ac.uk	AUTO_domain:N		
2570	intranet.salford.ac.uk	AUTO_domain:N		
2571	investing.reuters.co.uk	thomas:N	milos:N	
2572	investing.thisismoney.co.uk	mirel:N	alex:B	
2573	ioesolrserver.ioe.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2574	iq.ias.org.uk	omar:N	zoltan:N	
2575	ion.ls.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2577	ipac.brad.ac.uk	AUTO_domain:N		
2578	ipod.skdirectory.co.uk	mirel:S	luca:?	
2579	iq.lycos.co.uk	pascal:N	zoltan:N	
2580	ir.dcs.gla.ac.uk	AUTO_domain:N	AUTO_dmoz:N	
2581	irish-swingers.connect4fun.co.uk	tanguy:S	zoltan:S	
2583	isdocs.brighton.ac.uk	AUTO_domain:N		
2584	it.auctionsata.co.uk	zoltan:S	sebastiano:S	
2585	itreviews.co.uk	massimo:N	brian:N	
2586	itweek.directorym.co.uk	xiaoguang:N	thiago:N	

>200 man-hours later ...

- 6,500 host labels
  - 62% normal
  - 11% borderline
  - 22% spam
  - 5% can't classify
- More details:
  - C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna: “A Reference Collection for Web Spam”, SIGIR Forum Dec/2006.



Base features (optional)

- Link- and content-based features proposed in the literature, mostly from:
  - Z. Gyöngyi, H. García-Molina and J. Pedersen: “Compating Web Spam with TrustRank”, VLDB 2004
  - A. Ntoulas, M. Najork, M. Manasse and D. Fetterly: “Detecting Web Spam Pages through Content Analysis”, WWW 2006
  - C. Castillo, A. Gionis, D. Donato, V. Murdock and F. Silvestri: “Know your Neighbors: Web Spam Detection Using the Web Topology”, SIGIR 2007

Rules (webspam.lip6.fr)

- Submissions (max. 2 per team)
  - Spam/Nonspam labels and/or "*spamicity*" predictions for each host
- Metrics (computed using KDD-CUP's "perf")
  - F1 measure
    - Combines precision and true positives for a specific threshold
  - Area under ROC curve (for "spamicity" only)
    - ROC curve = false positives vs. true positives for different thresholds

Rules (webspam.lip6.fr)

- Test set labeled by participants
  - Considers all hosts with at least two labels from {normal, borderline, spam}
  - Ground-truth spamicity is average label: normal=0.0, borderline=0.5, spam=1.0
- Remove hosts with ground-truth spamicity=0.5 and evaluate using F1 and AUC (there can be more than one winner)

Rules (webspam.lip6.fr)

- Base evaluation is removing hosts with spamicity=0.5 in the test set
- Two scenarios for these hosts
  - Include them as normal, then test
  - Include them as spam, then test
- Significant difference (for ranking submissions): higher than the difference obtained using these scenarios; otherwise, consider it a tie

Participating Teams

- Tony Abou-Assaleh and Tapajyoti Das
- András A. Benczúr, István Bíró, Károly Csalogány, Miklós Kurucz and Tamás Sarlós
- Gordon Cormack
- Pascal Filoche, Tanguy Urvoy, Chauveau Emmanuel and Lavergne Thomas
- Dennis Fetterly, Steve Chien, Marc Najork, Mark Manasse and Alexandros Ntoulas
- Guanggang Geng, Chunheng Wang, Xiaobo Jin, Qiudan Li and Lei Xu