AIRWeb 2007

Third International Workshop on Adversarial Information Retrieval on the Web

Carlos Castillo
Yahoo! Research

Kumar Chellapilla
Microsoft Live Labs

Brian D. Davison
Lehigh University

8 May 2007 - Banff, Calgary, CANADA
16th International World Wide Web Conference (WWW2007)
The WWW has become the key source of information and medium of interaction for users.

Over the last decade it has grown in size, usage, and has become a means of commerce:
- Several tens of billions of web pages
- 100s of millions of queries per day

Web search is a huge business and attracts a lot of interest:
- Both for organic and paid search
Search Engines and Content Providers

- Search engines have become a major source of discovery and traffic for content providers

- Search Engine Providers
  - Want to provide relevant results to users

- Content Providers
  - Want to understand search ranking
  - Want to optimize their content to rank well for certain queries
  - How far is too far? (aggressive SEO vs Spam)
Both content providers and search engines affect quality of search results
  - Interactions can be complex

Adversarial Relationship
  - Improving relevance of results ⇔ demoting irrelevant spam results
  - Spam results ranking high => content provider receives high traffic

Central theme of AIRWeb
  - Understanding this relationship and improve web search
AIRWeb 2007

- Ten full papers and three short papers out of twenty one submissions
- Three categories
  - Temporal and Topological Factors
  - Link Spam
  - Tagging, P2P, and Cloaking
- Web Spam Challenge
  - Nine submissions from six teams
- Builds on successful past workshops at WWW’2005 and SIGIR’2006
Testing Web spam detection systems

- Novel element this year
- Supported by
  - EU Network of Excellence PASCAL Challenge Program
    - Organizer: Ludovic Denoyer
  - DELIS EU-FET research project
    - Web pages, web graph, human labels (UK-2006)

Participants helped evaluate submissions by labeling hosts
Acknowledgements

- Program Committee members
  - Lent their time, advice, and reputation
- Authors
  - Submitted interesting papers
- Workshop participants
  - Letting us all get to know you!
- Web spam challenge participants
  - Submitted predictions and labeled hosts

- Student travel support (four students)

Microsoft Live Labs™
### PC Members

- Einat Amitay
- Soumen Chakrabarti
- Nick Craswell
- Aaron D'Souza
- Edel García
- Zoltán Gyöngyi
- Ronny Lempel
- Marc Najork
- Erik Selberg
- Matt Wells
- András Benczúr
- Paul-Alexandru Chirita
- Matt Cutts
- Dennis Fetterly
- Natalie Glance
- Monika Henzinger
- Mark Manasse
- Jan Pedersen
- Mike Thelwall
- Baoning Wu
- Andrei Broder
- Tim Converse
- Ludovic Denoyer
- Tim Finin
- Antonio Gulli
- Jeremy Hylton
- Gilad Mishne
- Tamás Sarlós
- Andrew Tomkins
- Tao Yang

### Additional Reviewers

- Beitao Li, Giuseppe Ottaviano, Luca Foschini, Stefano Cataudella, and Xiaoguang Qi.
Workshop Schedule

- **8:30** Welcome
- **8:40** Session 1: Temporal and Topological Factors
- **10:00** Morning Break
- **10:30** Session 2: Link Farms
- **12:00** Lunch Break
- **1:30** Session 3: Tagging, P2P, and Cloaking
- **3:00** Afternoon Break
- **3:30** Web Spam Challenge
- **5:00** Discussion & Final Remarks
- Updated program is available online
  - http://airweb.cse.lehigh.edu/2007/program.html
  (order of papers on CD-ROM is out dated)
- Hard copies available