Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2011)

Carlos Castillo[1], Zoltan Gyongyi[2], Adam Jatowt[3] and Katsumi Tanaka[3]

[1]Yahoo! Research
Avinguda Diagonal 177, 08018
Barcelona, Spain
chato@yahoo-inc.com

[2] Google Research
1600 Amphitheatre Parkway,
Mountain View, CA 94043, USA
zoltang@google.com

[3]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
{adam, kanazawa, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

In this paper we overview the Joint WICOW/AIRWeb Workshop on Web Quality[1] (WebQuality 2011) that was held in conjunction with the 20th International World Wide Web Conference in Hyderabad, India.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Measurement, Experimentation

Keywords

Web information quality, web information credibility, spam detection, adversarial information retrieval,

1. OBJETIVES

WebQuality 2011 was held on March 28th, 2011 as a joint WICOW/AIRWeb workshop. WICOW (International Workshop on Information Credibility on the Web) workshops have addressed information credibility on the Web in 4 previous editions (2007-2010), while AIRWeb (Adversarial Information Retrieval on the Web) installments have covered adversarial information retrieval issues in 5 previous editions (2005-2009). The main topics of the two workshop series had been on a path of convergence, due to the continued diversification and fragmentation of web content, the increasing sophistication of manipulation attempts, and the growth in author base, particularly facilitated by emerging social media. Accordingly, a joint workshop catering for the larger research community interested in web content quality issues in general was held at WWW2011.

On one hand, the joint workshops aimed to cover the more blatant and malicious attempts that deteriorate web quality—such as spam, plagiarism, or various forms of abuse—and ways to prevent them or neutralize their impact on information retrieval. On the other hand, it also provided a venue for exchanging ideas on quantifying finer-grained issues of content credibility and author reputation, and modeling them in web information retrieval.

The main objective of the workshop was to provide the research communities working on web spam, abuse, credibility, and reputation topics with a survey of current problems and potential solutions. It was meant to present an opportunity for close interaction between practitioners who may have focused on more isolated sub-areas previously.

For an open publication platform such as the World Wide Web, content quality is a central issue. Low publishing barriers lead to very limited quality control, which results in the proliferation of mistaken, unreliable, and sometimes outright intentionally misleading information. Low quality (textual or multimedia) content can have detrimental effects on users, especially in the light of the ever-increasing role the Web plays in our daily lives. Content quality challenges call for technology that facilitates judging the trustworthiness of content and assessing the accuracy of the information. Some of these challenges and technologies are not fundamentally new: search engine spam is over a decade old now, and content credibility problems have received a fair share of research attention in the past few years as well. However, novel web content quality issues abound as various forms of adversarial behavior gain in sophistication, and as new groups of users and web platforms (such as microblogging services or local recommendation engines) emerge.

Besides the paper presentations we were pleased to have a keynote speech delivered by Elisa Bertino. Elisa is professor at the Computer at the Department of Computer Sciences, Purdue University and Research Director of CERIAS. Her main research interests cover many areas in the fields of information security and database systems.

2. TOPICS

The main themes of the workshop were that of evaluating web information credibility, and identifying and combating qualitatively extreme content (and related behavior), such as spam. These themes encompass a large set of often-related topics and subtopics, as listed below.

Assessing the credibility of content and people on the web and social media

Measuring quality of web content

- Detecting disagreement and conflicting opinions

[1] http://www.dl.kuis.kyoto-u.ac.jp/webquality2011/

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

- Information quality and credibility of web search results, on social media sites, of online mass-media and news, and on the Web in general
- Estimation of information age, provenance, validity, coverage, and completeness or depth
- Formation, change, and evolution of opinions
- Sociological and psychological aspects of information credibility estimation
- Users studies of information credibility evaluation

Uncovering distorted and biased content

- Detecting disagreement and conflicting opinions
- Detecting disputed or controversial claims
- Uncovering distorted or biased, inaccurate or false information
- Uncovering common misconceptions and false beliefs
- Search models and applications for finding factually correct information on the Web
- Comparing and evaluating online reviews, product or service testimonials

Modeling author identity, trust, and reputation

- Estimating authors' and publishers' reputation
- Evaluating authors' qualifications and credentials
- Transparent ranking/reputation systems
- Author intent detection
- Capturing personal traits and sentiment
- Modeling author identity, authorship attribution, and writing style
- Systems for managing author identity on the Web
- Revealing hidden associations between authors, commenters, reviewers, etc.

Role of groups and communities

- Role of groups, communities, and invisible colleges in the formation of opinions on the Web
- Social-network-based credibility evaluation
- Analysis of information dissemination on the Web
- Common cognitive or social biases in user behavior
- Credibility in collaborative environments (e.g., on Wikipedia)

Multimedia content credibility

- Detecting deceptive manipulation or distortion of images and multimedia
- Hiding content in images
- Detecting incorrect labels or captions of images on the Web
- Detecting mismatches between online images and the represented real objects
- Credibility of online maps

Fighting spam, abuse, and plagiarism on the Web and social media

Reducing web spam

- Detecting various types of search engine spam (e.g., link spam, content spam, or cloaking)
- Uncovering social network spam (e.g., serial sharing and lobbying) and spam in online media (e.g., blog, forum, wiki spam, or tag spam)
- Identifying review and rating spam
- Characterizing trends in spamming techniques

Reducing abuses of electronic messaging systems

- Detecting e-mail spam
- Detecting spit (spam over internet telephony) and spim (spam over instant messenger)

Detecting abuses in internet advertising

- Click fraud detection
- Measuring information credibility in online advertising and monetization

Uncovering plagiarism and multiple-identity issues

- Detecting plagiarism in general, and in web communities, social networks, and cross-language environments in particular
- Identifying near-duplicate and versioned content of all kinds (e.g., text, software, image, music, or video)
- High-similarity retrieval technologies (e.g., fingerprinting and similarity hashing)

Promoting cooperative behavior in social networks

- Monitoring vandalism, trolling, and stalking
- Detecting fake friendship requests with spam intentions
- Creating incentives for good behavior in social networks
- User studies of misuse of the Web

Security issues with online communication

- Detecting phishing and identity theft
- Flagging malware (e.g., viruses and spyware)
- Web forensics

Other adversarial issues

- Modeling and anticipating responses of adversaries to counter-measures
- New web infringements
- Web content filtering
- Bypassing censorship on the Web
- Blocking online advertisements
- Reverse engineering of ranking algorithms
- Stealth crawling

3. PC MEMBERS

We list the names and affiliations of PC members[2] below:

Andras Benczur (Hungarian Academy of Sciences)

James Caverlee (Texas A&M University)

Gordon Cormack (University of Waterloo)

Matt Cutts (Google)

Brian Davison (Lehigh University)

Dennis Fetterly (Microsoft)

Andrew Flanagin (University of California, Santa Barbara)

Miriam Metzger (University of California, Santa Barbara)

Andrew Tomkins (Google)

Masashi Toyoda (University of Tokyo)

Steve Webb (Georgia Institute of Technology)

Min Zhang (Tsinghua University)

Xiaofang Zhou (University of Queensland)

[2] As of January 31st, 2011.