Report on the 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,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract

The Joint WICOW/AIRWeb Workshop on Web Quality [i] (WebQuality 2011) was held in conjunction with the 20th International World Wide Web Conference in Hyderabad, India on the 28th March 2011. Seven full-papers presentations and a keynote talk were delivered in three sessions. This report briefly summarizes the workshop.

1 Introduction

WebQuality 2011 was held on 28th March 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 was held at WWW 2011, which catered for the larger research community interested in web content quality issues in general.

On one hand, the joint workshop 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 had 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 and local recommendation engines) emerge.

2 Keynote

The workshop commenced with an invited talk delivered by *Elisa Bertino*: "*Assuring Data Trustworthiness – Concepts and Research Challenges*". *Elisa Bertino* 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.

The talk motivated the investigation of data trustworthiness, classified existing approaches, and emphasized the multi-faceted semantics and versatility of trustworthiness as two particular challenges. It argued for a "trust fabric," the combination of identity, provenance, usage, and attack management for assuring trustworthiness. Two examples, a solution for provenance-based trust management for data streams and a high-level discussion of location data trustworthiness, concluded the talk.

3 Paper Presentations

The keynote was followed by two research sessions, with paper presentations briefly introduced next.

3.1 Web Spam Session

The first session started with the presentation of the paper entitled "*Web Spam Classification: a Few Features Worth More*" by *Miklós Erdélyi, Andras Garzo* and *András A. Benczúr*. The authors evaluated the relative impact of various available features on the classification accuracy of web spam. They argued for the importance of an appropriately selected learning method, and experimentally demonstrated that the right method along with a small and computationally inexpensive set of features can outperform previous, more expensive approaches.

Next was the presentation of a paper entitled "*Spam Detection in Online Classified Advertisements*" by *Hung Tran, Thomas Hornbeck, Viet Ha-Thuc, James Cremer* and *Padmini Srinivasan*. The authors introduced a novel set of features, specific to online advertisements, which can be used to discriminate between legitimate and spam posts on classified advertisement sites, such as Craigslist and Ebay Classifieds. They also presented experimental evidence supporting the effectiveness of the new features over classical web spam features.

In the last paper presented in the Web Spam session, entitled "*Improving Malicious URL Re-Evaluation Scheduling through an Empirical Study of Malware Download Centers*", *Kyle Zeeuwen, Matei Ripeanu* and *Konstantin Beznosov* observed the update behavior of malware download centers of the Web over a period of four month. Their empirical findings led to a classification of such centers and provided general guidelines and, more specifically, a scheduling algorithm for crawlers

that periodically re-query them for binary updates.

3.2 Web Quality Session

The second session started with the presentation of the paper *"Characterizing the Uncertainty of Web Data: Models and Experiences"* by *Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo* and *Paolo Papotti*. The paper described studies on the precision of structured web data. For a set of common objects, attribute values were extracted from different information providers in order to evaluate the accuracy of the provided information and the distribution of attribute values. The authors also considered in their approach data replication behaviors that are common on the Web.

The second paper of the Web Quality session, *"Modeling and Evaluating Credibility of Web Applications"* by *Adriano Pereira, Sara Guimarães, Arlei Silva* and *Wagner Meira Jr.* described a framework for the design, implementation, and evaluation of credibility models. The authors presented different credibility models using different types of information sources, such as attributes related to a (commercial) offer's characteristics or a seller's expertise and qualification, and then evaluated their effectiveness on a real dataset obtained from an electronic market.

The third paper, entitled "*Got Traffic? An Evaluation of Click Traffic Providers,*" was authored by *Qing Zhang, Thomas Ristenpart, Stefan Savage* and *Geoffrey M. Voelker*. The paper evaluates the quality of web traffic from different traffic providers. Since premium providers such as Google AdWords use a pay-per-click auction model, a variety of providers offer click traffic. The evaluation of the click traffic directed to sites relied on a variety of metrics, including timing properties, access patterns on the site, network properties of the hosts accessing the site, and correlation with blacklists.

The last paper presented in this session was "*Web Information Analysis for Open-domain Decision Support: System Design and User Evaluation*" by *Takuya Kawada, Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Yutaka Leon, Kentaro Inui, Sadao Kurohashi* and *Yutaka Kidawara*. The authors investigate the effectiveness of a system designed for supporting open-domain decision-making by multi-aspect analysis of information on the Web. The proposed system, called *WISDOM*, extracts author information and other types of data, assisting in the qualitative evaluation of web information, such as major opinions or their associated sentiment values. User experiments and task-focused comparison with a popular search engine were conducted to verify the effectiveness of the system.

4 Acknowledgements

We would like to thank the organizers of the WWW 2011 conference for accommodating our workshop. We also express our gratitude to the workshop PC members for their dedicated work and the participants for contributing to the workshop's success.

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