Report on the WebQuality 2012 Workshop

Carlos Castillo
Qatar Computing Research Institute
Doha, Qatar
chato@acm.org

Zoltan Gyongyi
Google Research
1600 Amphitheatre Pkwy,
Mountain View, CA 94043, USA
zoltang@google.com

Adam Jatowt
Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Katsumi Tanaka
Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

Abstract

The 2nd Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2012) was held in conjunction with the 21st International World Wide Web Conference in Lyon, France on the 16th April 2012. Seven full and three short paper presentations were delivered in three sessions. This report briefly summarizes the workshop.

1 Introduction

WebQuality 2012 was held on 16th April 2012 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 first joint workshop was held at WWW 2011 and catered for the larger research community interested in web content quality issues in general. This report summarizes the second joint workshop that was held in conjunction with the WWW 2012 conference.

For an open publication platform such as the World Wide Web, content quality is a central issue. Low publishing barriers lead to limited quality control, enabling the proliferation of mistaken, unreliable, and sometimes outright intentionally misleading information. Low quality (textual or multimedia) content can have detrimental effects on users relying on the Web as an increasingly important source of information in their daily lives. Content quality challenges call for technology that facilitates judging the trustworthiness of content and assessing the accuracy of the information. Some of the respective 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. At the same time, novel content quality challenges abound as various forms of adversarial behavior gain in sophistication, and as new groups of users, online publishing platforms, and interaction models emerge.

With a desire to give special attention to novel challenges, the subtitle of the WebQuality 2012 workshop was "The Antisocial Web: Credibility and Quality Issues on the Web and Social Media". As this subtitle indicates, the workshop recognized that the rapidly growing corpora of socially curated web content, as well as the user-generated content of online social media face specific quality and abuse problems. Indeed, approximately half of the workshop was dedicated to such problems, with relevant research presented and discussed in each of the sessions.

2 Paper Presentations

From among the 20 research paper submissions, 7 full and 3 short papers have been selected by the program committee for presentation at the event and for publication in the ACM Digital Library. Each submission was reviewed by at least three program committee members. The accepted papers were presented in three sessions: "Web Quality", "Online Credibility and Trust", and "Abuse Detection and Prevention". A brief summary of each session is provided next.

2.1 Web Quality Session

The first paper in the session, authored by Ricardo Baeza-Yates and Luz Rello, was titled "On Measuring the Lexical Quality of the Web". The authors did a large-scale quality measurement of web pages in English and Spanish, based on the frequency of lexical errors. They proposed a simple yet reliable method for evaluating the lexical quality of web documents, which can be applied to large data collections. A major finding was that authoritative websites have several orders of magnitude fewer misspellings than the Web overall. The results also included a geographical distribution analysis of lexical quality throughout English and Spanish speaking countries.

The second paper, "Measuring the Quality of Web Content Using Factual Information", was authored by Elisabeth Lex, Michael Voelske, Marcelo Errecalde, Edgardo Ferretti, Leticia Cagnina, Christopher Horn, Benno Stein, and Michael Granitzerg. It presented a statistical quality measure based on facts extracted from online content using Open Information Extraction. The measure was applied on Wikipedia for identifying featured/good articles and turned out to be about as efficient on average as a simple word count measure. Yet, for articles of similar lengths the word count measure is ineffective, while the proposed factual-density-based measure can still estimate article quality.

The next presentation, "A Breakdown of Quality Flaws in Wikipedia" by Maik Anderka and Benno Stein, provided a breakdown of Wikipedias quality aw structure. In contrast with prior research, the authors aimed to produce an extensive classification of quality flaws. Through exploratory analysis, they found the types and distribution of quality aws that exist in Wikipedia and quantified the extent of awed content. Two important nding were that over 25% of English Wikipedia articles contain at least one quality aw and that 70% of the flaws concern article veriability.

The presentation of the paper "A Deformation Analysis Method for Artificial Maps Based

on Geographical Accuracy and Its Applications" by Dasiuke Kitayama and Kazutoshi Sumiya concluded the first session. The paper proposed a prototype system for editing and navigating artificial digital maps. The objective of the work was to support users in estimating the credibility of maps found on the Web by considering the purpose of the maps. The authors proposed a deformation-analysis method based on geographical accuracy using optical character recognition techniques and comparison with gazetteer information.

2.2 Online Credibility and Trust Session

The second session commenced with the presentation of the paper "Game-theoretic Models of Web Credibility" by Thanasis Papaioannou, Katarzyna Abramczuk, Paulina Adamska, Adam Wierzbicki, and Karl Aberer. The authors approached the problem of credibility issues on the Web by modeling it as a game with two types of players: content producers and content consumers. Content producers want to communicate information, and content consumers want to receive credible information. In the model that was presented, content consumers can invest into generating signals for credibility that are independent of the actual veracity of their content (e.g., improving the visual aspect of their website). The authors described formally this game and the conditions for reaching equilibrium, and presented simulation results that yield insights into what conditions lead to certain outcomes.

The second paper of the session, "An Information Theoretic Approach to Sentimental Polarity Classification" by Yuming Lin, Jingwei Zhang, Wang Xiaoling, and Aoying Zhou described a method for sentiment analysis based on the careful weighting of the terms in documents. The weighting is done by measuring the mutual information between the terms and the labels on the training set. The weight of a term in a document also depends on the importance of the term for that document. The authors performed extensive experiments over a dataset of user reviews, where comments attached to reviews with low (high) scores were considered as having negative (positive) sentiment. These experiments indicated a significant improvement over standard state-of-the-art methods.

The third paper, authored by Dvid Siklsi, Blint Darczy, and Andrs A. Benczr, was entitled "Content-Based Trust and Bias Classification via Biclustering". This paper deepens previous work by the authors on issues that have been shown to be difficult to model statistically: neutrality, trust, and bias. The collection used for reference was the ECML/PKDD Discovery Challenge 2010 collection, in which, for instance, for the "bias" label the best system achieved only 0.558 AUC. The approach presented in their paper starts by performing a bi-clustering of the host-term matrix in order to represent each host in a space of clusters of terms, corresponding loosely to concepts. A distance function between such clusters is also described in the paper. Then, by the application of kernel methods and a fusion with pure text-based classifiers, the authors produced a classifier that performs significantly better than the earlier state of the art, e.g., 0.685 AUC for the same "bias" label.

2.3 Abuse Detection and Prevention Session

The last session started with the presentation of the paper "Detecting Collective Attention Spam" by Kyumin Lee, James Caverlee, Krishna Kamath, and Zhiyuan Cheng. The focus of the paper is a novel and ingenious form of spam, targeting social media where collective user attention coalesces and focuses quickly on emerging phenomena (e.g., trending topics on Twitter, breaking videos on YouTube, and popular profiles on Facebook). Successful spam

in these media is by definition short lived, and its detection requires fast-learning classifiers. Through the careful examination of a manually labelled Twitter dataset, the authors identified a viable feature set that can power classifiers enabling the automatic detection of such novel spam.

In the second paper of the session, "Identifying Spam in the iOS App Store", authors Rishi Chandy and Haijie Gu draw attention to the potential negative impact of fraudulent apps and reviews in the multimillion-dollar ecosystem of mobile phone apps. They assembled two new datasets by crawling the reviews submitted to Apples iOS App Store, performed manual labelling, and proposed a promising latent class model, tested in supervised and unsupervised learning experiments.

The final paper of the session, "kaPoW Plugins: Protecting Web Applications Using Reputation-based Proof-of-Work" by Akshay Dua, Wu-Chang Feng, and Tien Le, presented a new tool in the toolbox for fighting online comment spam: a browser plugin that requires client machines to solve computational puzzles as proofs-of-work on comment submission. Requiring no user intervention, such puzzles effectively slow down spammers attempting bulk, automated submissions, making review spamming less economical.

3 Acknowledgements

We would like to thank the organizers of the WWW 2012 conference for helping to organize our workshop. We also express our gratitude to the program committee members for their dedicated work and to the participants for their contribution to the workshop's success.

For more information about the workshop, please visit `http://www.dl.kuis.kyoto-u.ac.jp/webquality2012/`.