Report on the WebQuality 2014 Workshop

Adam Jatowt[1,2], Carlos Castillo[3], James Caverlee[4] and Katsumi Tanaka[1]

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

[2] Japan Science and
Technology Agency
4-1-8, Honcho, Kawaguchi-
shi, Saitama 332-0012
Tokyo, Japan

[3] Qatar Computing Research
Institute,
Doha, Qatar
chato@acm.org

[4] Texas A&M University
403 H.R. Bright Building
College Station, TX
77843-3112
caverlee@cse.tamu.edu

Abstract

The 4th Joint WICOW/AIRWeb Workshop on Web Quality[i] (WebQuality 2014) was held in conjunction with the 23rd International World Wide Web Conference in Seoul, South Korea on the 7th April 2014. This report briefly summarizes the workshop.

1 Introduction

WebQuality 2014 was held on 7th April 2014 as the 4th 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 joint workshops held at WWW 2011, WWW 2012 and WWW 2013 conferences attracted papers related to more general topics of web content quality and approaches to its measurement, improvement and analysis. This report summarizes the fourth joint workshop held in conjunction with the WWW 2014 conference.

Web content quality research is as relevant as ever and of increasing role especially in the context of social media that are growing both in size and complexity. The workshop covers 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. At the same time, the workshop also provides a venue for exchanging ideas on quantifying finer-grained issues of content credibility and author reputation, and modeling them in web information retrieval.

We were also pleased to invite Wendy Hall from University of Southampton for giving a keynote talk titled "***The Web Science of Web Quality***".

2 Paper Presentations

This year we have selected 5 full research papers based on the feedback from the program committee to be published in the ACM Digital Library and presented at the workshop. Each submission was reviewed by at least three program committee members. The accepted papers were presented in two sessions: *"Text Content Quality"* and *"Multimedia Content Quality"*. We briefly summarize the papers below.

2.1 Text Content Quality Session

In the first paper in this session, entitled *"**Learning Conflict Resolution Strategies for Cross-Language Wikipedia Data Fusion**" Volha Bryl* and *Christian Bizer* propose an approach to automatically learn the conflict resolution strategies in large-scale data integration. In many applications, Web data needs to be integrated in a way in which data quality and consistency aspects are taken care of. The authors demonstrate solution for resolving data conflicts during data fusion process which allows automatically learning an optimal combination of fusion functions for a set of data properties. In particular, they work with DBpedia[1] performing fusion of data about populated places such as cities and towns across 10 different language editions of Wikipedia. The evaluation shows that automatically learning the conflict resolution strategies leads to accuracy improvements of the integrated dataset. The authors also show a method for extracting rich provenance metadata for each DBpedia fact.

The second paper, *"**Incredible: is (almost) all web content trustworthy? analysis of psychological factors related to website credibility evaluation**",* was authored by *Maria Rafalak, Katarzyna Abramczuk* and *Adam Wierzbicki*. In this paper the authors find evidence for "positive bias" or "truth bias," which is a tendency to overestimate the credibility of information. About 2,000 participants hired through crowdsourcing provider Mechanical Turk were asked to indicate the credibility of a set of websites on a scale of 1 to 5. User-provided ratings were compared against the WOT index[2]. When considering psychological traits such as general attitudes to trust and risk-taking propensity, it appears that people who tend to overestimate more the credibility of websites are those who tend in general to trust more (which is expected), but also those who tend in general to take less risk (which is a bit counterintuitive). The interpretation of this finding is that "credible" is seen as a default, safe option by Internet users, who need reasons or evidence in order not to trust an online website.

The presentation of the paper *"**Predicting webpage credibility using linguistic features**"* by *Aleksander Wawer, Radoslaw Nielek* and *Adam Wierzbicki* concluded the first session. In this paper the authors improve supervised learning methods to infer the trustworthiness of content from text. A key source for this improvement is the application of the General Inquirer lexicon[3]. The supervision comes from a dataset of 1,000 labeled web pages in 3 levels of credibility: low, medium, and high. Interestingly, the authors also present results regarding which words are more associated to trustworthy or non-trustworthy content. People tend to trust less documents containing terms referring to financial services, in particular with respect to borrowing money (this could indicate they are savvy with respect to Internet scams on this topic). Documents that were deemed more trustworthy contain references to government and safety.

2.2 Multimedia Content Quality Session

The second session commenced with the presentation of the paper *"**Identifying Fraudulently Promoted Online Videos**"* by *Vlad Bulakh, Minaxi Gupta* and *Christopher Dunn*. The authors collect a sample of over 3,300 fraudulently promoted videos and 500 bot profiles that promote them. They analyze the fraudulent videos and profiles to identify their characteristic features. For example, an average fraud video is found to have shorter and fewer comments and is rated higher, while profiles which promote the fraudulent videos are more active in viewing and interacting with videos and

[1] http://dbpedia.org/About
[2] http://www.mywot.com/
[3] http://www.wjh.harvard.edu/~inquirer/

rarely upload any videos. Another contribution of the paper is creation of supervised machine learning classifier to differentiate between fraudulent videos and profiles, and the legitimate ones.

The last paper, authored by *Lei Li* and *Chengzhi Zhang*, is entitled ***"Quality Evaluation of Social Tags according to Web Resource Types"***. This paper looks into quality characteristics of tags used to annotate different kinds of web resources. The authors emphasize that tag quality should be measured in order to provide efficient tag recommendation services and to design better user tagging interfaces. They then selected web resources of five different types, blog, book, image, music and video, and explored the tag types used for annotating the different resources. Next, they explored the quality of each tag type and studied the relationship between tag type and quality. The conclusion of the paper is that the quality of each tag type is different according to the web resource.

3 Programme Committee

The following researchers and industry experts have served on the Programme Committee of WebQuality 2014:

Ching-man Au Yeung (Huawei Noah's Ark Lab)
Andras Benczur (Hungarian Academy of Sciences)
Brian Davison (Lehigh University)
Andrew Flanagin (University of California, Santa Barbara)
Zoltan Gyongyi (Google Research)
Kyumin Lee (Utah State University)
Pranam Kolari (Walmart Labs)
Miriam Metzger (University of California, Santa Barbara)
Meenali Rungta (Google)
Shazia Sadiq (University of Queensland)
David Siklosi (Hungarian Academy of Sciences)
Mozhgan Tavakolifard (Norwegian University of Science and Technology)
De Wang (Georgia Institute of Technology)
Steve Webb (Georgia Institute of Technology)

4 Acknowledgements

We would like to thank the organizers of the WWW 2014 conference for helping to organize our workshop. We also express our gratitude to all the program committee members for their dedicated work and to the participants for their contribution to the workshop's success.

[i] http://www.dl.kuis.kyoto-u.ac.jp/webquality2014/