Report on the WebQuality 2013 Workshop

Adam Jatowt[1,2], Carlos Castillo[3], Zoltan Gyongyi[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] Google Research
1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA
zoltang@google.com

Abstract

The 3rd Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2013) was held in conjunction with the 22nd International World Wide Web Conference in Rio de Janeiro, Brasil on the 13th May 2013. A keynote talk and four full and four short paper presentations were delivered in four sessions. This report briefly summarizes the workshop.

1 Introduction

WebQuality 2013 was held on 13th May 2013 as the 3rd 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 previous two joint workshops were held at WWW 2011 and WWW 2012 and attracted papers related to more general topics of web content quality and approaches to its measurement, improvement and analysis. This report summarizes the third joint workshop held in conjunction with the WWW 2013 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. WebQuality 2013 addressed a mix of quality issues, ranging from more blatant forms of abuse, such as web spam and malware distribution, to the finer grained topics of content credibility and author trustworthiness. The academic research contributions were complimented by a strong portfolio of industry experience reports. We were also pleased to have Ricardo Baeza-Yates from Yahoo! Research deliver a keynote talk titled "Measuring Web Quality".

2 Paper Presentations

This year we invited two types of submissions: research papers and practice & experience reports. From among 16 submissions, the workshop program committee selected 5 research articles and 3 practice & experience reports, representing 4 full and 4 short papers, both to be presented at the workshop and to be published in the ACM Digital Library as part of the Proceedings of the 22nd International Conference on World Wide Web Companion. Each submission was reviewed by at least three program committee members. The accepted papers were presented in three sessions: *"Web Content Quality"*, *"Industry Experience"*, and *"Web Spam Detection"*. Below we briefly summarize each session.

2.1 Web Content Quality Session

The first paper in this session, *"*Defending Imitating Attacks in Web Credibility Evaluation Systems*"* was authored by Xin Liu, Radoslaw Nielek, Adam Wierzbicki and Karl Aberer. The authors focus on a novel, particular type of attack on Web credibility evaluation systems, in which an attacker imitates the behavior of trustworthy experts by copying the system's credibility ratings to make she/he look credible and then to attack certain web contents. Their approach to combat such attacks relies on a two-stage process. In the first stage, they employ supervised learning algorithms to estimate the credibility of the target web content for detecting attackers. The second stage improves the decisions in low-confidence cases by analyzing users' past rating patterns through the application of a hierarchical clustering algorithm.

The second paper, *"*Trustworthiness Criteria for Supporting Users to Assess the Credibility of Web Information*",* was authored by Jarutas Pattanaphanchai, Kieron O'Hara and Wendy Hall. It looked into the factors that are important for helping to evaluate the trustworthiness of information based on the results of a survey administered to a panel of experts who had experience in assessing the credibility or the quality of web information. The analysis suggests that 10 pieces of information (including, for instance, author affiliation, references, and citations) are particularly useful for assessing trustworthiness. At the same time, it is common that users need to search for additional supporting information to assess trustworthiness when some of the 10 pieces are missing. The authors also investigate weighting factors for diverse evaluation criteria that can be used to support credibility assessment.

The presentation of the paper *"On the Subjectivity and Bias of Web Content Credibility Evaluations"* by Michal Kakol, Michal Jankowski-Lorek, Katarzyna Abramczuk, Adam Wierzbicki and Michele Catasta concluded the first session. The authors studied the influence of subjectivity and bias in the credibility ratings on the data coming from 1503 respondents who independently evaluated credibility of 154 web pages divided into several thematic categories. The study was supplemented by information about socioeconomic status and psychological features of evaluators. According to their findings, web content credibility evaluations are slightly subjective and exhibit a strong acquiescence bias.

2.2 Industry Experience Session

The industry experience session commenced with the presentation of the paper *"*Russian Web Spam Evolution: Yandex Experience*"* by Sergey Pevtsov and Sergey Volkov. This paper described current trends in web spamming activity and how these trends reflect changes in the anti-spam methods applied by search engines. For instance, domain name spam in the Yandex search engine went from a period of aggressive growth to one of stagnation after improvements to their anti-spam algorithms went live. On the other hand, Javascript tricks aiming to show innocent-looking contents to web spam analysts have been used increasingly often in the last few years. Currently, key challenges include overly aggressive advertising (often disguised as useful content), and social engineering, including phishing attempts to steal user credentials and to distribute malware.

The second paper of the session, *"Graph-based Malware Distributors Detection"* by Andrei Venzhega, Polina Zhinalieva and Nikolay Suboch described a method to detect malware (e.g., fake versions of the Firefox web browser that hijack a user's computer). This method is based on the existence of dense sub-graphs in the bipartite website/file-hosting graph: websites distributing malware apparently links to malware payloads that are served by certain file-hosting providers. Malwareness rank uses a propagation algorithm starting from a seed of known file-hosting providers serving malware to detect websites distributing malware.

The last paper, authored by Alexander Shishkin, Polina Zhinalieva and Kirill Nikolaev, was entitled *"*Quality-biased Ranking for Queries with Commercial Intent*"*. This paper studies a practical issue that plagues many search engine results pages: the top results being saturated with websites over-optimized to be relevant to the query. In this scenario, a search engine should not only rank by the relevance of the result for the query but accounting for user satisfaction as well. For queries with commercial intent, users are more satisfied when visiting websites that provide an assortment of goods presented in a visually attractive, well-designed, and useful manner. Some of these factors can be machine learned and inferred for large datasets. The results in their paper show measurable increases in user satisfaction with the results (quantified by the fraction of "long" clicks: clicks followed by a significant amount of time the user spent on the destination page) and a measurable decrease in the fraction of users that abandon a query.

2.3 Web Spam Detection Session

The last session started with the presentation of the paper *"*Cross-Lingual Web Spam Classification*"* by Andras Garzo, Balint Daroczy, Tamas Kiss, David Siklosi and Andras A. Benczur. The presented research focused on the challenges of web spam detection in web content locales and languages where spam training and evaluation data labeled by human raters is not readily available. The authors found that some well-known spam detection signals, especially link structure signals, transcend language boundaries. They also experimented with a variety of bag-of-words signals (filtering out non-English terms, or applying term-by-term translation, or a hybrid approach) training classifiers on a labeled Portuguese spam corpus. The experiments indicated that such bag-of-words signals combine well and that the role of multilingual sites in cross-lingual spam detection is significant.

In the final paper of this session, *"*Automatically Generated Spam Detection Based on Sentence-level Topic Information*"*, authors Yoshihiko Suhara, Hiroyuki Toda, Shuichi Nishioka and Seiji Susaki tackled content spam created through stitching together phrases from legitimate web sources. The authors identified frequent topic transitions across consecutive sentences as a telltale sign of such spam. Consequently, they applied a topic-voting heuristic with conventional LDA to achieve sentence-level topic assignments and constructed spam classification signals based on measures of topical drift across sentences. The paper describes a set of preliminary experiments performed on a Japanese blog corpus that highlight the potential of sentence-level topic signals in spam classification.

3 Acknowledgements

We would like to thank the organizers of the WWW 2013 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.

For more information about the workshop, please visit `http://www.dl.kuis.kyoto-u.ac.jp/webquality2013/`.