Report on the WebQuality 2015 Workshop

Radoslaw Nielek[1], Adam Wierzbicki[1], Adam Jatowt[2] and Katsumi Tanaka[2]

[1]Polish-Japanese
Academy of Information Technology
86 Koszykowa Street, 02-008 Warsaw, Poland
{nielek, adamw}@pja.edu.pl

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

Abstract

The 5th International Workshop on Web Quality (WebQuality 2015) was held in conjunction with the 24rd International World Wide Web Conference in Florence, Italy on the 18th May 2015. This report briefly summarizes the workshop.

1 Introduction

WebQuality 2015 was held on 18th May 2015 as the 5th International Workshop on Web Quality. The workshop has addressed information credibility and web quality issues and has been held since 2011 in co-location with WWW conferences (WWW 2011, WWW 2012, WWW 2013 and WWW 2014). This workshop series attracts papers related to uncovering distorted and biased content, measuring quality of web content and modeling author identity, trust, and reputation. In particular, in 2015 the majority of accepted submissions were focused on web spam detection and quality as well as the credibility of the web sites. This report summarizes the fifth workshop held in conjunction with the WWW 2015 conference.

Plethora of web sites and development of search engines results in that finding relevant information is often quite an easy task. The problem is that relevant information do not preclude attempts of manipulation or wrongdoings by the web site owner. The workshop covers 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 serves as a venue for exchanging ideas between information retrieval researchers and those focused on information quality.

We were also pleased to invite Cinzia Cappiello from Politecnico di Milano for giving a keynote talk titled "***On the Role of Data Quality in Improving Web Information Value***".

2 Keynote

The keynote talk given by Cinzia Cappiello from Politecnico di Milano titled "On the Role of Data Quality in Improving Web Information Value" was a brilliant attempt to combine formal approach excerpted from data quality research with more statistical, “best effort” oriented methods used in web mining. The first part of the keynote was devoted to discussion about possible definition of data quality in light of different types of data sources. The participants agreed that it seems unrealistic to expect a

single precise definition of data quality and the set of estimation metrics even just for the web sites. During the second part of the keynote more focus has been put on assuring data quality by combining data and information from many sources.

3 Paper Presentations

This year we have selected 4 full research papers from 12 papers which have been submitted based on the feedback from the program committee to be published in the ACM Digital Library and to be 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.

3.1 *Identification of Web Spam through Clustering of Website Structures*, Filippo Geraci,

In the paper entitled: "Identification of Web Spam through Clustering of Website Structures" Filippo Geraci studies methods for detecting spam websites defined as domains which are parked for future resale and which contain a large number of ads. Such web sites can deteriorate performance of web crawlers or other web mining tools as well as they typically introduce extra cost for ad bidders. The author demonstrates that spam websites managed by the same providers tend to have similar look-and-feel and defines an efficient metric for capturing structural similarity of web pages. Based on the metric, the Furthest Point First clustering is used for identifying spam websites. Finally, a manually curated dataset of spam websites is released for research purposes.

3.2 *Answer Quality Characteristics and Prediction on an Academic Q&A Site: A Case Study on ResearchGate*, Lei Li, Daqing He, Wei Jeng, Spencer Goodwin, Chengzhi Zhang,

Lei Li, Daqing He, Wei Jeng, Spencer Goodwin, Chengzhi Zhang in their paper: "Answer Quality Characteristics and Prediction on an Academic Q&A Site: A Case Study on ResearchGate" analyze factors determining the quality of answers in specialized QA sites. They investigate, in particular, an academic social network site ResearchGate extracting various Web-based features such as answer length or internal quality scores as well as human-coded features including the appearance of social elements or pointers to external resources. Using the obtained features they test three classification models to predict answer ratings and they discover that optimized SVM classifiers perform best, especially, when equipped with Web-based features. Another conclusion from their study is that ResearchGate has significant differences from standard generic QA sites in terms of factors contributing to answer quality.

3.3 *Text Classification Kernels for Quality Prediction over the C3 Data Set*, Balint Daroczy, David Siklois, Robert Palovics, and Andras A. Benczur,

In the paper: "Text Classification Kernels for Quality Prediction over the C3 Data Set" authors use the C3 dataset collected as a part of the Reconcile project to study the performance of algorithms for automatically assessing content credibility. The C3 dataset published on the WebQuality workshop web site[1] contains 22,325 evaluations (five dimensions: credibility, presentation, knowledge, completeness and intentions) of 5,704 pages given by 2,499 people and has been built with help of the mTurk[2] platform. Authors tested a variety of methods including: gradient boosted trees, biclustering based standard text classifiers and similarity kernel based on SVM on the Fisher Information Matrix. Although all the applied methods have a rich record of successful application to a variety of domains, actually, none of them was found to work sufficiently well on the C3 dataset. The best results have been obtained

[1] http://webquality.org/data-challenger/
[2] https://www.mturk.com/mturk/welcome

with the help of kernel method, yet, still AUC was only slightly over 0.7 for credibility dimension and 0.8 for presentation dimension.

3.4 *Characterizing Credit Card Black Markets on the Web*, Vlad Bulakh, and Minaxi Gupta,

The paper "Characterizing Credit Card Black Markets on the Web" authored by Vlad Bulakh and Minaxi Gupta presented thorough analysis of web sites focused on selling stolen credit and debit card information. Authors identified three web sites offering stolen credit cards details. Collecting data from these web sites required to overcome anti-scrapping mechanisms and to follow frequent domain name changes. At the end, a dataset containing almost 100 thousand unique credit cards has been created. Careful sanitization and analysis of the collected data revealed a lot of information about economy of this part of black market. Obvious misspellings in credit card information published on these web sites indicate that they are probably manual component within the entire workflow. It turns out that the prices of cards depend heavily on factors like the issuing bank or country of origin. Moreover, almost 70% of cards are priced below the cost that bank incurs in re-issuing them. It might be then the reason why the cards are still valid. The estimated gross revenue for single such web site varies from $181,047 to $276,783 per month.

4 Programme Committee

The following researchers and industry experts have served on the Programme Committee of WebQuality 2015:

Dávid Siklósi, Computer and Automation Research Institute, Hungarian Academy of Sciences,
Carlos Castillo, Qatar Computing Research Institute,
Kyumin Lee, Utah State University,
Andrew Flanagin, UCSB,
Luca Maria Aiello, Yahoo Labs, Barcelona
Andras A. Benczur, Institute for Computer Science and Control, Hungarian Academy of Sciences
Aleksander Wawer, Institute of Computer Science Polish Academy of Science, Warsaw
Liu Xin, Institute for Infocomm Research, Singapore,
Miriam Metzger, UC Santa Barbara, USA
De Wang, Georgia Institute of Technology,
Minaxi Gupta, Indiana University,

5 Acknowledgements

We would like to thank the organizers of the WWW 2015 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.