# The 2nd Workshop on Information Credibility on the Web (WICOW 2008)

**Adam Jatowt and Katsumi Tanaka**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, Japan
*{adam, tanaka}@dl.kuis.kyoto-u.ac.jp*
*Workshop Website: http://www.dl.kuis.kyoto-u.ac.jp/wicow2/*

**Abstract**

Research on credibility of web content is becoming increasingly important due to low publishing barriers and resulting abundance of untrustworthy or conflicting information on the web. On the 30th October 2008 the 2nd Workshop on Information Credibility on the web was held as part of CIKM 2009 conference in Napa Valley, USA. Nine full and six short papers were accepted and grouped into four sessions. In addition, two keynote speeches have been delivered. This report outlines the main results of the workshop.

## 1 Introduction

The web has started to play an important role for many people in delivering information related to their personal and professional lives. However, due to low publishing barriers there is usually inadequate quality control of web content. In result, a lot of mistaken or unreliable information appears on the web that can has detrimental effects on users. For example, web pages may provide inaccurate or incomplete information, their content can be obsolete or authors' opinions may be biased. The problem of information credibility is not only limited to textual content but also to other multimedia published on pages. This calls for methodologies that would facilitate judging the trustworthiness of content that users encounter on the web. Related issue is the analysis of user behavior and, in particular, the factors to which users pay attention when evaluating the credibility of information. In general, the information credibility is a complex, multi-dimensional issue and the effective methods for judging content's trustworthiness should combine technologies from diverse research areas such as information retrieval and extraction, information processing, web search, knowledge representation, etc.

The 2nd Workshop on the Information Credibility on the Web was held on October 30th in Napa Valley, California, USA. It was organized in conjunction with ACM 17th Conference on Information and Knowledge Management[i] and was attended by more than 40 participants. The aim of this workshop was to provide a platform for exchanging novel ideas and research findings as well as to promote discussions on various aspects of information credibility among both researchers and practitioners.

The submissions concerned with the following topics were encouraged:

- information credibility evaluation and its applications
- content analysis for credibility evaluation
- sentiment analysis of content
- intent and deception detection
- credibility of web search results
- search models and applications for trustworthy content
- conflicting opinion detection and analysis
- media and news credibility
- credibility evaluation of user-generated content (e.g., Wikipedia)
- information credibility evaluation in social networks
- analysis of information dissemination
- estimation of author and publishing venue reputation
- spatial and temporal aspects in information credibility
- estimation of information age, provenance and validity
- sociological and psychological aspects of information credibility
- users study for information credibility
- risk assessment of information credibility
- multimedia content credibility
- persuasive technologies
- information credibility in online advertising and Internet monetization
- object identification on the web
- spam detection

After a careful review process, with at least three reviews for each paper, the Program Committee has selected 9 full and 6 short papers covering a variety of topics related to information credibility. The accepted papers were thematically grouped into four sessions: Analyzing social networks and discussion forums, Web content analysis, Content aggregation on the web and Information quality: frameworks and theories. In addition, two keynote speakers were invited. The first keynote talk entitled: *Graph Mining and Influence Propagation* was delivered by *Christos Faloutsos* from Carnegie Mellon University, USA. *Yutaka Kidawara* from National Institute of Information and Communications Technology, Japan gave the second keynote talk entitled: *Information Credibility Analysis of Web Content*.

Furthermore, based on the review scores two best paper awards have been given to the following papers:

- *PodCred: A Framework for Assessing Podcast Credibility* by *M. Tsagkias*, *M. Larson*, *W. Weerkamp* and *M. D. Rijke*
- *Detecting Reviewer Bias through Web-Based Association Mining* by *J. Staddon* and *R. Chow*

The next edition of the workshop (WICOW 2009[ii]) is going to be held on April 20th, 2009 in Madrid, Spain in conjunction with the 18th International World Wide Web Conference[iii].

## 2 Paper Presentations

This section briefly introduces workshop papers arranged according to their corresponding sessions.

### 2.1 Analyzing Social Networks and Discussion Forums

The first session started with a keynote presentation: *Graph Mining and Influence Propagation* by *C. Faloutsos*. In his presentation, *Faloutsos* introduced several types of graphs observed in real networks and proposed a "Kronecker" graph generator that matches all of the known properties of real graphs.

The presentation also included some case studies such as influence and virus propagation on real graphs and the detection of "non-delivery" fraud types in eBay interaction graphs.

The full paper entitled: *Detecting Reviewer Bias through Web-Based Association Mining* by *J. Staddon* and *R. Chow* proposes evaluating the credibility of online reviews by analyzing the relationships between book authors and book reviewers. The potential, undisclosed relationships can be detected through mining association rules involving the occurrences of the names of authors and book reviewers on the web. The preliminary results on data from amazon.com demonstrate the high effectiveness of the proposed approach in disclosing strong relationships between reviewers and authors that may contribute to reviewer bias.

In the next full paper entitled: *Automatic Scoring of Online Discussion Posts* by *N. Wanas, M. El-Saban, H. Ashour* and *W. Ammar* introduce a method of automatically rating posts in online discussion forums based on their quality by using non-linear SVM. Posts are characterized by 22 features grouped into five categories: relevance, originality, forum-specific features, surface features and posting-component features. Based on these features the authors categorize the values of posts into high, medium or low. The experiments on about 20,000 rated posts show nearly 50% of categorization accuracy and indicate higher importance of structural features of posts compared to the features utilizing text analysis.

The full paper entitled: *Reasonable Tag-Based Collaborative Filtering* by *R. Nakamoto*, *S. Nakajima*, *J. Miyazaki*, *S. Uemura*, *H. Kato* and *Y. Inagaki* presents a system for tag-based collaborative filtering recommendation. The proposed application provides credible web page recommendations that match changing user interests and her or his bookmarking profile in online social tagging systems. The recommendation process is realized through finding users with similar preferences that like given information for the same reasons, and through considering the content of web pages currently viewed by the user.

## 2.2 Web Content Analysis

This session started with the second keynote talk: *Information Credibility Analysis of Web Content* by *Y. Kidawara*. The talk contained description of two large research projects in Japan undertaken by the National Institute of Communications and Information Technology (NICT) and the Ministry of Internal Affairs and Communications (MIC) aiming at credibility analysis of information. *Kidawara* introduced also some of the current initiatives by NICT and other research groups in Japan towards establishing state-of-the-art technologies for evaluating information credibility on the web. The important aspect of these approaches is analyzing information credibility according to the following criteria: content, sender, appearance and authenticity of content.

*R. Lopes* and *L. Carriço* in their full paper: *On the Credibility of Wikipedia: an Accessibility Perspective* investigate the influence of accessibility of referenced web pages in Wikipedia articles on the credibility of these articles. The key idea is that the readers must be able to access referred pages without any kinds of barriers if the original page can be considered credible. The paper proposes also a set of improvements which could help increasing the accessibility of references within Wikipedia articles.
The full paper entitled: *Extracting the Author of Web Pages* by *Y. Kato*, *D. Kawahara*, *K. Inui*, *S. Kurohashi* and *T. Shibata* describes a method of automatically detecting web page authors by SVM. The problem of author extraction is set in the wider context of information sender configuration of web pages. The proposed approach firsts identifies a set of candidates of page authors by linguistic analysis and then ranks the candidates based on local features such as the candidate's distance from the main content of web pages. The experimental evaluation shows precision of about 75% when considering 5 top-ranked candidates.

The last presentation in this session was the one of a short paper entitled: *A "Quick and Dirty" Website Data Quality Indicator* written by *I. A. Gelman* and *A. Barletta.* This paper advocates the idea of approximately evaluating page quality by analyzing the spelling error rate of page content. The error rate is derived using the reported hit counts obtained for search engine queries that are formed using commonly misspelled words. Initial results indicate a strong correlation between spelling errors and content quality of web pages.

## 2.3 Content Aggregation on the Web

The presentation of the full paper entitled: *ALPACA: A Lightweight Platform for Analyzing Claim Acceptability* by *J. King*, *J. Stoll*, *M. Hunter* and *M. Ahamad* opened the third session of the workshop. The authors first discuss the shortcomings of current systems for evaluating information credibility on the Internet. They then propose a set of design principles and introduce a system called ALPACA for examining credibility for claims based on these principles. This application organizes and presents claims and references through a graphical interface and allows manipulating complex information relationships for examining claim credibility.

The full paper entitled: *Using a Sentiment Map for Visualizing Credibility of News Sites on the Web* by *Y. Kawai*, *Y. Fujita*, *T. Kumamoto*, *J. Zhang* and *K. Tanaka* describes a system for assessing sentiment trends of news sites for the purpose of estimating news article credibility. The trend of a news site regarding a certain topic is calculated as an average sentiment of news articles published by a given news site. The sentiment values of news articles are determined using pre-constructed sentiment dictionary. The authors developed a map-based application for visualizing sentiment trends of news sites in different geographic locations for user queries.

*S. Nakamura*, *M. Shimizu* and *K. Tanaka* authored a short paper: *Can Social Annotation Support Users in Evaluating the Trustworthiness of Video Clips?* The authors propose a system for supporting trustworthiness evaluation of video clips on video sharing sites. The application visualizes the rate of negative and positive comments added by users arranged chronologically according to the posting date or the playback time of videos. It has a potential to assist users with determining the parts of videos that may lack credibility as evidenced by the changes in the aggregated sentiment value of posted user comments.

The last short paper of this session: *Web-based Evidence Excavation for Exploring Authentic Local Events* by *R. Lee*, *D. Kitayama* and *K. Sumiya* describes a method for extracting information about real-world events from the web and providing their reliable evidences. Events representations are mined from the web and clustered in order to discover their spatio-temporal characteristics as well as estimate their credibility. The authors describe also a similarity measure between events in order to perform clustering and searching.

## 2.4 Information Quality: Frameworks and Theories

The full paper entitled: *PodCred: A Framework for Assessing Podcast Credibility* by *M. Tsagkias*, *M. Larson*, *W. Weerkamp* and *M. D. Rijke* describes a framework for assessing credibility and quality of podcasts published on the Internet. The framework prescribes a series of indicators arranged into four groups pertaining to: podcast content, podcaster, podcast context and technical execution of the podcast. They are derived from a literature review on credibility, survey of prescriptive standards for podcasting and the analysis of award winning podcasts.

The full paper entitled: *A Metacognitive Approach to Credibility Determination* by *M. Iding*, *B. Auernheimer* and *M. E. Crosby* reports on the studies of user credibility judgements from an educational perspective. The studies have been conducted in different contexts including Hawaii, Norway and Singapore. Based on the results several user effective credibility determinants of web sites could be discovered. One outcome of this work is the generation of credibility-focused instructional recommendations from educational perspective.

The short paper: *Improving Information Quality in Email Exchanges by Identifying Entities and Related Objects* by *M. Kowalkiewicz* and *K. Juenemann* addresses the problem of emails' content quality evaluation. The authors propose to utilize techniques borrowed from information extraction and service oriented architecture fields for enriching emails with additional information on objects mentioned in emails. A proof-of-concept application designed to improve quality of email exchanges is presented in the paper.

*J. G. Conrad*, *J. Leidner* and *F. Schilder* in a short paper entitled: *Professional Credibility: Authority on the Web* investigate the problem of measuring the authority degree of authors in legal blogosphere. They report on the results of an experiment in which annotators examined the entries and comments to legal blogs from the viewpoint of authority. Based on the experimental results it has been then possible to identify several important conclusions and research questions involved in the quantification of human judgement of authority.

The short paper: *A Characteristic Based Information Evaluation Model* by *L. Tang*, *Y. Zhao*, *S. Austin*, *M. Darlington* and *S. Culley* raises a question whether it is possible to establish metrics to theoretically assess the credibility and value of information within the context of engineering design. The authors introduce Information Evaluation Model for assigning value to information in the area of global engineering and construction companies. The model is based on a field work and a review of prior works in the field of information value and quality estimation.

# 3 Acknowledgements

We convey our sincere thanks to the other organizers of WICOW 2008, Takashi Matsuyama and Ee-Peng Lim. We would like to thank CIKM 2008 organizers for letting us to organize this workshop in conjunction with CIKM 2008 conference and providing us with great support during the organization process. We also wish to express our gratitude to the sponsors of the workshop: National Institute of Information and Communications Technology and Kyoto University Global COE Program: Informatics Education and Research for Knowledge-Circulating Society.

[i] The ACM 17th Conference on Information and Knowledge Management (CIKM 2008): http://www.cikm2008.org/
[ii] The 3rd Workshop on Information Credibility on the Web (WICOW 2009): http://www.dl.kuis.kyoto-u.ac.jp/wicow3/
[iii] The 18th International World Wide Web Conference (WWW 2009): http://www2009.org/