Preface to the special issue on web information systems engineering (WISE 2018)

Hakim Hacid[1] · Wojciech Cellary[2] · Hua Wang[3] · Yanchun Zhang[3]

This special issue is dedicated to the best selected papers from WISE 2018. The main topic of the conference and this special issue is Web Engineering in the Era of Big Data and Artificial Intelligence (AI). This special issue focuses on selected problems in Web Engineering, a multidimensional area that involves big data, security, search algorithms, Internet of things, and data mining. Seven top ranked papers out of 48 full papers at WISE 2018 have been selected for the special issue of World Wide Web Journal (WWWJ). The selected papers underwent a rigorous extra refereeing and revision process. In particular, the issues papers have been extended with at least 30% new and unpublished material. Note that, adding more related work or extending the introduction was not considered in the 30%; rather the new content often includes more technical and implementation details, improved algorithms, more experiment results, etc.

A total of 209 research papers were submitted to the conference for consideration, and each paper was reviewed by at least three reviewers. Finally, 48 submissions were selected as regular papers (with an acceptance rate of 23% approximately), plus 21 as short papers. The research papers cover the areas of blockchain, security and privacy, social networks, microblog data analysis, graph data, information extraction, text mining, recommender systems, medical

---

This article belongs to the Topical Collection: *Special Issue on Web Information Systems Engineering 2018*
Guest Editor: Hakim Hacid, Wojciech Cellary, Hua Wang and Yanchun Zhang

✉ Hakim Hacid
Hakim.Hacid@zu.ac.ae

Wojciech Cellary
Cellary@kti.ue.poznan.pl

Hua Wang
Hua.Wang@vu.edu.au

Yanchun Zhang
Yanchun.Zhang@vu.edu.au

[1] College of Technological Innovation, Zayed University, Dubai, United Arab Emirates

[2] Department of Information Technology, Poznan University of Economics and Business, Poznań, Poland

[3] Institute for Sustainable Industries and Liveable Cities, Victoria University, Footscray, Australia

data analysis, web services, cloud computing, data stream, distributed computing, data mining techniques, entity linkage and semantics, web applications, data mining applications.

In addition to regular and short papers, WISE 2018 program also featured four workshops: (1) The 5th WISE Workshop on data quality and trust in big data (QUAT'18); (2) International workshop on edge-based computing for next generation wireless networks; (3) The 3rd international workshop on information security and privacy for mobile cloud computing, web and internet of things (ISCW'18); (4) The 1st international workshop on cloud computing economic impacts. This year's tutorial program included: (1) Text Mining for Social Media; (2) Towards privacy-preserving identity and access management systems for web developers and (3) From Data Lakes to Knowledge Lakes: The Age of Big Data Analytics.

Data is becoming large in context and diverse in type and nature. There have been a lot of efforts spent to analyze and automatically understand text data in the literature but more needs yet to be done. The paper by Wu et al. discusses text classification and focuses on capturing contextual information when using convolutional networks for text classification. To capture the contextual information, the authors propose to use the weighted sum operation to obtain contextual word representation. An implicit weighting method and two explicit category-aware weighting methods to assign the weights of the contextual information are proposed in the paper. Finally, experimental results on five text classification datasets show the effectiveness of the proposed methods.

Internet of Things (IoT) is gaining interest from the research community due to the opportunities it opens for, e.g., improving the life quality of people. He et al., propose a framework for cardiac arrhythmia detection from IoT-based ECGs. The paper proposes two solutions for the heartbeat classification task, namely: (i) Dynamic Heartbeat Classification with Adjusted Features (DHCAF) and (ii) Multi-channel Heartbeat Convolution Neural Network (MCHCNN). DHCAF is a feature-engineering based approach, in which a dynamic ensemble selection (DES) technique is introduced and develop a result regulator to improve classification performance is developed. MCHCNN is deep learning-based solution that performs multi-channel convolutions to capture both temporal and frequency patterns from heartbeat to assist the classification. The proposed framework has been evaluated with DHCAF and with MCHCNN on the well-known MIT-BIH-AR database, respectively. The results reported in this paper have proven the effectiveness of the proposed framework.

As an important source of social data, Twitter has been used by different research efforts to contribute to a better understanding of the social behavior of users on the Internet but also to evaluate and experiment different machine learning related problems. Ansah et al., focus on event detection using social media, with a special focus on Twitter. The paper presents a novel protest event detection framework called SensorTree. SensorTree utilizes the network structural connections among users in a community for protest event detection. The SensorTree framework tracks information propagation in Twitter network communities to model the sudden change in growth of these communities as burst for event detection. Once burst is detected, SensorTree builds a tensorized topic model to extract events. Extensive experiments on geographically diverse Twitter datasets using qualitative and quantitative evaluations have been performed. These experiments show the superiority of SensorTree compared to several existing state-of-the-art methods. The results further suggest that utilizing network community structure yields concise and accurate event detection.

Yet in relation with social networks and Twitter, Li et al., study the trust of users and tweets in social networks. In fact, with the generalization of Twitter and the large amounts of data generated daily on this social network, combined with the diversity of profiles and users, a

trustworthy evaluation of information and people become crucial for maintaining an open and healthy online social networking for our society. This work develops a Coupled Dual Networks Trust Ranking (CoRank) method to evaluate the trustworthiness of users and tweets by analyzing user/tweet behaviors on Twitter. A model to capture the complex characteristics and relations of both users and tweets and calculate their trust values is proposed. The approach goes beyond the existing solutions that use a single network to link both users and tweets. A set of experiments have been conducted against real data collected from Twitter. The experimental results show the effectiveness and robustness of the proposed method. A comparison study with three baseline methods: PageRank, TURank, and Weighted PageRank, has been performed and shows how the proposed approach outperforms the existing ones.

As social networks play in important role in Web communications nowadays, this other paper authored by Zhou et al. deals with the extracting a subset of representative users from the original set in social networks, which plays a critical role in Social Network Analysis. This paper proposes a novel approach and formulates the problem as RUS (Representative User Subset) problem that is proven to be an NP-Hard problem. To solve RUS problem, the authors propose two approaches KS (K-Selected) and an optimized method (ACS) that are both consisting of a clustering algorithm and a sampling model. In addition, the paper proposes a pruning strategy by taking advantages of MaxHeap structure. To validate the performance of the proposed approach, extensive experiments are conducted on two real-world datasets. Results demonstrate that our methods outperform state-of-the-art approaches.

Focusing on relations extraction and labelling, He et al., propose a multi-level distant supervision model for relation extraction, which divides the original categorization task into a number of subtasks in multiple levels of a constructed tree-like categorization structure. Each node in this tree-like structure is a sub-classifier trained with distant supervision. With the tree-like structure, an unlabeled relation instance would be categorized step by step along a path from the root node to a leaf node. In this paper, negative samples are automatically added according to predictive results of previous levels. That is, if an instance does not belong to the predicted class, the instance will be a negative sample for child nodes of this class in the next level. Furthermore, bootstrapped distant supervision is proposed to update the distant supervision model with new learned relation facts iteratively to further improve the extraction quality. Experimental results conducted on three real datasets prove that the proposed approach outperforms state-of-the-art approaches by reaching 12 + % better extraction quality.

Finally, Zhang et al., focus on adversarial attacks that may affect classifiers. In fact, recent studies claimed that ensemble classifiers tend to be more robust than single classifiers against evasion attacks. In this paper argues that this is not necessarily the case under more realistic settings of black-box attacks. In particular, it shows that a discrete-valued random forest classifier can be easily evaded by adversarial inputs manipulated based only on the model decision outputs. The proposed evasion algorithm is gradient free and can be fast implemented. Evaluation results show that random forests are even more vulnerable than SVMs, either single or ensemble, against such evasion attacks under both white-box and black-box settings.