Wireless Networks (2022) 28:1163–1167
https://doi.org/10.1007/s11276-021-02787-9

GUEST EDITORIAL

Editorial: Big data technologies and applications

Yulei Wu[1] · Yi Pan[2] · Payam Barnaghi[3] · Zhiyuan Tan[4] · Jingguo Ge[5] · Hao Wang[6]

Accepted: 7 September 2021 / Published online: 13 September 2021

Big data analytics has received numerous attentions in many areas [1–5]. This special issue contains 19 papers accepted by the 9th EAI International Conference on Big Data Technologies and Applications (BDTA-2018), which was held in Exeter, United Kingdom on 4–5 September 2018. This special issue, also the BDTA-2018 itself, is devoted to the most recent developments and research outcomes addressing the related theoretical and practical aspects on big data technologies and applications, and it also aims to provide worldwide researchers and practitioners an ideal platform to innovate new solutions targeting at the corresponding key challenges.

Jiang et al. [6] consider the structural information of the posts and the semantic information of users' interests and use UICD (User Interest Community Detection) model to analyze text streams from microblogging sites for detecting users' interest communities. Firstly, they propose modified HITS algorithm based on user interest filtering and LDA based interest detection model to distil emerging interests and high influence users by reducing negative impact of nonrelated users and its interests. Then they propose LPA and Collaborative Filtering Recommender based user interest community detection method named HLPA. The proposed method assigns a unique tag to each post, and then updates the post's label in the order of high to low. Finally, stable user interest communities can be obtained when the update is finished. The experimental results demonstrate the effectiveness of our model on users' interest community detection and in addressing the data sparsity problem of the posts.

Real-time online data processing is quickly becoming an essential tool in the analysis of social media for political trends, advertising, public health awareness programs and policy making. Traditionally, processes associated with offline analysis are productive and efficient only when the data collection is a one-time process. Currently, the cutting edge research requires real-time data analysis that comes with a set of challenges, particularly the efficiency of continuous data fetching within the context of present NoSQL and relational databases. In this paper, Shah et al. [7] from Lakehead University, Canada demonstrate a solution to effectively address the challenges of real-time analysis using a configurable Elasticsearch search engine. They propose a distributed database architecture, pre-build indexing and standardizing the Elasticsearch framework for large scale text mining. The results from the query engine are visualized in almost real-time.

Spam on social media platforms has led to a number of problems not only for social media users but also for researchers mining social media data. While there has been substantial research on automated methods of spam detection on Twitter, research on the lexical content of spam on the platform is limited. The authors in [8] presente a novel methodology that uses a dataset of 301 million generic tweets which were filtered through a URL blacklisting service to obtain 7207 tweets containing links to malicious web-pages. These tweets, considered spam, are combined with a random sample of non-spam tweets to obtain an overall dataset of 14,414 tweets. A total of 12 numerical tweet features are used to train and test a Random Forest algorithm with an overall classification accuracy of over 90%. In addition to the numerical features, the authors use the text of each tweet to create four frequency-mapped corpora pertaining uniquely to spam and non-spam data. The corpora of words, emoji, numbers, and stopwords for spam and non-spam are plotted against each other to visualize differences in usage between the two groups. A clear distinction between words, and emoji used in spam, and non-spam tweets is observed by the authors.

✉ Yulei Wu
y.l.wu@exeter.ac.uk

1 University of Exeter, Exeter, UK

2 Georgia State University, Atlanta, GA, USA

3 University of Surrey, Guildford, UK

4 Edinburgh Napier University, Edinburgh, UK

5 Chinese Academic of Sciences, Beijing, China

6 Norwegian University of Science and Technology, Trondheim, Norway

In today's society, mobile devices have become more and more popular, and social media generates large amounts of datasets every second. But the lack of geographical information on social media greatly affects the development of applications like location-based recommendation and crisis response. Location of twitter users has been found to correlate with the content of their tweets. Tang et al. [9] propose a multilayer recognition model to predict the city-level location for social network users, solely based on the user's tweet content. Through a series of optimizations such as entity selection, spatial clustering and outlier filtering, suitable features are extracted to model the geographic coordinates of tweet users. Then, the Multinomial Naive Bayes is applied to classify the datasets into different groups. The model is evaluated by comparing with an existing algorithm on twitter datasets. The experimental results reveal that this method achieves a better prediction accuracy of 54.82% on the test set, and the average error is reduced to 400.97 miles at best.

Recent developments in Big Data in financial industry has created a huge opportunity for design and development of effective aggregation (higher level) analytical measures (Fund, Portfolio, Sector, Industry etc.). Casturi and Sunderraman [10] propose a solution using rule-based selective aggregation architecture which is cost effective, efficient and building block for "Rapid Application and Decision Support Systems on Big Data". Their contribution to the paper is a new approach called "Selective Dimensional Cuboids" which enhances the existing selective dimensional selection by rule based framework to reduce the number of aggregation calculations by $(2n - 1)$ over n-dimensions. By reducing the overall calculations to $(2n - 1)$ saves computational costs, physical storage and also retrieval costs for data mining activities. The proposed architecture provides ability to set up aggregation rules by users, creates dynamic aggregation measures only for the selected dimension and avoids other dimensional measures. The solution is portable to any standard SQL platform with dimensional data characteristics.

The reengineering and redesigning work in a process-supported organization ought to be a measuring activity for the faithfulness of the underlying business processes based upon their enactment histories and event logs. This measuring activity is defined as a novel concept of workflow process fidelity, which can be concretized by comparing the original workflow process models discovered on Build-time to the mined workflow process models rediscovered from their event histories and logs recorded on Runtime. As one of the impeccable trials and outcomes addressing the workflow and business process fidelity issue, this paper [11] proposes an algorithmic approach for rediscovering all the primitive process patterns, such as sequential, exclusive-OR, parallel-AND, and iterative-LOOP process patterns, and their enactment proportions, and performs an experiment on a non-noise dataset of the XES-formatted event logs.

The authors in [12] develop an innovative machine learning heterogeneous ensemble algorithm to enhance the accuracy and reliability of classifying Multi-media Data (MMD) containing several different types of data such as numbers, text and images. The method they have developed consists of four consecutive stages: (1) features are extracted from each media data sub-set, then (2), modelling is performed independently on each of these datasets using a variety of base learning algorithms, (3) models are selected according to either their accuracy alone, or both their accuracy and diversity, and (4) an ensemble combines the outcomes from the selected models at the decision level. Hence the proposed method is called Decision Level Ensemble Method (DLEM). At the level of the accuracy and reliability, when tested on multi-media data, DLEM outperforms other machine learning methods created either by single model methods, or by other heterogeneous or homogeneous ensemble methods.

Carry-and-forward transmission method is still useful in today's Vehicular Ad Hoc Networks as it costs less than 5G technology and can support communications when infrastructures are destroyed by war or catastrophe. In the paper entitled "Optimal Throwboxes Assignment for Big Data Multicast in VDTNs" [13], Liu et al. consider using Throwboxes to help deliver data of large volume such as multimedia contents. The main contribution of this work is that the authors considered the volume of the data, which is not negligible, in practice. Throwboxes, for instance, roadside units, are common in VDTNs. Therefore, if vehicles cannot carry those big data, why not put them in throboxes? However, throwboxes themselves may have limited storage or will charge for data dropping. In this paper, the authors develop an optimal data packet assignment algorithm to achieve global maximum delivery utilities among ThrowBoxes. They evaluate the proposed scheme with the real world data trace. The results show that the proposed scheme achieves better performance in terms of delay, delivery ratio, cost and other metrics. This paper may inspire researchers in the related area to find new solutions.

Protecting Cyber-Physical systems (CPS) sensitive information from illegal access is still a big challenge because of their complexity and the big data generated from physical and digital entities. The existing privacy-preserving techniques have some limitations regarding their effectiveness in data analytics and keeping their utility. Therefore, in this paper the authors [14] study the role of big data component analysis for handling and protecting sensitive information from disclosure along with preserving the data utility. Independent Component

Analysis (ICA) technique can handle non-linear and non-normal data distributions, which is the case in CPS data. ICA transforms the raw CPS information into new shape, whilst keeping the data utility. The data utility is assessed by measuring the capability of some ML techniques to differentiate between normal and malicious patterns, as the transformed information is considered useless if the detection accuracy is too low. The mechanism is evaluated using the power CPS dataset in which the results show the effectiveness of this mechanism compared with four other privacy-preservation techniques, obtaining a higher level of privacy protection.

Research in financial domain has shown that sentiment aspects of stock news have a profound impact on volume trades, volatility, stock prices and firm earnings. These sources are mostly unstructured by nature and thus, need specific language processing tasks like polarity categorization for positive and negative sentiments in order to extract a meaningful information for future uses. Accordingly, this research [15] is organized to investigate the use of Natural Language Processing (NLP) in effort to improve the performance of sentiment classification in evaluating the information content of financial news. Sentiment classification is highly useful as an instrument in investment-based decision support systems. At present, feature extraction approach is mainly based on the occurrence frequency of words; therefore low-frequency linguistic features (bigram phrase) that could be critical in sentiment classification are typically ignored. The proposed approach to sentiment analysis for financial news classification focuses on informative low-frequency linguistic expressions in feature space. The proposed combination of low and high-frequency linguistic expressions contributes a novel set of features for text sentiment analysis and classification. The experimental results show that an optimal Ngram feature selection (combination of optimal unigram and bigram features) enhances the accuracy of sentiment classification as compared to other types of feature sets.

Continuous Restricted Boltzmann Machines are an extension of the conventional Restricted Boltzmann Machine to treat continuous data without first encoding it as a string of bits. They have great potential in machine learning and artificial intelligence as a generative network because this encoding step is inherently fragile and problem-specific. Harrison [16] develops an efficient algorithm for training the machine and demonstrates that it can be used effectively as a classification algorithm as well as a reconstruction algorithm. The algorithm constructs a pseudo-inverse to stably determine the updates in the internal weights. Unlike the accelerated discrete algorithm where an analytic expression can be derived, the continuous algorithm uses a numerical approximation to the difference in the partition function to avoid overtraining. An implementation of the algorithm is tested on standard datasets from the UCI repository and performed comparably, if not better than, other algorithms for dealing with uncertain and fuzzy data. Its application to the prediction of drug resistance in HIV protease is also demonstrated.

Continuous Ambulatory Peritoneal Dialysis (CAPD) is a treatment used by patients in the end-stage of Chronic Kidney Diseases (CKD). Those patients need to be monitored using blood tests and those tests can present some patterns or correlations. The main aim of the work in [17] is to find the best way to predict or classify the values of serum creatinine in patients undergoing CAPD procedures, through a classification method previously used in other studies, but also to understand the results expected and the results obtained. The classification process can find patterns useful to understand the patients' health development and to medically act according to such results. To accomplish such tasks, the Weka software is used in which Data Mining (DM) algorithms are already implemented. The research team can affirm that the results obtained in this paper present a solid foundation for further investigation, reaching a high performance in classifying the instances. The tested algorithms display good accuracy rates, reaching values of approximately 95%, and low relative absolute error values, which prove that the features chosen are the right choice.

The novelty of this study lies in developing a pervasive Web application based on business intelligence (BI) clinical indicators in an attempt to reduce the number of appointments, surgeries, and medical examinations that were not carried out in a Portuguese health institution most likely due to forgetfulness. It is important to note that most patients who attend the hospital are older adults and memory loss is very common with this age group. Therefore, the patients and/or their caregivers and family members are warned of their scheduled appointments, surgeries, and medical examinations via SMS in advance and appropriately by health professionals in order to reduce such numbers. On the other hand, it is also intended to strengthen the use of health information and communication technology (ICT) in healthcare environments since most current health ICT solutions are still immature according to the scientific community despise their great potential. Additionally, it is fundamental to refer that this study [18] is part of the research project "Mobile Collaborative Augmented Reality and Business Intelligence: A System to Support Elderly People's Self-care" that involves ensuring the continuity of care of patients (elderly people) and strengthening the communication strategies between health institutions, particularly nursing homes, and patients and their caregivers through technological innovation.

Allostatic State Mapping by Ambulatory ECG Repository (ALLSTAR) is a big data pro-ject of clinical Holter electrocardiograms (ECGs) in Japan. The database seems useful for investigating statistical features of 24-h ECG parameters such as heart rate variability (HRV). This research [19] examines the factors underlying 24-h HRV for which a wide variety of indices have been proposed. The authors calculate 4 time-domain, 4 frequency-domain, and 2 nonlinear HRV indices and a vagal reflex sensitivity index in 113,793 men and 140,601 women with sinus rhythm. Factor analysis delineated two factors with eigenvalue ≥ 1 ex-plaining 91% variance. Factor 1 is contributed by very-low-, low-, and high-frequency components and vagal reflex sensitivity. It increased from 0 to 20 yr, then decreased to 65 yr, and slightly increased after age 80 and increased with daily physical activity. Factor 2 is contributed by scaling exponent $\alpha 1$ and low-to-high frequency power ratio. It increases until age 35, plateaus between 35 and 55 yr, and then declines and increases with mild to moderate physical activity. The HRV of 24-h ECG consists of two major factors, reflecting cardiac vagal function and heart rate dynamics complexity, respectively, which have different relationships with age and physical activity.

Ren et al. [20] investigate the performance impact of co-running tasks on multicore computers. Machine-learning-based prediction frameworks are developed to predict the co-running performance. Historical tasks and new tasks are differentiated in the prediction frameworks, the difference between which is the framework can make use of the historical running information of historical tasks while there is no prior knowledge about new tasks. Given the limited information of new tasks, an online prediction framework is developed for new tasks by sampling the performance events on the fly for a short period and then feeding the sampled results to the prediction framework. Extensive experiments have been conducted with the SPEC2006 benchmark suite to compare the effectiveness of different machine learning methods considered in this paper. The results show that the prediction frameworks can achieve the accuracy of 99.38% and 87.18% for historical tasks and new tasks, respectively.

Owing to the time-varying network topology and constrained node resources in Mobile Opportunistic Networks (MONs), traditional social metrics are challenged to accurately measure social tie and behaviors between nodes. Furthermore, they are hard to represent contact opportunities among nodes precisely. For example, Centrality, a widely used social metric, is destination-agnostic since such metrics are usually measured without destination information. This may result in wrong forwarding decision and low routing performance. In this work, Zhang et al. [21] address the issue utilizing the destination-aware betweenness centrality (DBC) to select the right relays based on only local information given the specific destination node, then propose a Destination-Aware Social routing scheme for MONs, namely DAS. In DAS, each node independently chooses the right number of replicas for a message in respond to the network condition in a dynamic manner. Simulations results show improved performance with low overhead in comparison with existing social-aware routing schemes in various scenarios.

Big data analysis requires the speedup of parallel computing. Xen, one of the most popular virtualization platforms, is initially designed to target the management of serial jobs. Therefore, the power of parallel computing is not fully exploited in Xen due to its ineffective scheduling of parallel jobs. This paper [22] presents an application-level co-scheduler, called vChecker, which is able to co-schedule the tasks in a parallel job onto multiple CPU cores in a computer, and therefore mitigates the performance degradation of parallel jobs running in Xen. vChecker takes into account the number of CPU cores and the demand of parallel jobs, and assists the credit scheduler in Xen to schedule parallel jobs. vChecker is implemented at application level. There is no need to modify the Xen hypervisor. The experimental results show that vChecker improves the performance of parallel jobs in Xen and enhances system utilization.

As a prevailing recommendation technique that has been comprehensively explored to tackle the problem of information overload, the collaborative filtering (CF) algorithms are capable to perform adequately under various circumstances, but there still exist some shortcomings related to the lack of consideration of the semantic information about the given items. Yang et al. in the paper [23] propose a two-stage collaborative filtering approach driven by Simhash-based semantic feature analysis, of which the first stagey Simhash-based semantic feature extraction for items and categories, and the second stage is reinforced CF rating prediction driven by intensely compressed category features. The rich semantic features of vast items and their categories can be rapidly extracted and compressed in the first stage by employing the Simhash, with being utilized to promote the traditional collaborative filtering processes. Besides, to solve the problems pertaining to the Big Data context, the paper designs a parallel algorithm on Spark to accelerate the time-consuming process of semantic feature extraction for vast items. Finally, a comprehensive experiment is conducted to validate the reinforced CF approach by adopting practical datasets, and the results reveal that compared with the traditional CF algorithms it can accomplish a promising performance.

Multipath routing will incur adverse effects on the existing and emerging network measurement schemes, for example incomplete and inaccurate measurement results,

to understand network characteristics, since many of them commonly do the work under single-path routing rather than multipath-routing. In order to eliminate this emerging issue on single-path-based network measurement in Internet, it requires to identify whether there is multipath routing between two reachable hosts in the network. Notice that no out-of-order delivery among a strip of packets along multiple paths seldom occurs, in this paper [24], an efficient multipath routing identification approach has been proposed to achieve this goal, by introducing a composite probe built on out-of-order delivery. They have elaborated the proposed theoretical observation on the current probe composed of a strip of packets, and then presented the composite probe design in detail. The proposed approach not only can efficiently identify the existing multipath routing, but also accurately recognize its type, referring to flow-based or packet-based routing. Corroborated by experiments and simulations, conducted on Planetlab and NS2, respectively, the proposed approach outperforms other schemes in terms of effectiveness and accuracy.

Acknowledgements We would like to express our deep thanks to the Editor-in-Chief, Professor Imrich Chlamtac, for providing us with the opportunity to host this special issue in Wireless Networks Journal. We also thank all the authors who submitted their papers. Last but not least, we thank the thoughtful work of the many reviewers who have provided invaluable evaluations and recommendations.

References

1. Wu, Y., Hu, F., Min, G., & Zomaya, A. (Eds.). (2017). *Big data and computational intelligence in networking*. Taylor & Francis/CRC. ISBN: 9781498784863
2. Zuo, Y., Wu, Y., Min, G., & Cui, L. (2019). Learning-based network path planning for traffic engineering. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2018.09.043
3. Wu, Y., Dai, H.-N., & Tang, H. (2021). Graph neural networks for anomaly detection in industrial internet of things. *IEEE Internet of Things Journal*. https://doi.org/10.1109/JIOT.2021.3094295
4. Huang, H., Yin, H., Min, G., Jiang, H., Zhang, J., & Wu, Y. (2017). Data-driven information plane in software-defined networking. *IEEE Communications Magazine, 55*(6), 218–224.
5. Cheng, X., Wu, Y., Min, G., & Zomaya, A. Y. (2018). Network function virtualization in dynamic networks: a stochastic perspective. *IEEE Journal on Selected Areas in Communications*. https://doi.org/10.1109/JSAC.2018.2869958
6. Jiang, L., Shi, L.-L., Liu, L., Yao, J., & Yousuf, M. A. (this issue). User interest community detection on social media using collaborative filtering
7. Shah, N., Willick, D., & Mago, V. (this issue). A framework for social media data analytics using elasticsearch and Kibana.
8. Robinson, K., & Mago, V. (this issue). Birds of prey: Identifying lexical irregularities in spam on twitter.
9. Tang, H., Zhao, X., & Ren, Y. (this issue). A multilayer recognition model for twitter user geolocation.
10. Casturi, R., & Sunderraman, R. (this issue). Cost effective, rule based, big data analytical aggregation engine for investment portfolios.
11. Kim, K., Lee, Y., Ahn, H., & Kim, K. P. (this issue). An experimental mining and analytics for discovering proportional process patterns from workflow enactment event logs.
12. Alyahyan, S., & Wang, W. (this issue). Decision level ensemble method for classifying multi-media data.
13. Liu, P., Ding, Y., & Fu, T. (this issue). Optimal throwboxes assignment for big data multicast in VDTNs.
14. Keshk, M., Moustafa, N., Sitnikova, E., & Turnbull, B. (this issue). Privacy-preserving big data analytics for cyber-physical systems.
15. Yazdani, S. F., Tan, Z., Kakavand, M., & Mustapha, A. (this issue). NgramPOS: A bigram-based linguistic and statistical feature process model for unstructured text classification.
16. Harrison, R. W. (this issue). Continuous restricted Boltzmann machines.
17. Brito, C., Esteves, M., Peixoto, H., Abelha, A., & Machado, J. (this issue). A data mining approach to classify serum creatinine values in patients undergoing continuous ambulatory peritoneal dialysis.
18. Esteves, M., Abelha, A., & Machado, J. (this issue). The development of a pervasive web application to alert patients based on business intelligence clinical indicators: A case study in a health institution.
19. Yuda, E., Kisohara, M., Yoshida, Y., & Hayano, J. (this issue). Constituent factors of heart rate variability ALLSTAR big data analysis.
20. Ren, S., He, L., Li, J., Chen, Z., Jiang, P., & Li, C.-T. (this issue). Contention-aware prediction for performance impact of task co-running in multicore computers.
21. Zhang, J., Huang, H., Yang, C., Liu, J., Fan, Y., & Yang, G. (this issue). Destination-aware metric based social routing for mobile opportunistic networks.
22. Jiang, P., He, L., Ren, S., Chen, Z., & Mao, R. (this issue). vChecker: An application-level demand-based co-scheduler for improving the performance of parallel jobs in Xen.
23. Yang, P., Gu, L., & Liu, X. (this issue). Collaborative filtering driven by fast semantic feature analysis on spark.
24. Huang, H., Pan, S., & Zhang, J. (this issue). Multipath routing identification for network measurement built on end-to-end packet order.