J Grid Computing (2019) 17:205–208
https://doi.org/10.1007/s10723-019-09484-0

Special Issue on Knowledge Discovery in Big Data (KDBD)

Sajid Anwar · Álvaro Rocha

Received: 27 May 2019 / Accepted: 27 May 2019 / Published online: 10 June 2019
© Springer Nature B.V. 2019

Technological advancements pertaining to big data such as artificial intelligence, business analytics, data mining and machine learning, help greatly in making the right decisions at the right time. The WWW and social area networks (SANs) have contributed greatly to the volume and heterogeneity of data available, (e.g., text, images, videos, audio and drawings). Using big data to obtain greater insight poses significant challenges in terms of computing efficiency, business analytical problem-solving and knowledge discovery. This special issue is intended to be a meeting point between researchers and big data analysts. The issue focuses on problems relating to big data, giving the opportunity to researchers to propose new methods of transforming technological frameworks into big data techniques and to produce new research results pertaining to knowledge discovery.

For this special issue on knowledge discovery in big data (KDBD), we have collected 10 articles that address relevant developments and advances in the area. These articles were selected and improved during a two-round review process, in accordance with the standard practices of the *Journal of Grid Computing*. The resulting 10 contributions cover some key aspects and developments regarding big data technologies in the realm of knowledge discovery.

In first paper, entitled “A Dynamic Profile Questions Approach to Mitigate Impersonation in Online Examinations”, the authors have aimed to strengthen the authentication of students via the use of dynamic profile questions during an online examination, to reduce collusion. Collusion is seen as a major security threat in such examinations. It occurs when a student invites a third party to impersonate or abet him or her in a test. The results of the usability and security analysis are reported. It was found in this study that the dynamic profile questions were more usable than both text-based and image-based questions. It is also revealed that impersonation attack via email was not successful, and that students were able to share answers to dynamic profile questions with a third-party impersonator in real time, which resulted in 93% correct answers. The study also revealed that a response-time factor may be implemented to identify and report impersonation attacks.

The second paper, entitled “Beyond Homology Transfer: Deep Learning for Automated Annotation of Proteins” addresses the need to create sequence-based computational techniques that can precisely annotate uncharacterized proteins. In this paper, the authors propose Deep Seq—a deep learning architecture that utilizes only the protein sequence information to pre-

S. Anwar (✉)
Institute of Management Sciences, Peshawar, Pakistan
e-mail: sajid.anwar@imsciences.edu.pk

Á. Rocha
University of Coimbra, Coimbra, Portugal
e-mail: amrocha@dei.uc.pt

dict its associated functions. The prediction process does not require handcrafted features; rather, the architecture automatically extracts representations from the input sequence data. The results of experiments with Deep Seq indicate significant improvements in terms of prediction accuracy, compared with other sequence-based methods. The deep learning model achieves an overall validation accuracy of 86.72%, with an F1 score of 71.13%. Improved results for the protein function prediction problem were achieved through Deep Seq, utilizing sequence information only. Moreover, using the automatically learned features, and without any changes to Deep Seq, the authors successfully solved a different problem, i.e., protein function localization with no human intervention. Finally, they also discuss how the same architecture can be used to solve even more complicated problems, such as prediction of 2D and 3D structures, as well as protein-protein interactions.

In the third paper, "Optimized Gabor Feature Extraction for Mass Classification Using Cuckoo Search for Big Data E-Healthcare", the authors present a computer-aided system for healthcare that may be an effective tool for automatically processing big data relating to breast cancer. They are of the opinion that, that in the field of medicine, sensitivity to false positives is very high because it results in false diagnosis and can lead to serious consequences. Therefore, the challenge for researchers, is to correctly distinguish between normal and affected tissues, to increase the detection accuracy in patients with breast cancer. Radiologists use Gabor filter banks for feature extraction, applied to the entire input image, yielding poor results. The authors have devised a system for optimizing the Gabor filter bank to select the most appropriate Gabor filter using a metaheuristic algorithm known as "cuckoo search". The proposed algorithm is run over subimages, in order to extract more descriptive features. Moreover, feature subset selection is used to reduce features. The algorithm is considered to be more efficient, faster and less complex, leading to improved results. The authors tested the proposed method on 2,000 mammograms taken from the DDSM database and found that it outperformed some of the best techniques used for mammogram classification, based on sensitivity, specificity, accuracy and area under the curve (ROC).

In the fourth paper, "An Improved Method to Deploy Cache Servers in Software-Defined Network-Based Information-Centric Networking for Big Data" the authors have addressed the issues of scalability and optimality in connection with information-centric networking (ICN), with a centralized cache server architecture that transfers big data from one node to another and provides in-network caches. To resolve these issues, they deployed multiple cache servers based on joint optimization of multiple parameters, namely: (i) closeness centrality, (ii) betweenness centrality, (iii) path-stretch values and (iv) load balancing in the network. In this research, they first compute the locations and the number of cache servers based on the network topology information in an offline manner, and then place the cache servers at their corresponding locations in the network. Next, the controller installs flow rules at the switches, such that a switch can forward the request for content to one of its nearest cache servers. If the content request matches the contents stored at the cache server, the content is delivered to the requesting node; otherwise, the request is forwarded to the controller. In the next step, the controller computes the path, such that the content provider first sends the content to the cache server. Finally, a copy of the content is forwarded to the requesting node. The authors have confirmed through simulations that the approach performs better in terms of traffic overhead and average end-to-end delay, compared with an existing state-of-the-art approach.

The fifth paper in the special issue is entitled "What Is Happening around the World? A Survey and Framework on Event Detection Techniques on Twitter". In this study, the authors have used Twitter data for detecting events happening around the world. The content (tweets) published on Twitter are short and pose diverse challenges for detecting and interpreting event-related information. This article provides insights into ongoing research and helps in understanding recent research trends and techniques used for event detection using Twitter data. The authors classify the techniques and methodologies according to event types, orientation of content, event detection tasks, task evaluation and common practices. Then, they highlight the limitations of existing techniques and propose solutions accordingly, to address the shortcomings. They propose a framework called

EDoT, based on research trends, common practices and techniques used for detecting events on Twitter. EDoT can serve as a guideline for developing event detection methods, especially for researchers who are new in this area. This research article describes and compares data collection techniques and the effectiveness and shortcomings of various Twitter-based and non-Twitter-based features and discusses various evaluation measures and benchmarking methodologies. Finally, the paper discusses the trends, limitations and future directions for detecting events on Twitter.

In sixth paper, "Application of Parallel Vector Space Model for Large-Scale DNA Sequence Analysis" the challenges relating to the large scale and complex structure of DNA datasets are studied. This paper presents a novel parallel vector space model (PVSM) approach that supports the analysis of large-scale DNA sequences by taking advantage of a multicore system. The proposed approach is built on top of a modified vector space model (VSM). In order to assess the performance of PVSM, the proposed technique is extensively evaluated in the context of computational efficiency and accuracy, using DNA sequences of various sizes. The performance of PVSM is compared with sequential modified VSM. The sequential VSM is implemented on a single processor, whereas the method presented in this paper is initially parallelized on four processors and subsequently on 12 processors. The experimental results show that PVSM performed better than the sequential VSM. The proposed method achieved approximately twice the speed of the sequential approach, without affecting the accuracy level. Moreover, the proposed PVSM is highly scalable by increasing the number of processing cores, supporting the analysis of large-scale DNA sequences.

In the paper "Enhancing Text Using Emotion Detected from EEG Signals", published as the seventh paper, the authors have proposed an end-to-end system which aims to enhance user-input sentences according to the user's current emotional state, while communicating via social media. It works by a) detecting the emotion of the user and b) enhancing the input sentence by inserting emotive words, to make the sentence more representative of the emotional state of the user. The emotional state of the user is recognized by analysing electroencephalogram (EEG) signals from the brain. For text enhancement, the words corresponding to the detected emotion are modified using a correlation-finder scheme. Next, verification of the correctness of the sentence is performed using a long short-term memory (LSTM) network-based language modelling framework. An accuracy of 74.95% was recorded for the classification of five emotional states in a dataset consisting of EEG signals from 25 participants. Similarly, promising results were obtained for the task relating to text enhancement and the overall end-to-end system. To the best of our knowledge, this work is the first to enhance text according to the emotional state detected by EEG brainwaves. The system releases an individual from thinking of and typing words, which may sometimes be a complicated procedure.

In the eighth paper, "An Efficient Real-Time Data Dissemination Multicast Protocol for Big Data in Wireless Sensor Networks", the role of wireless sensor networks in data collection and dissemination of big data is examined, especially in real-time communication where current schemes are assumed to use a general traffic model for real-time delivery, and therefore lack adaptability. To solve this problem, a few routing protocols were designed to accommodate the new real-time model, (m,k)-firm, which is regarded as the most applicable scheme for event-based as well as query-based applications in wireless sensor networks. However, since current schemes for (m,k)-firm streams support unicast communications only, they cannot be applied to multicast communications where many group members are willing to receive data packets from the sink node. To overcome this problem, the authors propose a new multicast scheme for (m,k)-firm streams, to deliver data packets to group members. To construct a multicast tree, different types of overlay tree are constructed according to their distance-based priority (DBP) value. Simulation results prove that the proposed scheme can meet (m,k)-firm requirements and that it has a longer network lifetime than existing schemes.

In the ninth paper, entitled "Big Data Analytics, Text Mining and Modern English Language" the authors used text mining techniques to analyse old and modern English, in order to examine the evolution of the modern English language. They introduce a common-words counting algorithm that identifies common words of the 15th century that diminished gradually in usage in later centuries. The authors computed the speed of linguistic changes and identified the reasons behind them. For this purpose, 34,000 textbooks were downloaded from Project Gutenberg with different authors from the 15th to the 19th centuries. These books were categorized into five centuries. In their study, the authors selected common words from the books of the 15th century and calculated their frequencies in other centuries, calculating the sum of term frequency-inverse document frequency (TFIDF) values for these words and proving that the frequencies of the words decreased from the 15th century to the 19th century, with some words, such as 'doth', 'hath', 'punt', 'guise' and 'selfe' even disappearing during different centuries. The authors then calculated the speed of change of words using the slope formula. This proved that words were changing in each century, with the speed of change of words being lowest between the 16th and 17th centuries and highest between the 18th and 19th centuries, showing that the old words, or their spellings, changed to the modern words mostly during the 18th and the 19th centuries. The authors attribute these changes to industrialization, modernization and the expansion of the British Empire, seeing these as key factors that changed the old English language into the modern English language.

In the final paper, "Semantic Orientation-Based Decision-Making Framework for Big Data Analysis of Sporadic News Events" the authors investigate the growing public endorsement of social media, which has changed public life dramatically, through sentiment analysis. They use big data analytics to offer remarkable opportunities to individuals and organizations by providing proficient decision-making frameworks and improved forecasting models, through sentiment analysis of social media. In this research, they analyse public views, sentiments and opinions shared on social media about a democratic participatory activity called the 'Azadi march', which was held in Pakistan with the participation of online users from all over the world. The authors carried out computational semantic orientation on public tweets for analysing public awareness and the effects of online communication through social media on real-world public decision-making. In the study, the authors employed an unsupervised approach for identification and scoring of tweets. They then used a lexicon-based approach, in which annotated lexicons are used for scoring verbs, adverbs and other parts of speech. A corpus is used for scoring adjectives and informal opinion indicators. Emojis, exclamatory statements and other additional features are incorporated as a supplementary analysis. In the study, it is shown that emoticons and NetLingo play a significant role in sentiment orientation. Opinion groups are generated from all retrieved tweets and the aggregate sentiment weights of the opinion groups are computed. The findings of this study indicate that the proposed lexicon-based approach outperforms contemporary machine learning techniques, achieving 86% average accuracy for sentence-level sentiment analysis.

After a thorough and rigorous review process, we believe that these 10 articles constitute an interesting set of contributions addressing the theme of **Knowledge Discovery in Big Data (KDBD)**. We would like to thank the authors who have submitted their work to this special issue, for their valuable insights and contributions. A special thanks also to the reviewers who, with their commitment and expertise, have helped us to select the best articles and to improve their content. We hope that readers find this special issue an interesting and valuable source of information relating to their own work. Finally, we would like to express our gratitude to Prof. Peter Kacsuk, Editor-in-Chief of the Journal of Grid Computing, for giving us the opportunity to organize this special issue and for his continuous support during the whole process.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.