Editorial

Lorna Uden

Faculty of Computing Engineering and Sciences (FCES),
School of Computing,
Staffordshire University,
College Road, Stoke-on-Trent, Staffordshire ST4 2DE, UK
Email: L.uden@staffs.ac.uk

Welcome to V13 N4 issue of *IJWET*. It consists of four papers. The first paper is 'Stable web scraping: an approach based on neighbour zone and path similarity of page elements', by Peng Gao, Hao Han, Junxia Guo and Motoshi Saeki.

According to these authors, extracting data from the web is not a trivial task because most of them are presented in a semi-structured manner, typically in HTML format with no structured interface. Web scraping techniques based on XPath enable users to consistently extract information of interest from webpages that do not provide a structured interface. However, XPath-based extraction is likely to fail when encountering page variants, resulting in a high cost of repair. Countermeasures based on pattern matching or model learning often require careful pre-processing, which is not suitable for cases where the target data is frequently re-designated. The authors of this paper present a new extraction method for the stable scraping of arbitrary designated data from webpages. Instead of attempting to find the desired data directly, they first determine its approximate location in the changed page, called the neighbour zone. Then they search for the precise location by ranking the path similarity of page elements within the neighbour zone. Experiments on a large set of real-world webpages show that their method has better stability for web scraping, compared with the XPath-based extraction. In the two datasets, 0.118 and 0.891 F1-scores were increased respectively. Further work is needed to develop the calculation of unchanged node pairs and the semantic similarity of attributes to further improve the effectiveness of the SSN and SSNP methods.

The second paper is 'Malicious behaviour classification in web logs based on an improved Xgboost algorithm' by Jiaming Song, Xiaojuan Wang, Lei Jin and Jingwen You. According to these authors, attacks against web servers are one of the most serious threats in security fields. Analysing the web logs is one of the most effective methods to identify malicious behaviours. At present, web attacks are so complex that single layer classification model is unable to deal with the emerging attacks, in particular, there is a limitation that category features cannot be added to single layer model.

This paper presents an improved Xgboost classification algorithm for malicious behaviour detection. The malicious behaviour classification model takes web server log files that conform to common log format as input data and performs feature extraction for each web request. More precisely, the classification model utilises a specific HTTP query structure to identify malicious behaviour. The authors argue that experimental results show that, compared to other machine learning algorithms, the improved Xgboost algorithm performs better. Further research is needed to verify the results.

The third paper is 'Correlation-based feature subset selection technique for web spam classification' by Surender Singh and Ashutosh Kumar Singh. According to the authors, machine learning algorithms and web spam features has been used to recognise spam. Feature selection (FS) method has been already applied in many data mining problems to reduce the classification error or increase accuracy. FS can be divided into feature ranking and feature subset selection.

In this paper, correlation-based feature selection (CFS) technique (with best-first search) is used which selects features that are most efficient. The main research focus of this paper to investigate various FS methods related to the problem. The author argues that results from the proposed model show that the CFS using BFS removes irrelevant features of different feature sets of used web spam datasets efficiently and improves the accuracy for classifiers. In case of WebSpam-UK2006, the proposed approach achieved the maximum AUC in case of J48 and naïve Bayes, for the link-based features and combination of content with link-based respectively. While in case of WebSpam-UK2007, the maximum value of AUC is shown by naïve Bayes for content-based features. Time complexity is also reduced significantly because the number of features in all feature sets are decreased to a great extent (details in Appendix) while improving the accuracy of the classifier. CFS gives better results but these outcomes are specific to the type of feature sets and classifiers used. Further work is required to verify such claims. For future work, it would be good to propose a wrapper and hybrid-based FS method.

The fourth paper is 'Sentiment analysis based on the domain dictionary: a case of analysing online apparel reviews' by Ran Tao, Yuanguo Luo and Guohua Liu. These authors suggested that past customer emotion outlined in their reviews plays an important role in not only the purchasing decisions of potential consumers, but also in manufacturers' production plans and in businesses, maintenance of their shopping environments. The amount of online reviews is too large to obtain useful information using only traditional methods. Rather, it is necessary to look for help from the computer technologies to collect and handle the massive online reviews. Text mining and sentiment analysis (SA) are emerging as appropriate methods.

This paper proposes a SA approach based on the domain dictionary and a case study is used as an example to help enterprises and researchers more effectively apply social media analytics in practice. The approach uses web text data acquisition, natural language pre-processing, a domain dictionary, emotion calculation rules, SA, and visualisation techniques for extracting structured subjective customer emotion and objective commodity metadata from unstructured review pages and analysing the relationships between them. The value of the proposed approach was demonstrated through a case study by using online apparel reviews of an anonymous brand in China in order to understand the relationship between the subjective customer emotion and the objective commodity metadata. The study found that domain dictionaries can be more effective in extracting emotional information. More empirical studies are necessary to verify the results.