Introduction to Data, Text, and Web Mining for Business Analytics Mini-track

Dursun Delen
Oklahoma State University

Hamed Zolbanin
Ball State University

This mini-track has a total of nine papers that are about developing analytics systems for decision support by means of data, text, or web mining. Five of the nine papers focus on a variety of interesting text mining, natural language processing, and sentiment analysis. The remaining four papers deal with deep learning for recommendation identification, stream analytics for adverse event detection, statistical analysis for ranking of multi-attribute numerical objects, and finally using individual and ensemble models to predict the outcome of the competitive sporting events.

The paper by Lutz *et al.* propose the use of distributed text representations and multi-instance learning to analyze, with high interpretability, the sentiment of individual sentences in financial news. In contrast to previous approaches, which merely predict the stock market's reaction to the news on a document level, the method proposed in this study transfers information from the document level to the sentence level. This method improves the performance of the existing approaches by more than 3.80%.

The paper by Gan *et al.* proposes a deep neural network based recommendation model, named Convolutional Dense-layer Matrix Factorization (CDMF), for the context-aware recommendation. This model extracts hidden features from item description and then fuses it with tag information. The proposed model is a multi-sourced model that generates a comprehensive feature vector. CDMF makes rating predictions based on the fused information of both users and items and outperforms the state-of-the-art recommendation methods.

The paper by Alattas *et al.* develops an unsupervised algorithm to rank numerical observations, an important application in information retrieval. This algorithm uses the correlation coefficients between attribute values and magnetic properties to rank multi-attribute numerical objects. The proposed algorithm improves upon the existing techniques and is able to handle missing values.

Nakayama and Wan analyze the sentiment of online reviews to examine their biases and helpfulness at the aspect level across two different cultures. They find several differences between the Japanese and Western consumers in regards to their emotions about their experience at restaurants or about what each consumer group valued the most. The authors call for more research on the use of sentiment analysis to explore how cultural differences may influence online reviews for experience / subjective goods.

The paper by Liu and Yoon employ the integrated power of computational syntax, semantics, and indexical pragmatics to propose an ontology-driven framework for the retrieval of domain-specific content on the Internet. This design addresses some of the shortcomings of the traditional methods, such as heavy dependence on query expansion, lexical analysis of text, and the need for large amounts of training data.

Shi *et al.* use three data stream regression algorithms to propose a novel data stream approach for the prediction of adverse events in a war theater. Compared to previous studies, the proposed approach is able to include a greater number of input variables, and therefore, improve the results over the machine learning methods developed in previous studies. Other advantages of the proposed stream based method are its reduced time and space complexity.

The paper by Wambsganß and Fromm design and prototypes an artifact that automatically extracts user-generated repair instructions from the web and uses Natural Language Processing methods to transform these data into numerical features. The system then classifies the content into either repair instructions or not, thereby saving a lot of time and money for companies that had to do this task manually.

In another text mining study, Liu proposes a new topic extraction schema to identify the key noun-phrases by constructing a context-free grammar (CFG) from input documents. In the proposed method, documents are reconstructed as a set of CFG rules, which in turn, have a hierarchical structure. Using the hierarchical structure of the input document, Liu designs a new algorithm to identify and extract key noun-phrases.

The last, but not the least, paper by Eryarsoy and Delen illustrates another interesting application of advanced analytics to predict the outcomes of the European football games. The specific objective of their study was to develop and compare a number of advanced analytics models to predict the outcomes of soccer (or association football) games (win, loss or draw), and determine the dominant factors/attributes that influence the game outcomes. They used 10 years of comprehensive game-level data spanning the years 2007-2017 in the Turkish Super League and tested a variety of individual and ensemble classifier methods to identify the most promising methods for outcome predictions.