Introduction to Data, Text, and Web Mining for Business Analytics Minitrack

Dursun Delen
Oklahoma State University
dursun.delen@okstate.edu

Enes Eryarsoy
Sehir University, Turkey
eneseryarsoy@sehir.edu.tr

Şadi E. Şeker
Istanbul Medeniyet University
sadi.seker@medeniyet.edu.tr

Abstract

Data mining (along with its derivatives that include text mining and Web mining) is one of the most popular enablers of business analytics. Although its roots dates back to late 1980s and early 1990s, most noteworthy/impactful outcomes of data mining come out after the turn of this century. Many believe that the recent popularity of analytics can largely be credited to the increasing use of data mining, which is capable of extracting and providing much needed insight and knowledge to decision makers at any and all levels of managerial hierarchy. The term data mining was originally used to describe the process through which previously unknown patterns in data were discovered. This definition has since been stretched beyond those limits by software vendors and consultancy companies to include most forms of data analysis in order to increase its reach and capability. With the emergence of analytics as an overarching term for all data analyses, data mining is put back into its proper place—a critical part of analytics continuum where the new discovery of knowledge happens.

This mini-track has six papers, collectively illustrating the depth and breadth of data, text and Web mining, and their innovative applications to interesting and highly challenging business problems.

1. Introduction to the papers

The six research papers accepted for this minitrack can be divided into two groups—the first group of papers are mostly related to development of data mining methods, methodologies and algorithms, and their applications to complex real-world problems; and the second group of papers are related to text mining and its derivative application area, sentiment analysis, and their applications to real-world problems where textual data is collected from variety of sources including Internet and social media.

In the first paper, Erraguntla *et al.* propose an innovative applied data analytics and data mining research project, which was funded by U.S. Army Medical Research and Materiel. The goal of the project, creatively named as Data Integration and Predictive Analysis System (IPAS), is to enable prediction, analysis, and response management for incidents of infectious diseases. IPAS collects and integrates comprehensive datasets of previous disease incidents and potential influencing factors to facilitate multivariate, predictive analytics of disease patterns, intensity, and timing. IPAS supports comprehensive epidemiological analysis—exploratory spatial and temporal correlation, hypothesis testing, prediction, and intervention analysis. Innovative machine learning and predictive analytical techniques like support vector machines (SVM), decision tree-based random forests, and boosting are used to predict the disease epidemic curves. Predictions are then displayed to stakeholders in a disease situation awareness interface, alongside disease incidents, syndromic and zoonotic details extracted from news sources and medical publications. Data on Influenza Like Illness (ILI) provided by CDC was used to validate the capability of IPAS system, with plans to expand to other illnesses in the future. This paper presents the ILI prediction modeling results as well as IPAS system features.

The second paper, which is written by Albashrawi *et al.*, investigates mobile banking (MB) usage through the theoretical lens of UTAUT model with its four pillars. The research model proposes to be tested using a hybrid neural networks-based structural equation modeling (SEM-NN) framework to identify the significant contributors/factors. Universal structural modeling (USM) can then be utilized to find the hidden paths and nonlinearity in our research model. To the best of authors' knowledge, this is the first study to examine the role of subjective and objective experience on MB usage by utilizing a multi-analytical approach. Neural network (NN) and USM can identify the most significant determinants and hidden interaction effects, respectively. Thus, both techniques would help to complement SEM and increase our understanding of the influential factors on MB usage. The paper presents their preliminary results and discusses the implications.

The third paper, written by Egger and Schoder, is about consumer-oriented tech mining, a relatively new an innovative concept in the world of analytics. According to the authors, to avoid missing

URI: http://hdl.handle.net/10125/41284
ISBN: 978-0-9981331-0-2

technological opportunities and to counteract risks, organizations ought to scan and monitor developments in the external environment through a structured process of technology intelligence. Previous approaches in tech mining—the application of text mining for technology intelligence—have primarily focused on the elicitation of technical or legal information from Web, patent, or research databases. However, knowledge of consumers' needs, fears, and hopes is a prerequisite for the success of an emerging technology in the marketplace. Thus, the authors claim that technology intelligence needs to also consider consumers' technology perceptions. Hence, they propose a novel and comprehensive approach to collect user-generated content from the Web and apply text mining to derive consumer perceptions. In doing so, they align with an established tech-mining process. This paper illustrates their approach on the emerging technology of autonomous driving and provides an initial indication of concurrent validity.

The fourth paper, by Jaakonmäki *et al.*, is about social media marketing. According to the researchers, social media has become an important tool in establishing relationships between companies and customers. However, creating effective content for social media marketing campaigns is a challenge, as companies have difficulty understanding what drives user engagement. According to the authors, one approach to addressing this challenge is to use analytics on user-generated social media content to understand the relationship between content features and user engagement. In this paper authors report on a quantitative study that applies machine learning algorithms to extract textual and visual content features from Instagram posts, along with creator- and context-related variables, and to statistically model their influence on user engagement. They claim that their findings can guide marketing and social media professionals in creating engaging content that communicates more effectively with their audiences.

The fifth paper, which is written by Märkle-Huß *et al.*, propose a method/methodology to improve sentiment analysis with document-level semantic relationships from rhetoric discourse structures. According to the researchers, the conventional sentiment analysis usually neglects semantic information between (sub-)clauses, as it merely implements so-called bag-of-words approaches, where the sentiment of individual words is aggregated independently of the document structure. Instead, they claim to advance sentiment analysis by the use of rhetoric structure theory (RST), which provides a hierarchical representation of texts at document level. For this purpose, texts are split into elementary discourse units (EDU). These EDUs span a hierarchical structure in the form of a binary tree, where the branches are labeled according to their semantic discourse. Accordingly, this paper proposes a novel combination of weighting and grid search to aggregate sentiment scores from the RST tree, as well as feature engineering for machine learning. They apply their algorithms to the especially hard tasks of prediction, such as stock returns subsequent to financial disclosures. As their results show, machine learning improves the balanced accuracy by 8.6 percent compared to the baseline.

The sixth and the last, but not the least, paper is written by Liu et al., and it is about Natural Language Processing (NLP) research in information systems. According to the authors, NLP is now widely integrated into web and mobile applications, enabling natural interactions between human and computers. Although many NLP studies have been published, none have comprehensively reviewed or synthesized tasks most commonly addressed in NLP research. The authors conduct a thorough review of IS literature to assess the current state of NLP research, and identify 12 prototypical tasks that are widely researched. Their analysis of 238 articles in Information Systems (IS) journals between 2004 and 2015 shows an increasing trend in NLP research, especially since 2011. Based on their analysis, they propose a roadmap for NLP research, and detail how it may be useful to guide future NLP research in IS. In addition, they employ Association Rules (AR) mining for data analysis to investigate co-occurrence of prototypical tasks and discuss insights from the findings.

2. Summary and conclusion

This minitrack has been a part of HICSS since 2007, celebrating its 10th anniversary this year along with HICSS' 50th anniversary. The success of the minitrack can largely be attributed to the attractiveness and increased popularity of business analytics as a rewarding research topic in information systems and systems sciences as well as the quality and quantity of papers submitted by the reputable researchers/authors of the topic area all over the world.

This year, Data, Text and Web Mining for Business Analytics minitrack had six research papers accepted for presentation at the HICSS conference and publication in the conference proceeding. These six high-quality papers dealt with a variety of algorithmic and methodological issues in this domain, and also provided real-world examples to showcase the novelty and contribution of their proposed solution approaches.