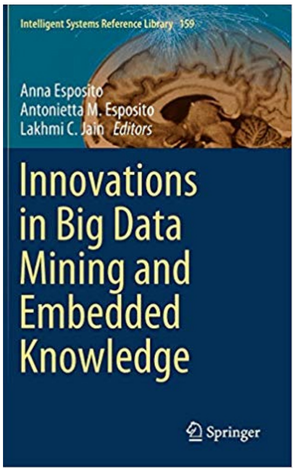
Intelligent Decision Technologies 15 (2021) 5–8
DOI 10.3233/IDT-219003
IOS Press

Book Review

Abstract. This is a review of a book that discusses knowledge discovery using data mining and knowledge embedding through models. A number of scheme are reported in the book to explain how data mining and discovered embedded knowledge can be beneficial to social organizations, domestic sphere and ICT market. Each chapter of the book presents a unique problem of data mining philosophies from an embedded point of view. It will help researchers to understand the current status of big data mining and embedded knowledge, discover new research opportunities and gain more information about this field.

Keywords: Big data, data mining, embedded knowledge, machine learning

"Innovations in Big Data Mining and Embedded Knowledge" by Anna Esposito (Editor), Antonietta M. Esposito (Editor), Lakhmi C. Jain (Editor), 1st edition (16 July 2019), Pages 276, ISBN 978-3-030-15938-2, Vol. 159 in Springer Book Series *Intelligent Systems Reference Library* (https://www.springer.com/series/8578)


Intelligent Systems Reference Library 159
Anna Esposito
Antonietta M. Esposito
Lakhmi C. Jain Editors
Innovations in Big Data Mining and Embedded Knowledge
Springer

1. Introduction

Blackler [1] identifies five types of interrelated knowledge that play an important role in our life. These are, embodied, encultured, embedded, encoded and embrained knowledge. Embedded knowledge refers to the knowledge that is locked in processes, products, culture, routines, artifacts, or structures [2]. The challenges in managing embedded knowledge vary considerably as culture, principle, emotion, ethics and routines can be difficult to understand, hard to change and changing with time. It is important to note, that while embedded knowledge can exist in explicit sources, the knowledge itself is not explicit. Due to the difficulty in effectively managing embedded knowledge, organizations that succeed to do that might enjoy a significant competitive advantage.

In addition to that we are living in a big data era, as available data from various sources are growing at an explosive rate. With the availability of vast amounts of data, most organizations have started focusing on leveraging valuable insights from it. Those insights can give businesses a competitive advantage in the market. On the one hand, big data mining has been demonstrated as the 21st century's most powerful tool; on the other hand, this high-level tool is too complicated to mine embedded knowledge. With many vivid real-world examples, the authors fill the gap by sharing their extensive experience. The authors go into great detail to elaborate as to what this new frontier upholds in the modern world.

The book titled "Innovations in Big Data Mining and Embedded Knowledge" edited by Anna Esposito, Antonietta M. Esposito and Lakhmi C. Jain explores new directions in research area of embedded knowledge and big data mining techniques. This type of knowledge discovery is beneficial to support humans in resolving various issues and problems. In recent year, few research articles are available which amalgamate big data mining and embedded knowledge. Thus, the book comes at right time when there is a great demand for big data mining and knowledge discovery is only increasing. The book focuses on several types of applications rather than concrete theory. This book updates students, researchers and teachers with the various research problems and applications showing how data mining and discovering embedded knowledge can be beneficial in every sphere of our life. Presenting the problems as well as solutions in a lucid manner makes this book unique.

ISSN 1872-4981/$35.00 © 2021 – IOS Press. All rights reserved.

2. Content

The book contains thirteen chapters where chapter one makes groundwork and provides overview of each chapter. Chapter 2 designs a recommender system on a big data platform for tourism industry. This chapter starts with big data characteristics and focuses on future tourism from big data point of view. The authors introduce fundamental knowledge of recommender system and its evaluation metrics to the readers and slowly make a transition to three layer lambda architecture. The underlying technology supports data acquisition, storage and processing. A multi-agent tourism recommender system is proposed that combines collaborative filtering, content-based and demographic techniques.

Chapter 3 is devoted in applying machine learning (ML) techniques in big data applications. The real challenge of applying ML to big data is to handle scalability of training data set and understandability of the models. Here authors address the issues for classification of sequence data in biological domain. Initially, machine learning (ML) techniques are used to produce classification model and then learned model is applied to make prediction. Readers are also introduced with the statistical classification methods and feature based classification methods. An evolutionary machine learning framework (EML) is proposed by the authors where evolutionary feature generator (EFG) uses genetic programming (GP) algorithm to extract useful features from training datasets and partial spatial boosting machine learner (PSBML) algorithm performs scalable classification. Although algorithm and experimental sections show the result but the visualization of ROC curve is missing. Then Chapter 4 and 5 make assessment survey.

Chapter 4 presents a data analysis on the relation of achievement in International Large-Scale Assessment (ILSA) and culture, and designing culturally appropriate educational technology. Two datasets ILSA [3] and Cultural Dimensions [4] are used for data analysis. Linear regression model is used to examine relation between two scores IRT and PC of ILSA dataset. The proposed methodology shows how the culture affects Educational Assessment Technology (EdAT). Further it emphasizes the Educative Relationship Triangle where Learner, Instructor and Knowledge are the main components.

A survey is made on students pursuing Higher Education to analyse the distribution of preferences across 109 subjects in Chapter 5. Data analytics methodologies addresses three main issues ie. preferences across different subjects, gender segmentation and socio economic status of the students. Data is collected through questionnaires and interviews. Results are presented through various charts and graphs.

Chapter 6 introduces a multimodal data analysis of human behavioral data from the big data perspective. Authors discuss about the challenges of analyzing multimodal data and create a sample data set from big data. Study focuses on silent and breathe pause during collaborative group dialogs. MULTISIMO [5] corpus was used for experimental purpose. Various data analysis such as extraction of pause, speech pause frequency, silence intervals and dialog topics, language effects on pause duration are presented.

Knowledge from bio-medical databases is mined and in this context two case studies are highlighted in Chapter 7. First case study reports to detect food adulteration using various data mining tools. Initially relevant features are extracted from the data set and reduced data is used to form clusters using two clustering approaches: k-means and Expectation Minimization (EM). Later kernel-based and random forest classifiers are applied on the result. Second case study provides a framework for semantic medical process mining where the framework exploits a knowledge-based abstraction of event log traces in the medical field of stroke management.

Chapter 8 proposes a plan to bridge the gap between society and technology using natural language processing (NLP) which is a major part of artificial intelligence (AI). At the outset authors explain the importance of AI in society and how NLP has wide scope in this domain. Authors suggest to enhance NLP research using various means such as corpus driven view of language, reproducibility study, software protototypes, open source software. The proposal follows the philosophy of agile development in the life cycle and demands the collaboration of research community, non-expert community and general public. Special emphasis is put on the choice of a strategy for knowledge integration.

Chapter 9 focuses on humans face-to-face communication through audio (speech) and gesture modalities. A brief presentation on multimodal communications makes readers familiar about the current studies and connection with big data. In this context authors explain multimodal corpora and processing big data using various machine learning tools. Authors also highlight principles and rules of using personal data downloaded from internet.

Chapter 10 proposes a web based application to detect and recognize spontaneous emotions using electroencephalogram (EEG) signals. Preparation of EEG data of different emotions is explained in introduction.

The application adopts cloud based model and the components of the system are elaborated in detail. Structure of metadata is presented in the form of relational model. Java Enterprise solution is used to develop web application and roles, privileges, access rights are clearly mentioned among different users. Computational module of the system follows Hadoop [6] framework and Map/Reduce [7] approach to perform distributed and parallel processing on large data sets in a distributed environment

Chapter 11 is related to the study and development of systems and devices that can recognize, interpret, process, and simulate human affects called affective computing. Since communication is influenced by Human-Human-Interaction (HHI) and Human-Computer-Interaction (HCI) authors introduce the concept of enriched data to capture human affective state. In this context art of corpora collection and recognition techniques are explained. Two datasets LMC [8] and iGF [9] are used for experimental purpose. Corpus of both the datasets is clearly explained. Socio-demographic features are also considered to improve the performance of disposition modeling. The necessity of multimodal investigations of disposition is indicated which then will be heading towards an improvement of overall performance.

Authors propose a system where data is collected from online social media and deep learning techniques are applied to extract meaningful representation from the collected dataset in Chapter 12. CAS^{2}T [10] toolkit is used for data collection and its network analyser component identifies related multimedia data from YouTube. Then the video clips are sent to crowdsourcing platform iHEARu-PLAY [11] for annotation. Trustability-based Dynamic Active Learning (TDAL) algorithm is used for this work. Series of experiments are conducted by authors to evaluate the quality of collected data and analyse the performance of machine learning approach.

Chapter 13 throws an open discussion on shortcomings of data driven approaches to dialogue management. In this context, components of conversational agent architecture are explained in detail. Authors throw a light on how large scale automatic production of dialogue corpora lacks external criteria. Hidden Markov Model (HMM) and Neural Network based approach are heavily criticized and compared for dialogue management. The authors' suggestions on how to choose the right method are likely to be of outstanding importance for most readers.

Indeed, the book identifies a number of holistic approach to create knowledge and how mined data leads to innovation. Extensive references are provided at the end of each chapter, presenting a roadmap for readers who want to learn more from the literature. The book has several good features that I found very helpful. Every chapter begins with a section of "Chapter Objectives" and ends with a section of "References for Further Study". The chapter objectives are listed roughly in parallel with the sections in each chapter, outlining the purpose of each section and serving as an overview of the main content that the reader is expected to grasp. Algorithms are illustrated in pseudo-code and are easy to be translated into concrete programming languages. This may be especially helpful to data mining practitioners.

3. Concluding remark

The contents of the book are oriented towards applied research. But all the research methods are not confined to any particular area or discipline of research. In view of this fact, the book can widely be referred in all kinds of courses and researches. Every chapter of the book contains case illustration so that readers can better appreciate how big data analytics is deployed. All chapters follow a common format: learning objectives, text, case illustrations, summary, future work and reference. Each chapter has a references (or Bibliography) section pointing to a great mass of literature. Exhaustive references after every chapter made the book unique. As for the presentation, the book is the most readable and its narration is lively.

Overall, this is a good book that could benefit big data mining researchers, practitioners, and anyone who wants to learn something about big data mining and embedded knowledge. Above all the book fulfils its purpose of making innovations in embedded knowledge applicable in a very efficient manner.

References

[1] Blackler F. Knowledge, knowledge work and organization: An overview and interpretation, Organization Studies. 1995; 16(6): 1021.

[2] Gamble PR, Blackwell J. Knowledge Management: A State of the Art Guide, Kogan Page, London, 2001.

[3] Ganimian AJ, Koretz DM. Dataset of International Large-Scale Assessments. Harvard Graduate School of Education, Cambridge, MA, Last updated, 8 Feb 2017.

[4] Hofstede G, Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations: Sage Publications, Thousand Oaks, CA, 2001.

[5] https://www.scss.tcd.ie/clg/MULTISIMO.

[6] Hadoop.apache.org: “Welcome to Apache^{TM} Hadoop$^{®}$!” (2015). [Online]. https://hadoop.apache.org.

[7] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters, Communications of the ACM. 2008; 51(1): 107.

[8] Rösner D, Frommer J, Andrich R, Friesen R, Haase M, Kunze M, Lange J, Otto M. LAST MINUTE: a novel corpus to support emotion, sentiment and social signal processing, In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC), 2012, pp. 82-89.

[9] Tornow M, Krippl M, Bade S, Thiers A, Siegert I, Handrich S, Krüger J, Schega L, Wendemuth A. Integrated health and fitness (iGF)-corpus – ten-modal highly synchronized subject-dispositional and emotional human machine interactions, In Proceedings of Multimodal Corpora: Computer vision and language processing, 2016, pp. 21-24.

[10] Amiriparian S, Pugachevskiy S, Cummins N, Hantke S, Pohjalainen J, Keren G, Schuller B, CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms, In Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII), 2017, pp. 340-345.

[11] https://www.ihearu-play.edu.

Jhimli Adhikari
Associate Professor
Narayan Zantye College
Affiliated to Goa University, Goa, India
E-mail: jhimli_adhikari@yahoo.co.in