IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON UTILITY PATTERN MINING: THEORETICAL ANALYTICS AND APPLICATIONS

Digital Object Identifier 10.1109/ACCESS.2021.3051446

EDITORIAL

IEEE ACCESS SPECIAL SECTION EDITORIAL: UTILITY-PATTERN MINING: THEORETICAL ANALYTICS AND APPLICATIONS

Utility-pattern mining in data has received a lot of attention from the knowledge discovery in database (KDD) community due to its high potential for many applications such as finance, biomedicine, manufacturing, e-commerce, and social media. Current research in utility-pattern mining primarily focuses on discovering patterns of high value (e.g., high profit) in large databases and analyzing/learning important factors (e.g., economic factors) in a data mining process. One of the most popular applications of utility mining is the analysis of large transactional databases to discover high-utility itemsets, which are sets of items that yield a high profit when purchased together.

This Special Section was organized to bring together academic and industrial researchers and practitioners from the data mining, machine learning, and other interdisciplinary communities, in a collaborative effort to identify and present major technical challenges, recent results, and potential topics in the emerging field of utility-pattern mining, especially focusing on theoretical analytics and applications.

Eighteen papers were submitted to this Special Section, 13 of which were accepted for publication. All the submissions were peer-reviewed by at least two researchers to ensure the quality of the publication. The accepted articles in this Special Section are briefly introduced below.

In the article ''An effective approach for the diverse group stock portfolio optimization using grouping genetic algorithm,'' by Chen *et al.*, the authors develop an algorithm for dealing with diverse group stock portfolio optimization (DGSPO). The proposed algorithm is based on the grouping genetic algorithm (GGA), with chromosome representation and fitness function designed for the purpose of finding a suitable DGSP. In summary, this article presents an application of utility-pattern mining.

The article ''Efficient discovery of weighted frequent neighborhood itemsets in very large spatiotemporal databases,'' by Kiran *et al.*, presents a novel model of weighted frequent neighborhood itemsets that may be found in a spatiotemporal database. An efficient algorithm is also proposed to find all weighted frequent neighboring itemsets in a database. In summary, this article extends utility-pattern mining to spatiotemporal databases to help users find hidden information in very large databases.

The article ''Research on urban resident activity patterns and hotspot area based on GPS floating car data,'' by Li *et al.*, studies the hot spots of urban residents' activities and enforces effective planning and decision-making for urban and traffic departments. An improved Fuzzy c-means algorithm is proposed, which adopts adaptive distance norm and adds its own norm induction matrix to each cluster to ensure global optimization. In summary, this article takes the time-series data problem into clustering analysis, to reveal relevant utility patterns in a clustering model for urban planning and commercial layouts.

In the article ''An efficient method for mining closed potential high-utility itemsets,'' Vo *et al.* present a novel and efficient method, named CPHUI-List, to mine closed potential high-utility itemsets (CPHUIs) from uncertain databases without generating candidates. The article proposes a novel type of pattern named CPHUI, along with a data structure named PEU-List, for mining CPHUIs. A pruning strategy named Pr-Prune is proposed to prune the search space and reduce the cost of database scans by utilizing the proposed PEU-List. Based on the proposed PEU-List, Pr-Prune strategy, an effective algorithm named CPHUI-List algorithm is developed to directly mine CPHUIs in uncertain databases. In summary, this article considers the issue of closed utility-pattern mining in an uncertain environment.

In the article ''Efficient approach for damped window-based high utility pattern mining with list structure,'' by Nam *et al.*, the authors propose a novel list-based algorithm to mine high utility patterns while considering the arrival time of each transaction in an incremental database environment. The designed algorithm efficiently performs pattern pruning by using a damped window model that considers previously received data to be less important than recently inserted data and identifies high-utility patterns. In summary, the suggested mining techniques are about utility-pattern mining using a damped window model which is useful for applying utility-pattern mining to streaming data.

In the article ''Wireless transmitter identification based on device imperfections,'' by Li *et al.*, the authors analyzed the generation mechanism of the RF fingerprint and constructed a mathematical model of the transmitted signal. A real communication transmitter link with different internal

components was established to acquire the signals and study the effect of different components on the characteristics of the transmitter. The spectral feature and the permutation entropy feature are extracted as RF fingerprints based on the actual inherent nonlinear dynamical characteristics of the transmitter. In summary, this article applies utility-pattern mining task to wireless device identification.

In the article ''Activity pattern mining for healthcare,'' by Jin *et al.*, the authors focus on discovering patterns containing useful information about human health and analyzing/learning important factors in a data mining process. First, wireless sensing technology is used to collect large volumes of experimental data to address real-world problems. These complex data contain rich environmental information and human activity pattern information. By analyzing the channel information of wireless signals, key features of the activity are discovered. Then, activity patterns are established and different kinds of algorithms are used to classify the data. In summary, this article extends utility-pattern mining to a specific medical task to identify useful activity patterns.

In the article ''Incrementally updating the discovered high average-utility patterns with the pre-large concept,'' by Wu *et al.*, the authors utilize the pre-large concept to update the discovered high average-utility itemsets (HAUIs) in newly inserted transactions and reduce the time of the database rescanning process. To further improve the performance of the developed algorithm, two new upper-bounds are also proposed to decrease the number of candidates for HAUIM. An equation is also specified to ensure that an additional database scan is not unnecessarily performed, and a linked-list structure is utilized in the designed algorithm to ensure that each transaction is scanned at most one time, thus reducing the number of database scans in the maintenance progress. In summary, this article extends utility-pattern mining to consider the concept of average utility in incremental databases.

In the article ''Visualization of generic utility of sequential patterns,'' by Wiktorski *et al.*, the authors present an approach to generating utility bitmaps that provide a visual representation of the numeric data obtained using generic pattern utility algorithms. Based on the presented validation in the context of physical activity monitoring, utility bitmaps allow for immediate separation of various physical activities and indicate age and fitness differences between the participants, even though this information was not available to the algorithm. The approach seems to be promising for exploratory analysis of large collections of long time series and possibly other sequential patterns such as distance series common in sports data analysis, and depth series common in petroleum engineering. In summary, this article contributes to the body of knowledge in applications of utility pattern mining. In particular, for utility pattern mining in large data sets and visualization techniques.

The article ''A data-driven approach for Twitter hashtag recommendation,'' by Belhadi *et al.*, deals with the hashtag recommendation problem and proposes a high average-utility pattern mining-based solution to derive the recommended hashtags from a large collection of tweets. The framework called PM-HRec (Pattern Mining for Hashtag Recommendation) first collects tweets and converts them into a transactional database while considering the temporal information of the tagged tweets. The temporal top-k high average utility patterns are then discovered. The utility patterns, ontology, and irrelevantly tagged tweets are used to extract the most relevant hashtags for a given orpheline tweet. In summary, in this article, a data-driven solution for hashtag recommendation is developed by integrating high utility pattern mining in the searching process.

The article ''A multi-core approach to efficiently mining high-utility itemsets in dynamic profit databases,'' by Vo *et al.*, handles dynamic profit transaction databases to discover high-utility patterns. An algorithm called MCH-Miner is then presented to utilize the computing power of the currently available multi-core processor architecture by adapting the iMEFIM algorithm to be able to efficiently mine HUIs from dynamic profit databases. In summary, this article develops a new model for utility-pattern mining, which considers dynamic profit tables as an extension.

In the article ''PrefixSpan-based pattern mining using time sliding weight from streaming data,'' by Kang *et al.*, the authors propose a prefixSpan-based pattern mining algorithm using time sliding weight from streaming data. It applies a time sliding weight to create a structure called projected DB Tree. A time window is applied to the sequential data to calculate the label and support of the window. When a projected DB tree is created, the time weight calculated for each pattern is inserted in a table. At this time, the tree is updated by deleting the node whose time weight is less than the reference value. For this reason, whenever data is updated, the tree is sorted again. The reordering process removes the pattern having less influence by applying time weights. Therefore, it is possible to construct a projected DB Tree that can extract influential patterns. In summary, this article discusses the generic pattern mining progress regarding the stream and sequence data.

In the article ''Proof learning in PVS with utility pattern mining,'' by Nawaz *et al.*, the authors present a proof process learning approach for the PVS (Prototype Verification System) proof assistant. The approach is based on high-utility itemset mining (HUIM), focusing on finding not only frequent proof steps/patterns in the proofs but also proof steps/patterns with high importance (utility). In the proposed approach, the PVS proofs for theories were converted into a proof corpus that was suitable for learning. HUIM techniques were applied on the corpus to find frequent proof steps/patterns with high utility, which are used in the proofs. Moreover, relationships between proof steps/patterns were discovered through sequential rule mining. Some interesting proof patterns were found with HUIM, and the obtained results showed that the number of proof steps in each proof and the utility value assigned to each proof command have

a direct correlation to the efficiency of HUIM algorithms. In summary, this article devises a utility-pattern mining based approach for proof process learning to derive frequent proof steps of high importance.

In conclusion, the Guest Editors would like to thank all the authors who submitted their research articles to this Special Section. The Guest Editors highly appreciate the contributions of the reviewers for their constructive comments and suggestions. The Guest Editors would also like to thank the guidance from IEEE Access Editor-in-Chief and staff members. The Guest Editors hope that this Special Section will benefit the scientific community and contribute to the knowledge base, particularly in the field of pattern analytics and mining.

JERRY CHUN-WEI LIN, *Guest Editor*
Department of Computer Science, Electrical Engineering and Mathematical Sciences
Western Norway University of Applied Sciences
5063 Bergen, Norway

PHILIPPE FOURNIER-VIGER, *Guest Editor*
School of Humanities and Social Sciences
Harbin Institute of Technology, Shenzhen
Shenzhen 518055, China

VINCENT S. TSENG, *Guest Editor*
Department of Computer Science
National Chiao Tung University
Hsinchu 30010, Taiwan

PHILIP S. YU, *Guest Editor*
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA

JERRY CHUN-WEI LIN (Senior Member, IEEE) is currently working as a Full Professor with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. He has published more than 400 research papers in refereed journals and international conferences. He is also the Leader of the well-known SPMF Project. His research interests include data mining, soft computing, artificial intelligence, social computing, and privacy-preserving and security technologies. He is also a Senior Member of ACM and a Fellow of IET (FIET). Moreover, he had been awarded as the Most Cited Chinese Researcher in 2018 and 2019 by Elsevier. He is also the Editor-in-Chief of *Data Science and Pattern Recognition (DSPR)* journal, an Associate Editor/Editor/Guest Editor of the *Journal of Internet Technology*, IEEE Access, *PLOS One*, *Intelligent Data Analysis*, *International Journal of Interactive Multimedia and Artificial Intelligence*, *ACM Transactions on Management Information Systems*, IEEE Transactions on Industrial Informatics, IEEE Transactions on Fuzzy Systems, and *ACM Transactions on Management Information Systems*.

PHILIPPE FOURNIER-VIGER is currently a Full Professor with the Harbin Institute of Technology, Shenzhen, China, where he is also a National Talent. He has participated in more than 280 research articles, which have received more than 6100 citations. His research interests include data mining, algorithm design, pattern mining, sequence mining, sequence prediction, itemset mining, graph mining, big data, and applications. He is also an Associate Editor-in-Chief of the *Applied Intelligence* journal, the Co-Editor-in-Chief of the *Data Science and Pattern Recognition (DSPR)* journal, an Associate Editor of the *Array* journal, and the Founder of the popular SPMF data mining library, which offers more than 180 algorithms, and has been used in more than 850 research articles since 2010. He has been Keynote Speaker for several conferences, such as ICGEC 2018, MIWAI 2019, CCNS 2020, and IFID 2020.

VINCENT S. TSENG (Fellow, IEEE) received the Ph.D. degree with a major in computer science from National Chiao Tung University, Taiwan, in 1997. He is currently a Distinguished Professor with the Department of Computer Science, National Chiao Tung University. He has published more than 400 research papers in peer-reviewed journals and conferences and holds 15 patents. His research interests include data mining, big data, biomedical informatics, and mobile and Web technologies. He is also a Senior Member of ACM. He was a recipient of the 2014 K. T. Li Breakthrough Award and the 2015 Outstanding Research Award from the Ministry of Science and Technology Taiwan. He has served as chairs/program committee members for several premier international conferences related to data mining and machine learning, including KDD, ICDM, SDM, PAKDD, ICDE, CIKM, and IJCAI. He has also served as the Chair for IEEE CIS Tainan Chapter from 2013 to 2015, and as the President of the Taiwanese Association for Artificial Intelligence from 2011 to 2012. He has been on the editorial board of a number of journals, including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Knowledge Discovery from Data*, and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS.

PHILIP S. YU (Life Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, the M.S. and Ph.D. degrees in electrical engineering from Stanford University, and the M.B.A. degree from New York University. He is currently a Distinguished Professor of computer science with the University of Illinois at Chicago (UIC) and holds the Wexler Chair of information technology with UIC. Before joining UIC, he was with IBM, where he was a Manager of the Software Tools and Techniques Department, Thomas J. Watson Research Center. He has published more than 1300 papers in peer-reviewed journals [i.e., IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS (TPDS), *ACM Transactions on Knowledge Discovery from Data (TKDD)*, and *The VLDB Journal—The International Journal on Very Large Data Bases (VLDBJ)*] and conferences (i.e., SIGMOD, KDD, ICDE, WWW, AAAI, SIGIR, and ICML, etc.) He holds or has applied for more than 300 U.S. patents. His research interests include databases, data mining, artificial intelligence, and privacy. He is also a Fellow of ACM. He received the ACM SIGKDD 2016 Innovation Award and the IEEE Computer Society 2013 Technical Achievement Award. He was the Editor-in-Chief of *ACM Transactions on Knowledge Discovery from Data* from 2011 to 2017 and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING from 2001 to 2004.