Transactions on Data Analysis in Social Science 1(4) (2019) 186–195

ISSN Online: 2821-1936

Transactions on Data Analysis in Social Science

Journal Homepage: https://transoscience.ir

Basket Purchase Prediction Using Association Rule Mining Based on the FP-Growth Algorithm

S. Dami[1,*], A. Moraveji[2], A. Ghasemnejad[3]

[1] Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Islamic Azad University, Tehran West Branch, Tehran, Iran
[2] Instructor, Department of Computer Engineering, Faculty of Engineering, Islamic Azad University, Tehran West Branch, Tehran, Iran
[3] M.Sc. Student in Information Technology, Faculty of Engineering, Islamic Azad University, Tehran West Branch, Tehran, Iran

ARTICLE INFO

Article History:
Received 29 May 2019
Received in revised form 5 July 2019
Accepted 29 October 2019
Available online 3 December 2019

Keywords:
Data Mining; Basket Purchase Prediction; Association Rule Discovery; Frequent Pattern Growth

ABSTRACT

Market basket analysis helps marketing analysts understand customer behavior such as identifying which products are frequently purchased together. Various data mining techniques and algorithms have been developed to perform such analyses. The present study introduces an innovative approach that applies the FP-Growth algorithm to discover associations among users' purchases in order to enhance the efficiency of e-commerce systems. In the proposed method, all user transactions are utilized in the basket analysis process. In other words, even purchases that appear unrelated to the user's current transactions can provide valuable information, contributing to a deeper understanding of customer purchase patterns and improving the overall performance of sales systems. To evaluate the effectiveness of the proposed method, its results were compared with those of the Eclat and Apriori algorithms. Experimental analyses revealed that, on average, the proposed method outperformed the compared approaches.

1. INTRODUCTION

Data mining refers to the extraction of useful knowledge from large volumes of data. Market basket analysis (MBA) is one of the key data mining techniques used to discover associations among datasets. Association rule mining identifies relationships within large sets of data items [6]. As massive amounts of data are continuously collected and stored in databases, many industries have become increasingly concerned with uncovering meaningful association rules from their data repositories [7].

For example, discovering interesting relationships among vast amounts of commercial transaction data can significantly aid in catalog design, cross-marketing, and various decision-making processes in business. A typical

* Corresponding Author: dami@wtiau.ac.ir
Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Islamic Azad University, Tehran West Branch, Tehran, Iran

http://dx.doi.org/10.47176/TDASS.2019.186

example of association rule mining is *market basket analysis* [8]. This approach examines customer purchasing patterns by identifying associations among products frequently placed together in shopping baskets. Recognizing such associations enables retailers to develop more effective marketing strategies by gaining insights into the products commonly purchased together by customers. Consequently, this facilitates a better understanding of consumer purchasing behavior and contributes to increased sales performance. Furthermore, this area of research provides a broad and valuable field for developing improved data mining algorithms [9].

Market basket analysis (MBA), also referred to as *association rule learning* or *dependency analysis*, is a data mining technique [10] that can be applied in various domains such as marketing, bioinformatics, education, and nuclear science. The primary objective of MBA in marketing is to provide retailers with valuable information to understand buyer behavior, thereby assisting them in making more effective business decisions [16].

As industries continue to accumulate vast quantities of transactional and operational data in their databases, the discovery of hidden relationships among these data items becomes increasingly important. Identifying such associations in large-scale datasets can aid in designing catalogs, implementing cross-selling strategies, and enhancing decision-making processes across business operations. Market basket analysis, as a prominent example of association rule mining, serves as an effective tool for analyzing customer purchasing behavior and identifying product associations [17]. Such insights can help retailers expand their marketing strategies by understanding which products are frequently purchased together, ultimately leading to improved customer targeting and sales performance [18].

2. LITERATURE REVIEW

Today, a vast amount of data is stored in databases across various fields such as retail markets, banking, healthcare, and other industries [12]. However, not all available data are necessarily useful to end-users. Therefore, extracting valuable and meaningful information from massive datasets has become essential. This process, known as *data mining* or *Knowledge Discovery in Databases (KDD)*, aims to identify hidden patterns and relationships within large datasets [21].

The general process of finding and interpreting patterns in data involves multiple stages, including selection, preprocessing, transformation, data mining, and interpretation [13]. In business contexts, data mining supports marketing strategies, for instance, through *market basket analysis (MBA)*, a method widely employed in management research (e.g., Aguinis et al.). Market basket analysis, also referred to as *association rule mining*, assists marketing analysts in understanding customer behavior specifically, which products are frequently purchased together [22]. Various techniques and algorithms have been developed to perform data mining effectively [23].

The identification of frequent patterns plays a crucial role in uncovering dependencies, correlations, and significant associations among large datasets [15]. Frequent pattern mining seeks recurring itemsets within given data, which helps identify relationships among items in large transactional or relational datasets. A well-known example of frequent itemset mining is market basket analysis [14].

These patterns may include itemsets, subsequences (e.g., purchasing a PC, then a digital camera, and subsequently a memory card if this sequence frequently occurs in historical sales data), and substructures, which can take various structural forms such as subgraphs, subtrees, or subnetworks. In general, association rule mining involves two main steps:

1. Identifying all frequent itemsets, each satisfying a minimum support threshold.
2. Generating strong association rules from these frequent itemsets that meet both minimum support and minimum confidence levels.

A rule that satisfies both criteria is referred to as a *strong rule*. There are several methods for mining frequent patterns, including Boolean association rule mining, which deals with dependencies among items based on their presence or absence in transactions.

Hiroshi Ishikawa and colleagues proposed two complementary web-based methods to address this challenge. First, they introduced an adaptive recommender system called the L-R System, which integrates user modeling with web page classification. By analyzing user access logs, this system identifies access patterns and recommends web pages based on transition probabilities and relationships between user models and web structures. Second, they developed a clustering-based approach for constructing descriptive user models by grouping website access logs according to users' browsing behaviors. Such user models can also be utilized to detect unexpected navigation paths resulting from poor website design [3].

Ashok Kumar and his collaborators demonstrated that automated data collection which generates vast amounts of web access information can be efficiently processed using a new algorithm called K-APRIOR. This algorithm effectively identifies frequently accessed web pages within large-scale binary weblog datasets [2]. Web mining, in general, focuses on extracting information from websites worldwide. Traditional web search engines often ignore user preferences when delivering results. To address this limitation, Kumar and his team applied *web content mining* techniques for personalization. Through personalization, web access and content retrieval can be tailored to users' needs, potentially including the generation of user-specific web pages or retrieval of documents based on individual interests. Their work also analyzed the PageRank algorithm for web mining applications based on user preferences [1].

Murat and colleagues introduced a hybrid recommender system that combines the outputs of multiple web usage mining–based recommender techniques. With the rapid expansion of the internet, finding useful information online has become a critical challenge. Web recommender systems assist users in navigating this complex informational space by identifying the most relevant sources. Recently, the number of web-based recommender systems has increased substantially, focusing on analyzing user navigation behavior and predicting subsequent page requests. Each system has its own criteria and limitations; however, Murat's hybrid approach combines several recommendation techniques to produce a unified set of recommendations for new users. Comparative evaluations of four combination methods demonstrated how hybrid techniques improve prediction accuracy. Further improvements were proposed based on these results, confirming that hybrid recommender systems outperform single-method models in predicting user requests [4].

Zihayat et al. employed Big Data analytics to enhance news recommendation systems. They implemented a two-stage approach for recommending news and online articles, achieving significant improvements in accuracy [5]. Similarly, Zhao et al. (2019) proposed a method called OPE for market basket analysis in the Chinese food industry. Through empirical evaluations of customer purchase recommendations, their approach demonstrated favorable outcomes [13].

3. PROPOSED METHOD

Association rule mining enables organizations such as government agencies, retail stores, and banks to discover and predict future patterns and behaviors by analyzing data stored in their systems [11]. By examining past transactions, association rule mining can forecast future customer or system behavior and provide answers to questions that were previously difficult or time-consuming to address [12].

In this study, the use and discovery of relationships among users' different purchase patterns to improve the performance of e-commerce systems through the FP-Growth algorithm represent one of the main innovations. In the proposed method, all customer purchases are considered in the shopping basket analysis process, meaning that even purchases seemingly unrelated to the user's current selections can still contain valuable information. Such data may contribute to more accurate customer basket analysis and enhance overall sales system performance.

This research was conducted using a library-based approach, where data were collected from well-known digital repositories, scientific websites, and published academic papers. Additionally, part of the data was derived from the results of implementing the proposed process model. The study employed an experimental approach, using publicly available datasets on the Internet. After collecting the required data, data preprocessing was performed to prepare them as inputs for the system. Once the data were refined, association rules related to users' shopping baskets were extracted. The general workflow of the proposed research is illustrated in Figure 1.

During the data preprocessing phase, all input data were carefully examined. Duplicate records were removed, incomplete data (i.e., records containing null values) were excluded, and any low-quality or inconsistent data were eliminated to ensure data reliability and integrity.

The FP-Growth algorithm, employed in this study to identify association rules, was originally introduced by *Han et al.* It is a scalable and efficient method for mining a complete set of frequent patterns through a process called *pattern fragment growth*, utilizing an extended prefix-tree structure known as the Frequent Pattern Tree (FP-Tree). This tree structure stores essential compressed information about frequent patterns in a highly compact form.

The popularity and effectiveness of the FP-Growth algorithm have led to numerous studies proposing variations aimed at improving its performance. FP-Growth simplifies most of the challenges associated with traditional association rule mining methods by using the FP-Tree structure. In an FP-Tree, each node represents an item along with its current count, while each branch corresponds to a distinct association among items.

The pseudocode of the FP-Growth algorithm used in this study is presented in Figure 2.

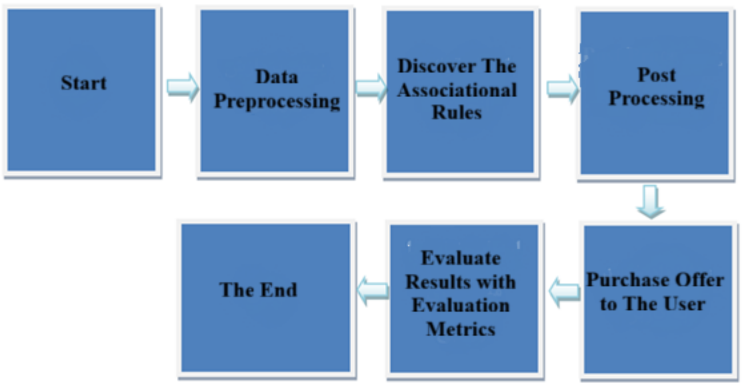

Fig. 1. Process of the Proposed Method

The FP-tree is mined by calling FP_growth(*FP_tree*, *null*), which is implemented as follows.

```
procedure FP_growth(Tree, α)
    if Tree contains a single path P then
        for each combination (denoted as β) of the nodes in the path P
            generate pattern β ∪ α with support_count = minimum support count of nodes in β;
    else for each a_i in the header of Tree {
        generate pattern β = a_i ∪ α with support_count = a_i.support_count;
        construct β's conditional pattern base and then β's conditional FP_tree Tree_β;
        if Tree_β ≠ ∅ then
            call FP_growth(Tree_β, β); }
```

The FP-growth algorithm for discovering frequent itemsets without candidate generation.

Fig. 2. Pseudocode of the FP-Growth Algorithm

The advantages of the proposed method are summarized as follows:

1. The method does not fragment long transaction patterns.
2. It preserves complete information required for mining frequent patterns.
3. It effectively eliminates irrelevant or infrequent items from the dataset.
4. Items are arranged in descending order of frequency, increasing the likelihood of shared patterns.

5. The method does not expand the transaction set beyond the size of the original database.
6. It operates faster than the Apriori algorithm.

During the post-processing phase, the extracted association rules are refined to enhance their quality. Rules that appear trivial or self-evident are removed. Additionally, rules that do not meet the desired support or confidence thresholds are filtered out, ensuring that only high-quality and meaningful association rules remain for further analysis.

4. EVALUATION OF RESULTS

4.1. Datasets

The datasets used to evaluate the quality of the extracted association rules were selected according to the following criteria:

- A total of five datasets were selected from a larger pool of available data sources. These datasets are denoted as D1 to D5, and the following procedures were applied to each of them.
- Each dataset contained 500 transactions, which were used to assess the quality of the association rules between items.
- Each transaction was required to contain at least three items.
- A total of 200 two-item sets were appropriately distributed among the transactions, ensuring a balanced relationship between intra-transaction and inter-transaction itemsets.
- A total of 100 three-item sets were also evenly distributed among the datasets to maintain a balance between intra-transactional and inter-transactional associations. These three-item sets could partially overlap with some of the two-item sets.
- Additionally, 50 unrelated three-item sets were distributed across the transactions to ensure proper balancing between intra-transactional and inter-transactional patterns. Some degree of overlap between these sets and the two-item sets was allowed.

4.2. Evaluation Method

To evaluate the proposed approach, several performance metrics were used, as described below.

4.2.1. Rule Accuracy

This metric measures the correctness of the generated association rules and is calculated as follows:

$$Acc(H \leftarrow B) = p(H|B) \tag{1}$$

This metric indicates how accurately the extracted rules reflect the true relationships among items in the dataset.

$$Err(H \leftarrow B) = 1 - Acc(H \leftarrow B) = p(\overline{H}|B) \tag{2}$$

In the above equation, the conditional probability of H given B is used. The underlying concept is that the rule accuracy represents the proportion of itemsets in which, if B occurs, H also occurs.

Formally, this corresponds to the ratio of True Positives (TP) to the total number of Positives (P).

Based on this concept, to assess the validity of the extracted association rules, the error rate can be defined as follows:

$$Err(H \leftarrow B) = 1 - Acc(H \leftarrow B) = p(\overline{H}|B) \tag{3}$$

This measure indicates the proportion of incorrectly predicted associations among the evaluated rules.

4.2.2. Negative Reliability

The purpose of this metric is to determine the proportion of instances that actually belong to the negative class and have been correctly identified as such. In other words, it measures how reliably the system can recognize negative instances. The formal definition of Negative Reliability is given as follows:

$$NegRel(R) = \frac{TN}{TN+FN} = \frac{TN}{N_p} = \frac{n(\overline{HB})}{n(\overline{B})} = \frac{p(\overline{HB})}{p(\overline{B})} = p(\overline{H}|\overline{B}) \tag{4}$$

This metric reflects the accuracy of the system in identifying non-associated or irrelevant itemsets within the dataset.

4.2.3. Sensitivity

This metric is equivalent to Recall, a commonly used measure in Information Retrieval (IR) and Classification tasks. Sensitivity evaluates how effectively the association rules respond to true conditions in other words, it measures the probability that the antecedent of a rule is correctly identified given that its consequent holds true.

Put differently, this metric assesses the accuracy of the rule $B \leftarrow H$, which is the inverse of the rule $H \leftarrow B$. The mathematical expression of Sensitivity is presented as follows:

$$Sens(R) = \frac{TP}{TP+FN} = \frac{TP}{P_a} = \frac{n(HB)}{n(HB)+n(H\overline{B})} = \frac{n(HB)}{n(H)} = \frac{p(HB)}{p(H)} = p(B|H) \tag{5}$$

This measure indicates how effectively the proposed model identifies valid associations when they truly exist in the dataset.

4.3. Experimental Results

To evaluate the performance of the proposed method, it was compared with two well-known algorithms: Eclat and Apriori.

Table 1 presents the obtained results for each of the three methods across the five datasets. The final evaluation metric was defined as the average of all performance measures introduced in Section 4.2.

Table 1. Evaluation results of the proposed method compared with other approaches

Dataset	Proposed	Eclat	Apriori
D1	97.289	88.448	91.325
D2	97.821	96.939	90.552
D3	90.888	90.392	93.013
D4	51.126	45.269	51.155
D5	78.521	57.347	63.660

Figure 3 illustrates a comparative chart of the obtained results for all three methods across the five datasets.

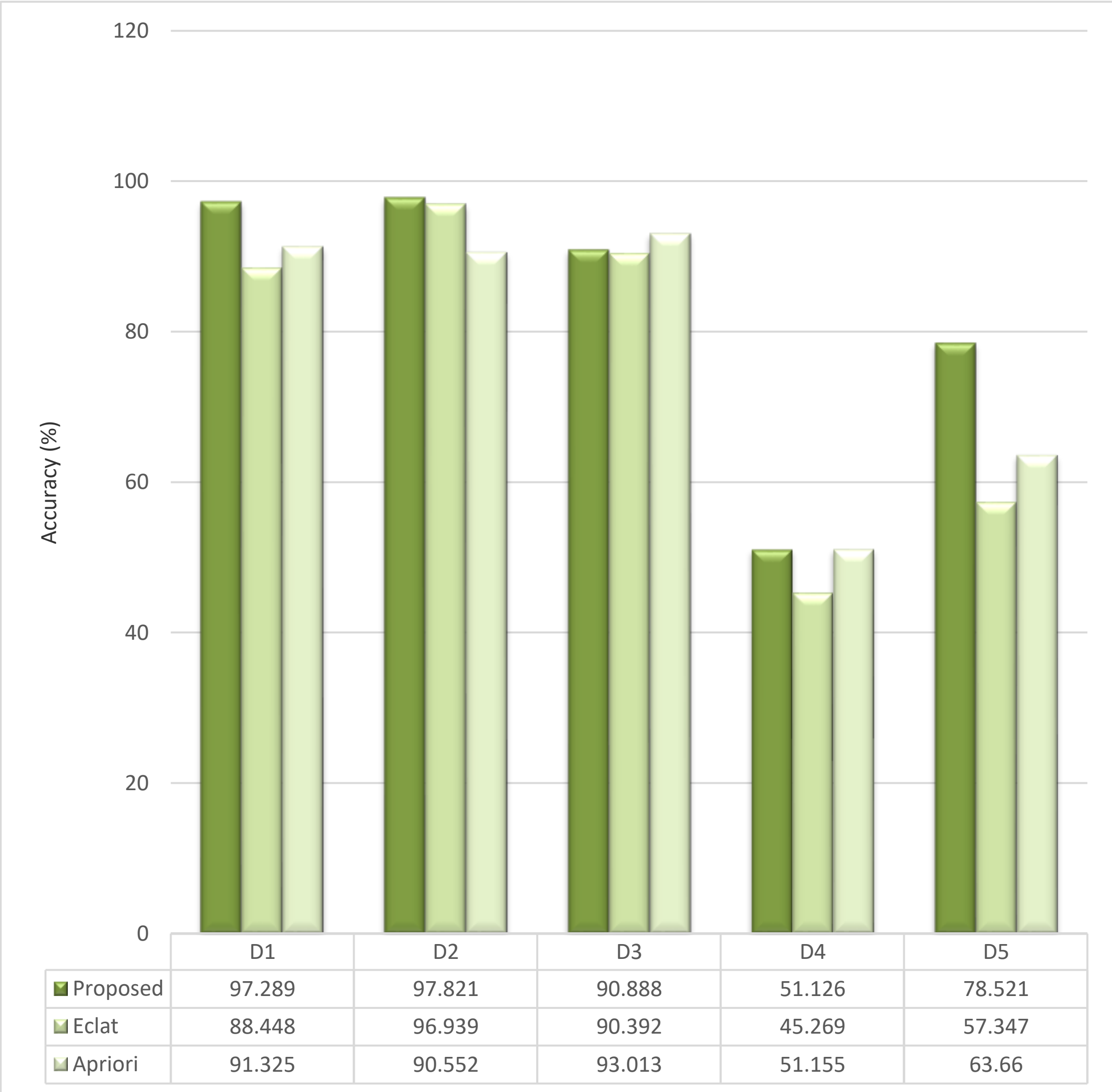

Fig. 3. Comparison of the obtained results for the evaluated methods

As shown in the above figure, for Dataset D1, the Apriori algorithm performed slightly better than Eclat; however, the proposed method outperformed both.

For Dataset D2, Eclat performed marginally better than Apriori, yet the proposed method again achieved the highest performance.

In Dataset D3, the proposed method and Eclat showed very close results, while Apriori performed slightly better than both.

For Dataset D4, the proposed method and Apriori exhibited nearly identical performance, and Eclat performed slightly worse than both.

Finally, for Dataset D5, Apriori outperformed Eclat; nevertheless, the proposed method significantly surpassed both competing algorithms by a considerable margin.

Table 2 summarizes the average performance of the proposed method compared to the two benchmark algorithms. As observed, the proposed method consistently achieved higher overall performance.

Table 2. Average performance of the proposed method compared with benchmark algorithms

Metric	Proposed	Eclat	Apriori
Average	**83.129**	75.679	77.941

5. CONCLUSION

Today, a vast amount of data is stored across various domains such as retail markets, banking, healthcare, and many others. However, not all of this information is necessarily useful to users. Therefore, extracting meaningful and valuable information from large datasets is crucial. This process of discovering useful patterns from data is known as *data mining* or the *Knowledge Discovery in Databases (KDD)* process.

Market Basket Analysis (MBA), also known as *association rule learning* or *dependency analysis*, is a data mining technique that can be applied in diverse fields such as marketing, bioinformatics, education, and nuclear sciences. The main goal of MBA in marketing is to provide retailers with insights into customer purchasing behavior, enabling them to make better-informed decisions.

One of the main innovations of this study lies in utilizing and discovering relationships among different customer purchases to enhance the efficiency of e-commerce systems through the *FP-Growth algorithm*. In the proposed approach, all user transactions can contribute to the basket analysis process. In other words, even purchases that appear unrelated to the user’s current transactions may contain valuable information that supports customer basket analysis and improves the performance of the sales system.

To evaluate the effectiveness of the proposed method, it was compared with the *Eclat* and *Apriori* algorithms. Five datasets, labeled D1 through D5, were selected for this comparison.

- For dataset D1, the Apriori method performed slightly better than Eclat, but the proposed method outperformed both.
- For dataset D2, Eclat slightly outperformed Apriori, while the proposed method achieved the best performance among all.
- For dataset D3, the proposed method and Eclat showed very similar performance, whereas Apriori performed marginally better than both.
- For dataset D4, the proposed method and Apriori demonstrated nearly identical performance, with Eclat performing slightly worse.
- For dataset D5, Apriori outperformed Eclat, but the proposed method achieved a significantly higher performance than both algorithms.

Overall, the proposed approach demonstrated superior average performance compared to the other methods across all datasets.

6. FUTURE WORK

Future research could focus on extending the current study by integrating multiple hybrid algorithms. Additionally, adapting and developing the proposed method for implementation in *parallel processing* or *distributed computing environments* can further enhance its scalability and efficiency.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Aruk Kumar, S., & Niraj, S. (2012). A novel PageRank algorithm for web mining based on user's interest. International Journal of Emerging Technology and Advanced Engineering, 2(9), 1–6. ISSN 2250-2459.

[2] Ashok Kumar, D., & Loraine Charlet Annie, M. C. (2012). Web log mining using K-Apriori algorithm. International Journal of Computer Applications, 41(11), 1–5.

[3] Ishikawa, H., Ohta, M., Yokoyama, S. H., Nakayama, J., & Katayama, K. (2002). Web usage mining approaches to page recommendation and restructuring. International Journal of Intelligent Systems in Accounting, Finance and Management, 11, 137–150.

[4] Murat, G., & Sule, G. (2010). Combination of web page recommender systems. Expert Systems with Applications, 37(4), 2877–2885.

[5] Zihayat, M., Ayanso, A., Zhao, X., Davoudi, H., & An, A. (2019). A utility-based news recommendation system. Decision Support Systems, 117, 14–27. https://doi.org/10.1016/j.dss.2018.12.001

[6] Abdallah, M. A., & Covaci, A. (2014). Organophosphate flame retardants in indoor dust from Egypt: Implications for human exposure. Environmental Science & Technology, 48(9), 4782–4789. https://doi.org/10.1021/es501078s

[7] Ali, N., Dirtu, A. C., Van den Eede, N., Goosey, E., Harrad, S., Neels, H., Mannetje, A. T., Coakley, J., Douwes, J., & Covaci, A. (2012). Occurrence of alternative flame retardants in indoor dust from New Zealand: Indoor sources and human exposure assessment. Chemosphere, 88(11), 1276–1282. https://doi.org/10.1016/j.chemosphere.2012.03.100

[8] Ali, N., Ali, L., Mehdi, T., Dirtu, A. C., Al-Shammari, F., Neels, H., & Covaci, A. (2013). Levels and profiles of organochlorines and flame retardants in car and house dust from Kuwait and Pakistan: Implication for human exposure via dust ingestion. Environment International, 55, 62–70. https://doi.org/10.1016/j.envint.2013.02.001

[9] Bollmann, U. E., Moler, A., Xie, Z. Y., Ebinghaus, R., & Einax, J. W. (2012). Occurrence and fate of organophosphorus flame retardants and plasticizers in coastal and marine surface waters. Water Research, 46(2), 531–538. https://doi.org/10.1016/j.watres.2011.11.028

[10] Brandsma, S. H., de Boer, J., Leonards, P. E. G., Cofino, W. P., & Covaci, A. (2013). Organophosphorus flame-retardant and plasticizer analysis, including recommendations from the first worldwide interlaboratory study. Trends in Analytical Chemistry, 43, 217–228. https://doi.org/10.1016/j.trac.2012.12.004

[11] Brommer, S., Harrad, S., Van den Eede, N., & Covaci, A. (2012). Concentrations of organophosphate esters and brominated flame retardants in German indoor dust samples. Journal of Environmental Monitoring, 14(9), 2482–2487. https://doi.org/10.1039/c2em30303e

[12] Chen, M., Jiang, J., Gan, Z., Yan, Y., Ding, S., Su, S., & Bao, X. (2019). Grain size distribution and exposure

evaluation of organophosphorus and brominated flame retardants in indoor and outdoor dust and PM10 from Chengdu, China. Journal of Hazardous Materials, 365, 280–288. https://doi.org/10.1016/j.jhazmat.2018.10.082

[13] Zhao, L., Jian, K., Su, H., Zhang, Y., Li, J., Letcher, R. J., & Su, G. (2019). Organophosphate esters (OPEs) in Chinese foodstuffs: Dietary intake estimation via a market basket method, and suspect screening using high-resolution mass spectrometry. Environment International, 128, 343–352. https://doi.org/10.1016/j.envint.2019.04.055

[14] Moreno, M. N., Segrera, S., & López, V. F. (n.d.). Association rules: Problems, solutions, and new applications. Universidad de Salamanca.

[15] Kumar, K. P., & Arumugaperumal, S. (2013). Association rule mining and medical application: A detailed survey. International Journal of Computer Applications, 80(17), 1–8. https://doi.org/10.5120/13967-1698

[16] Hao, Z., Wang, X., Yao, L., & Zhang, Y. (2009). Improved classification based on predictive association rules. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2009) (pp. 1165–1170). https://doi.org/10.1109/ICSMC.2009.5345954

[17] Niu, Q., Xia, S. X., & Zhang, L. (2009). Association classification based on compactness of rules. In Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining (WKDD 2009) (pp. 245–247). https://doi.org/10.1109/WKDD.2009.160

[18] Wang, Y. J., Xin, Q., & Coenen, F. (2007). A novel rule weighting approach in classification association rule mining. In Proceedings of the Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007) (pp. 271–276). https://doi.org/10.1109/ICDMW.2007.126

[19] Ehsani Chimeh, H., & Karami, M. (2018). Spam detection from big data based on evolutionary data mining systems. Transactions on Machine Intelligence, 1(1), 1–9. https://doi.org/10.47176/TMI.2018.1

[20] Borhani, S., Mohammadi Zanjireh, M., & Haj Ali Asgari, F. (2019). Determining the factors affecting the incidence of hypertension in pregnant women using data mining techniques. Transactions on Data Analysis in Social Science, 1(2), 59–70. https://doi.org/10.47176/TDASS.2019.59

[21] Bartik, V. (2009). Association-based classification for relational data and its use in web mining. In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2009) (pp. 252–258). https://doi.org/10.1109/CIDM.2009.4938657

[22] Sumithra, R., & Paul, S. (2010). Using distributed Apriori association rule and classical Apriori mining algorithms for grid-based knowledge discovery. In Proceedings of the International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1–5). https://doi.org/10.1109/ICCCNT.2010.5591577

[23] Trnka, A. (2010). Market basket analysis with data mining methods. In Proceedings of the International Conference on Networking and Information Technology (ICNIT) (pp. 446–450). https://doi.org/10.1109/ICNIT.2010.5508476