Analysis of Recent Maximal Frequent Pattern Mining Approaches

Gangin Lee and Unil Yun(✉)

Department Of Computer Engineering, Sejong University, Seoul, Korea
ganginlee@sju.ac.kr, yunei@sejong.ac.kr

Abstract. Since the concept of representative pattern mining was proposed to solve the limitations of traditional frequent pattern mining, a variety of relevant approaches have been developed. As one of the major techniques in representative pattern mining, maximal frequent pattern mining provides users with a smaller number of more meaningful pattern mining results. In this paper, we analyze characteristics of recent maximal frequent pattern mining methods using various concepts and techniques.

Keywords: Data mining · Knowledge discovery · Maximal frequent pattern · Pattern mining · Representative pattern

1 Introduction

The concept of representative pattern mining was proposed to overcome the fatal problems of previous traditional frequent pattern mining such as generating an excessive number of frequent patterns and degrading mining performance. Maximal frequent pattern mining [4, 11, 12], which is one of the major techniques in representative pattern mining, is known for extracting representative patterns more efficiently at the cost of accuracy in pattern restoration. Such an advantage has attracted development of various relevant applications such as bio data analysis [2, 9], uncertain data analysis [5], privacy preserving [6], distributed processing [7], social network analysis [10], and hypergraph dualization [8]. In this paper, we analyze characteristics of recent maximal frequent pattern mining techniques.

The remainder of this paper is as follows. Section 2 introduces the basic concept of frequent pattern mining and its important related works. Section 3 describes recent approaches of maximal frequent pattern mining. Section 4 finally concludes this paper.

2 Frequent Pattern Mining

Since the Apriori algorithm was devised [1], various frequent pattern mining approaches have been proposed. FP-Growth [3] is a tree-based approach that can solve the fatal problems of Apriori such as excessive database scans and candidate pattern creation. Its own tree structure, called FP-tree, and mining techniques have attracted many research attentions. As a result, a variety of variations and applications have been developed. From given databases, frequent pattern mining methods find all of the

J.J. (Jong Hyuk) Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*, Lecture Notes in Electrical Engineering 421, DOI 10.1007/978-981-10-3023-9_135

possible patterns such that their support values are not smaller than user-given minimum support threshold. Frequent patterns can be expressed as the following definitions.

Definition 1. (*Support of a pattern*) A database, $DB = \{T_1, T_2, \ldots, T_n\}$, is composed of multiple transactions, Ts, and each transaction $T_k = \{i_1, i_2, \ldots, i_m\}$ also has a number of items, is. Then, pattern $A = \{i_1, i_2, \ldots, i_j\}$, which can be generated from DB, is a subset of at least one T. The support of A, $Sup(A)$, is calculated as follows:

$$Support(A) = \sum_{i=1}^{n} f(A, T_i), f(A, T_i) = \begin{cases} 1, \text{ if } A \subseteq T \\ 0, \textit{ otherwise} \end{cases} \quad (1)$$

Definition 2. (*Frequent pattern*) A measure, called *minimum support threshold* (denoted as δ), is set by a user and employed to judge whether or not a pattern is frequent. This measure is the product of a user-given percent value and the number of transactions in DB. Then, pattern A is considered as a frequent pattern if the following condition is satisfied:

$$Support(A) \geq \delta \quad (2)$$

Therefore, the main goal of frequent pattern mining is to find all of the possible patterns that satisfy the above condition from a given database.

However, they may extract an excessive number of pattern results depending on features of databases and threshold settings. Since it is difficult to analyze all the mined patterns, we need to consider another approach.

3 Recent Maximal Frequent Pattern Mining Techniques

Maximal frequent pattern mining is a method that mines a smaller number of representative patterns instead of extracting all of the possible frequent patterns. There are two ways to mine representative patterns: closed frequent pattern mining and maximal frequent pattern mining. Although they can extract pattern mining results that can represent frequent patterns, maximal frequent pattern mining can guarantee better pattern condensing effect. Closed frequent pattern mining guarantees complete pattern restoration from closed frequent patterns to original frequent patterns; however, its pattern condensing effect is worse than that of maximal frequent pattern mining. This paper focuses on recent techniques of maximal frequent pattern mining. The maximality feature of a pattern is defined as follows.

Definition 3. (*Maximal frequent pattern*) Let X be a pattern, $\Gamma = \{X'_1, X'_2, \ldots, X'_k\}$ be a set of supersets of X, X's, and δ be a user-given minimum support threshold. Then, X becomes a maximal frequent pattern if the following conditions are satisfied:

$$Sup(X) \geq \delta, \Gamma = \{X' | Sup(X') < \delta\} \quad (3)$$

Through these constraints, we can effectively reduce the number of generated patterns.

The maximality characteristic has the following effect. The left side of Fig. 1 shows an example of frequent patterns. The right side of Fig. 1 is a result of expressing the frequent patterns as representative ones using the maximality feature of them. As shown in the figure, a large number of frequent patterns can be expressed as a smaller number of maximal frequent patterns according to their support characteristics.

$WFPmax_{WA}$ and $WFPmax_{SD}$ [12] are tree-based algorithms that extract maximal frequent patterns considering weight factors of items. They use a recursive divide-and-conquer manner and employ different item sorting orders according to algorithm types. They can effectively be utilized in various areas dealing with static data. Theses algorithms scan a given database twice in order to extract weighted maximal frequent patterns. Their basic frameworks follow that of FP-Growth [3]. Therefore, in the process of the first database scan, they calculate a weight ascending order (in the case of $WFPmax_{WA}$) or support descending order (in the case of $WFPmax_{SD}$) from the given data and prune invalid items. In the second database scanning process, they construct their own tree structures and perform pattern growth works recursively. IM_WMFI [11] is a weighted maximal frequent pattern mining algorithm for processing incremental databases. In contrast to the above ones, the method can be employed effectively in environments where data are continually accumulated. It also follows a tree-based pattern growth manner. Additionally, IM_WMFI constructs its own tree structure and performs mining operations within a single database scan in order to handle such dynamic data efficiently. Since all of the necessary works for mining weighted maximal frequent patterns have to be conducted within a single database scan, the algorithm directly stores given data into its own tree structure at first, and then it performs additional tasks for tree restructuring. In order to increase efficiency of tree restructuring, IM_WMFI divides its tree (the entire task) into a number of paths (smaller tasks) and performs item resorting operations for each path (a divide-and-conquer manner). AWMax [4] is an approach integrating the concepts of approximate pattern mining and maximal frequent pattern mining. The method mines maximal frequent patterns considering error tolerance and weight conditions on noise environments. Accumulated data may have unexpected errors such as noise, device

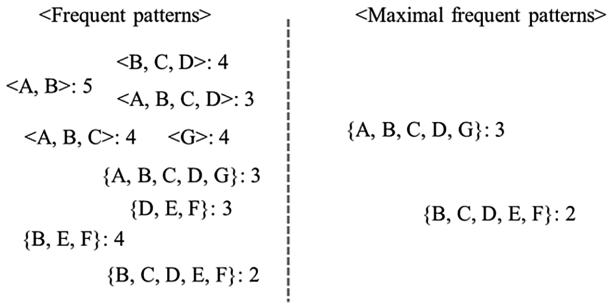

Fig. 1. Example of relations among frequent patterns and maximal frequent patterns

Table 1. Features of recent maximal frequent pattern mining algorithms

Algorithm	Stream processing	Considering weight factor	Tree-based approach	Approximate method
$WFPmax_{WA}$	N	Y	Y	N
$WFPmax_{SD}$	N	Y	Y	N
IM_WMFI	Y	Y	Y	N
AWMax	N	Y	Y	Y

malfunction or failure, record error, etc. In these cases, AWMax can effectively extract weighted maximal frequent patterns considering error tolerance.

Table 1 shows major features of the aforementioned algorithms. According to types of given data and purposes of pattern mining results, users can select and utilize proper algorithms.

4 Conclusion

In this paper, we explained the concept and characteristics of maximal frequent pattern mining and conducted analysis of recent relevant methods. Extensive usability of maximal frequent pattern mining led to various techniques and applications. Therefore, users can select and employ appropriate algorithms according to their own situations. The maximality property of patterns can also effectively be applied in various areas that may cause excessive computational overheads such as graph pattern mining, distributed processing of big data, etc. We are scheduled to conduct such expanded studies in our future work.

Acknowledgments. This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 20152062051 and NRF No. 20155054624) and the Business for Academic-industrial Cooperative establishments funded Korea Small and Medium Business Administration in 2015 (Grant no. C0261068).

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Goparaju, A., Brazier, T., Salem, S.: Mining representative maximal dense cohesive subnetworks. Netw. Model. Anal. Health Inf. Bioinform. **4**(1), 29 (2015)
3. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining Knowl. Discov. **8**(1), 53–87 (2004)
4. Lee, G., Yun, U., Ryang, H., Kim, D.: Approximate maximal frequent pattern mining with weight conditions and error tolerance. Int. J. Pattern Recogn. Artif. Intell. **30**(6), 1650012:1–1650012:42 (2016)

5. Li, H., Zhang, N.: Probabilistic maximal frequent itemset mining over uncertain databases. In: 21st International Conference on Database Systems for Advanced Applications, pp. 149–163 (2016)
6. Karim, M., Rashid, M., Jeong, B., Choi, H.: Privacy preserving mining maximal frequent patterns in transactional databases. In: 17th International Conference on Database Systems for Advanced Applications, pp. 303–319 (2012)
7. Necir, H., Drias, H.: A distributed maximal frequent itemset mining with multi agents system on bitmap join indexes selection. Int. J. Inf. Technol. Manage. **14**(2/3), 201–214 (2015)
8. Nourine, L., Petit, J.: Extended dualization: application to maximal pattern mining. Theor. Comput. Sci. **618**, 107–121 (2016)
9. Salem, S., Ozcaglar, C.: MFMS: maximal frequent module set mining from multiple human gene expression data sets. In: 12th International Workshop on Data Mining in Bioinformatics, pp. 51–57 (2013)
10. Stattner, E., Collard, M.: MAX-FLMin: an approach for mining maximal frequent links and generating semantical structures from social networks. In: 23rd International Conference on Database and Expert Systems Applications, pp. 468–483 (2012)
11. Yun, U., Lee, G.: Incremental mining of weighted maximal frequent itemsets from dynamic databases. Expert Syst. Appl. **54**, 304–327 (2016)
12. Yun, U., Lee, G., Lee, K.: Efficient representative pattern mining based on weight and maximality conditions. Expert Syst. (2016). (in press)