Research Frontiers in Sequential Pattern Mining

Ritika[1], Sunil Kumar Gupta[2]
{ritikasood1987@gmail.com[1], skgbcetgsp@gmail.com[2]}

I. K. Gujral Punjab Technical University, Kapurthala, Punjab, India [1],
Department of Computer Science & Engineering, Sardar Beant Singh State University, Gurdaspur, Punjab, India[2]

Abstract. The area of sequential pattern mining is progressing rapidly day by day. The aim of sequential pattern mining is to extract subsequences satisfying a threshold parameter value. Researchers are addressing the challenges and problems occurring in various domains of sequential pattern mining(SPM) by defining new constraints, algorithms to increase the quality of patterns. This paper presents an organized study on the research in the varied domains of sequential pattern mining such as Traditional SPM, Time-interval SPM, High Utility SPM and High Utility Hierarchical SPM. The fields are classified on the basis of support, time or utility as the threshold parameter. A taxonomy of the popular algorithms in these diversified fields of sequential pattern mining is also presented. The paper aims to act as a valuable guide for researchers by suggesting future potentials in sequential pattern mining.

Keywords: Utility, Sequential pattern mining, Utility mining

1 Introduction

Data mining techniques perform analysis of the data stored in the databases to generate knowledgeable and useful patterns. The databases can be transactional, healthcare databases, educational databases or financial databases, etc. The field of data mining emerged with text mining, then phrase mining, and progressed towards the mining of sequential patterns[1]. Sequential pattern mining(SPM) focuses on the extraction of sequential patterns from the databases. Earlier, SPM was performed based on support as threshold and later the concept of utility mining come forth. A sequential pattern can be like:{desktop, printer} a customer after buying a desktop computer mostly buys a printer. So it will be more useful to recommend buying a printer to a customer who is buying the desktop computer.

The era of sequential pattern mining began in the year 1995[2]. The authors developed an Apriori algorithm which lacked the concept of time constraints and taxonomy. These limitations were overcome by the popular Generalized Sequential Pattern mining algorithm *GSP*[3]. The GSP also suffered from the limitations of the Apriori algorithm. One of the limitations of the Apriori algorithm was the generation of numerous candidates which was overcome by the pattern growth approaches[4, 5]. These pattern growth based algorithms considered frequency or support as a

THEETAS 2022, April 16-17, Jabalpur, India

DOI 10.4108/eai.16-4-2022.2318239

threshold. These algorithms were further extended to extract patterns that include information about time intervals between items in a pattern.

Later, the research journey moved towards enhancing the significance of patterns by mining them on the basis of utility rather than support. A foundational mathematical model was proposed for the mining of itemsets based on utility[6]. Then, the utility mining was performed for mining of patterns[6, 7, 8]. Time interval sequential pattern mining was also performed by Wang and Huang taking utility as the threshold instead of support[9]. Some researchers also worked towards the hybridization of constraints by taking support and frequency both as thresholds. Literature is enriched with research articles favouring the utilization of effective pruning strategies in the mining process [10, 11].

This paper presents an overview of various research frontiers in sequential pattern mining by surveying the articles relating to its diversified domains. There is also a discussion of future research prospects in these areas. The remaining paper is organized into various sections. The mathematical constructs related to SPM are discussed in Section 2. In Section 3, some research drifts in SPM are discussed along with their taxonomy. A study of research challenges and future scope is given in Section 4. The last section presents the conclusion of the work.

2 Sequential Pattern Miining

Sequential pattern mining discovers subsequences that are interesting to users. The interest of a sub-sequence can expressed in terms of various parameters such as support, profit, discount, utility, etc.

2.1 Mathematical Constructs

Definition 1 *Quantitative Sequence Database*

A quantitative sequence database QS_{DB} consists of a set of sequences $QS_{DB} = < seq_1, seq_2, \cdots, seq_n >$. Each sequence is identified by a sequence-identifier $1, 2, \cdots, n$. Table 1 is an example of a sequence database having five sequences representing the time and quantity information related to the purchase of items from p to t.

Every row of Table 1 contains information about the sequence identifier and the corresponding sequence. The sequences of the table are written as $< (x_1, t_1, q_1), (x_2, t_2, q_2), \cdots, (x_n, t_n, q_n) >$, where x represents an item, the symbol t represents the time of purchase and the symbol q represents the purchased quantity corresponding to x.

Definition 2 *Time intervals*

$TI = \{TI_0, TI_1, TI_2, \cdots, TI_k\}$ represents the set of k time intervals. The time ti lies in the range as given in Table 2.

Definition 3 *Itemset*

An Itemset is defined as the subset of a set I_set containing items given as: $I\_set = \{itm_1, itm_2, \cdots, itm_m\}$.

Table 1: Quantitative Sequence Database

Sequence Identifier	Sequence
sq_1	$< (p,1,3)\{(q,2,4),(s,3,1)\} >$
sq_2	$< \{(q,2,4),(s,4,5),(t,5,1)\},(p,1,2) >$
sq_3	$< (q,2,3),(p,1,4),(q,3,3),\{(r,2,3),(s,1,1)\} >$
sq_4	$< (p,2,1),(r,2,4) >$
sq_5	$< \{(r,1,2),(s,2,3)\},(t,1,6) >$

Table 2: Time Intervals

Time Interval	Range
TI_0	$ti = 0$
TI_1	$0 < ti \leq TI_1$
$\cdots$	$\cdots$
TI_j	$TI_{j-1} < ti \leq TI_j$
TI_k	$TI_{k-1} < ti < \infty$

Definition 4 *Sequence*

A sequence consists of the itemsets arranged in alphabetical order. It is given as $SEQ =< I\_set_1, I\_set_2, \ldots, I\_set_n >$.

Definition 5 *Time Interval Sequence*

Given an itemset I_set as given in Definition 3 and a set of time-intervals TI as specified in Definition 2, the time interval sequence is of the form $TIS =< itm_1, ti_1, itm_2, ti_2, \ldots, itm_{s-1}, ti_{s-1}, itm_s >$ such that $itm_i \in I\_set$ for $1 \leq i \leq s$ and $ti_i \in TI$ for $1 \leq i \leq s-1$

Definition 6 *Support Value*

$$support\_value(pt) = \{(sq_{id}, sq) | (sq_{id}, sq) \in QS\_DB \wedge ptiscontainedinsq\}$$

Definition 7 *Internal Utility*

The sequences stored in a quantitative sequence database also contain information about the purchase quantities related to an item which are considered to be its internal utility.

Definition 8 *External Utility*

The information regarding external utility is stored in a separate utility table as given in Table 3:

Definition 9 *Utility*

It is calculated by multiplying the internal utility of an item with its corresponding external utility.

Table 3: Utility Table

Item	External Utility
p	2
q	3
r	1
s	4
t	9

3 Research Frontiers in Sequential Pattern Mining

3.1 Traditional Sequential Pattern Mining

The traditional SPM approaches mine patterns considering frequency as the measure of interestingness. The concept of sequential pattern mining was introduced in 1995[2]. Two algorithms, AprioriAll and AprioriSome were proposed for mining sequential patterns from transactional databases. This work was extended by proposing a GSP algorithm that generalized the problem of SPM through the inclusion of time constraints and taxonomies[3]. These algorithms were slow in execution as they suffered the limitations of the Apriori approach involving multiple scans and the generation of numerous candidates.

Later, researchers developed pattern growth-based approaches for sequential pattern mining. In the year 2000, a pattern growth-based method, FreeSpan was proposed which recursively mined projected databases[4]. Experimental evaluation showed that the FreeSpan method performed faster than the previous Apriori-based algorithms. A very popular pattern-growth based algorithm, PrefixSpan performs prefix-projection[5] and executed faster than the FreeSpan[4] and GSP algorithm[3]. The pseudo and level-by-level projection techniques were also proposed to further enhance the mining efficiency.

An organized study covering the introduction of the pattern growth methodology and its principles is given by Han in [12]. The authors discussed GSP[3], SPADE based on candidate generate and test strategy, and FreeSpan[4], PrefixSpan[5], CloSpan[13] algorithms based on pattern growth methodology. Various extensions of the SPM are discussed covering constraint-based mining[14], mining of top-k sequential patterns, multi-dimensional mining and multi-level mining. A study and framework for constraint-based mining through the incorporation of constraints is given in [14]. These constraints help to define new interestingness measures and reduce the candidate set.

A systematic review of the literature related to various domains such as multi-dimensional mining, mining without candidate generation, closed pattern mining, mining using vertical data format, multi-level mining, structural pattern mining, constraint-based mining, approximate pattern mining, maximal frequent pattern mining[15] was presented. P-PrefixSpan approach, an extension of PrefixSpan can also be applied to extract reliable sequential patterns as proposed in [16]. This algorithm incorporates the time-probability constraint and helps to make effective marketing strategies based on probability and monetary constraints. The study given in [17] presents an introduction, strategies

and research scope in SPM like mining of closed patterns, approximate patterns, maximal patterns and sequential generators.

3.2 Time-Interval based SPM

The traditional SPM algorithms such as Apriori[2], GSP[3], FreeSpan[4] and PrefixSpan[5] as discussed in previous subsection do not discover time intervals between items in succession. The database as given in Table 1 contains information about time of purchase of items which can be utilized in making marketing plans. Algorithms for generation of sequential patterns including time information using the Apriori and the PrefixSpan algorithm is given in [18]. A time-interval sequential pattern as defined in Section 2 would be like a customer who bought a computer, again buys a printer after 7 days. The time intervals are also specified in advance as defined in Section 2. Approaches for mining with multi-time intervals were also proposed in [19]. These algorithms suffered from the sharp boundary problem which can be solved using fuzzification. Fuzzy theory concepts can be embedded in time-interval SPM as implemented in FTI-Apriori algorithm[5] and FTI-PrefixSpan[20] algorithms.

3.3 High Utility Mining

High Utility mining refers to the process of extraction of subsequences satisfying the minimum utility threshold value. The foundation of utility mining was laid in the year 2004 through the development of a model for itemset mining using the concept of transaction and external utility[6]. Two-phase algorithm, given in the year 2005 also worked on the utility mining of itemsets in two phases[7]. Another approach for mining of itemset utilities showed improved performance due to the incorporation of novel pruning strategies[8]. The concept of sequence utility is also defined for representing the utility while proposing the UtilityLevel and UtilitySpan algorithms for high utility SPM[21].

GPA Algorithm, an extension of the two-phase algorithm performs level wise mining and applies a rigorous upper bound[22]. PB algorithm applies the projection mechanism to enhance the efficiency of mining. Both GPA and PB algorithms perform better than the Two-Phase algorithms due to the use of projection strategy. Tree based algorithms also show improved efficiency as compared to the two-phase algorithm through the use of effective pruning strategies.

There are various algorithms in literature which use the tree data structure for representing the utility information. UP-Growth[23] and UP-Growth+60 algorithms incorporate pruning strategies while performing high utility itemset mining. The information is served in a UP-tree data structure. Several other researchers have also proved the effectiveness of pruning strategies in the mining process. Krishnamoorthy also proposed a novel method using lookahead and partitioned utility pruning strategies for mining of itemsets[11]. USpan algorithm constructs a lexicographic sequence tree which is integrated with concatenation and pruning strategies to improve the performance[24]. The high utility sequential patterns are found by performing the depth first search traversal of the tree. HUS-Span shows improved performance as compared to the USpan algorithm by employing the RSU and PDU pruning strategies[25].

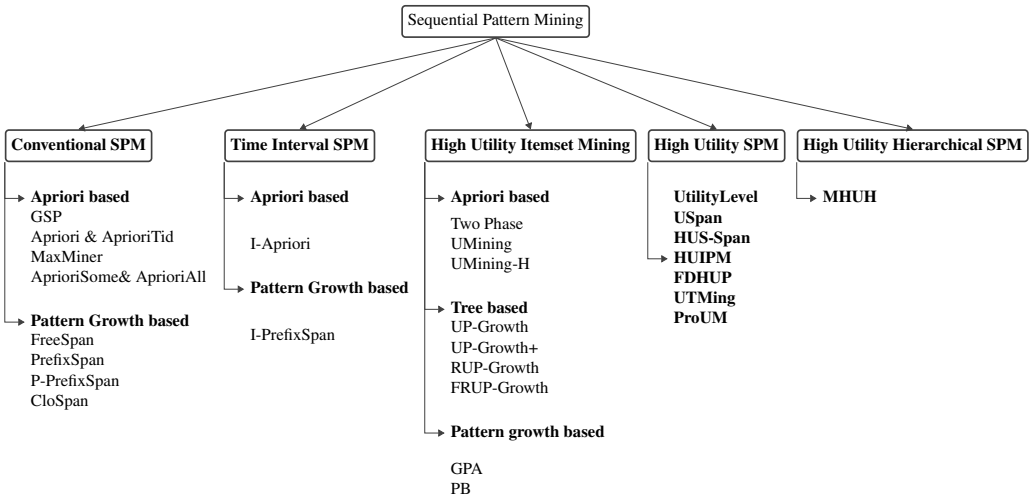

Fig. 1. A Taxonomy of SPM algorithms

A survey in 2013 suggested that the hybrid algorithms satisfying more than one interestingness measure can give more useful results[26]. UTMing algorithm[9] combines utility and time information for mining of sequential patterns giving improved accuracy. HUIPM algorithm bring out a novel metric frequency affinity to represent the value of frequency for each item pertaining to a transaction[27]. This frequency affinity is then integrated with utility. Based on the similar strategy, FDHUP algorithm gives faster results as compared to the HUIPM algorithm[28]. An extension of the UP-Growth algorithm, FRUP algorithm considers utility as well as support during mining giving more useful results. Pro-UM, a projection based utility mining approach uses the utility array structure for representing utility information[29]. The use of projection technique results in significant performance improvement.

The traditional High Utility SPM algorithms do not use the knowledge of hierarchical relationship among items. MHUH algorithm[30] utilizes the information given in taxonomies and incorporates it into high utility sequential pattern mining. A novel pruning strategy, taxonomy sequence weighted utility is also employed during exploration of candidates. Gan et. al. presents a well organized study on various methods of Utility Pattern Mining and presents a taxonomy[31]. This study provides a discussion of various research challenges and scope in the domain of utility pattern mining.

A taxonomy of most widely used algorithms in the domain of SPM is given in Figure 1. At the first level, the algorithms are classified with respect to the threshold parameter, and at the second level, they are classified on the basis of the underlying principle.

4 Challenges

SPM has many open challenges which can be taken as a research opportunity by researchers. This section presents a discussion of some of those challenges.

- Handling Uncertain, unstructured, heterogeneous data: There is need of efficient algorithms, especially in the field of Utility mining to handle uncertain or unstructured data including various types of data such as records, images, audio and video data etc. Data structures can be defined for representing such data efficiently.
- Defining new Constraints: New constraints can be defined and applied in the field of sequential pattern mining. This will reduce the number of candidates generated during exploration.
- Imposing multiple constraints: Researchers apply constraints in SPM as per the requirements of the application to specify the interest. Hybrid algorithms can be developed by specifying more than one constraint in the algorithm to extract more useful patterns.
- Effective Pruning Strategies: Several pruning strategies have been defined in literature, but there is still scope left for improvement. New pruning strategies or hybridization of existing pruning strategies can be performed to have a gain in efficiency.
- Real-time Results: There is a need for algorithms in sequential pattern mining which gives results in real time.

.

5 Conclusion

The aim of sequential pattern mining is to extract patterns or subsequences which satisfy a threshold value such as support, profit or utility. This paper begins with an introduction, classification and research opportunities in Sequential pattern mining. The related literature is discussed and classified into Traditional SPM which is mined on the basis of support, Time Interval SPM which includes time information and High Utility SPM which is mined on the basis of utility. A taxonomy of some popular algorithms is presented. This field is progressing day by day but still there is space for further exploration. After reviewing the literature, some research challenges have been found and discussed in this paper.

References

[1] Han J, Kamber M, Pei J. Third Edition : Data Mining Concepts and Techniques. Journal of Chemical Information and Modeling. 2012;53(9):1689-99.

[2] Agrawal R, Srikant R. Mining sequential patterns. Proceedings - International Conference on Data Engineering. 1995:3-14.

[3] Ramakrishnan Srikant RA. Mining Sequential Patterns: Generalizations and Performance Improvements. In: In 5th International Conference on Extending Database Technology,(EDBT'96), P. M. G. Apers, M. Bouzeghoub, and G. Gardarin, Eds. LNCS. Avignon, France; 1995. p. 3-17.

[4] Han J, Pei J, Mortazavi-Asl B, Chen Q, Dayal U, Hsu MC. FreeSpan: Frequent Pattern-projected Sequential Pattern Mining. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '00. New York, NY, USA: ACM; 2000. p. 355-9.

[5] Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, et al. Mining sequential patterns by pattern-growth: the PrefixSpan approach. IEEE Transactions on Knowledge and Data Engineering. 2004 nov;16(11):1424-40.

[6] Yao H, Hamilton HJ, Butz CJ. A Foundational Approach to Mining Itemset Utilities from Databases. In: SIAM International Conference on Data Mining; 2004. p. 482-6.

[7] Ying Liu WkL, Choudhary A. A Two Phase Algorithm for Fast Discovery of High Utility Itemsets. In: Advances in Knowledge Discovery and Data Mining; 2005. p. 689-95.

[8] Yao H, Hamilton HJ. Mining itemset utilities from transaction databases. Data & Knowledge Engineering. 2005;59:603-26.

[9] Wang W, Huang AY. Considering high utilities for time interval sequential pattern mining. In: Proc. Conf. Technologies and Applications of Artificial Intelligence (TAAI); 2015. p. 412-8.

[10] Ahmed CF, Tanbeer SK, Jeong BS, Lee YK. HUC-Prune: an efficient candidate pruning technique to mine high utility patterns. Applied Intelligence. 2011;34(2):181-98.

[11] Krishnamoorthy S. Pruning strategies for mining high utility itemsets. Expert Systems with Applications. 2014.

[12] Han J, Pei J, Yan X. Sequential Pattern Mining by Pattern-Growth: Principles and Extensions. Foundations and Advances in Data Mining. 2005 sep:183-220.

[13] Yan X, Han J, Afshar R. CloSpan: Mining: Closed Sequential Patterns in Large Datasets; 2003. .

[14] Pei J, Han J, Wang W. Mining sequential patterns with constraints in large databases. In: International Conference on Information and Knowledge Management, Proceedings; 2002. .

[15] Burdick D, Calimlim M, Gehrke J. MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases; 2001.

[16] Shyur HJ, Jou C, Chang K. A data mining approach to discovering reliable sequential patterns. Journal of Systems and Software. 2013.

[17] Fournier-Viger P, Lin JCW, Kiran RU, Koh YS, Thomas R. A Survey of Sequential Pattern Mining. Data Science and Pattern Recognition. 2017 feb;1(1).

[18] Yen-Liang Chen Mei-Ching Chiang MTK. Discovering time-interval sequential patterns in sequence databases. Expert Systems with Applications. 2003 oct;25(3):343-54.

[19] Hu YH, Wu F, Yang CI. Mining multi-level time-interval sequential patterns in sequence databases. In: 2nd International Conference on Software Engineering and Data Mining, SEDM 2010; 2010. p. 416-21.

[20] Chen YL, Huang TC. Discovering fuzzy time-interval sequential patterns in sequence databases. Part B (Cybernetics) IEEE Transactions on Systems, Man, and Cybernetics. 2005 oct;35(5):959-72.

[21] Ahmed CF, Tanbeer SK, Jeong B. A Novel Approach for Mining High-Utility Sequential Patterns in Sequence Databases. ETRI Journal. 2010 oct;32(5):676-86.

[22] Hong TP, Hsu JH, Yang KJ, Lan GC, Lin JCW, Wang SL. Mining high utility partial periodic pattern by GPA. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017. p. 820-4.

[23] Tseng VS, Wu CW, Shie BE, Yu PS. UP-Growth: An Efficient Algorithm for High Utility Itemset Mining. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '10. New York, NY, USA: ACM; 2010. p. 253-62.

[24] Han J, Pei J, Yin Y, Mao R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery. 2004;8(1):53-87.

[25] Wang JZ, Yang ZH, Huang JL. An Efficient Algorithm for High Utility Sequential Pattern Mining; 2014. p. 49-56.

[26] Parmar DK, Rathod YA, Patel MM. Survey on high utility oriented sequential pattern mining. In: Proc. IEEE Int. Conf. Computational Intelligence and Computing Research; 2013. p. 1-7.

[27] Ahmed CF, Tanbeer SK, Jeong BS, Choi HJ. A framework for mining interesting high utility patterns with a strong frequency affinity. Information Sciences. 2011.

[28] Lin JCW, Gan W, Fournier-Viger P, Hong TP. Mining Discriminative High Utility Patterns. Intelligent Information and Database Systems. 2016 jan.

[29] Gan W, Lin JCW, Zhang J, Chao HC, Fujita H, Yu PS. ProUM: High Utility Sequential Pattern Mining. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC); 2019. p. 767-73.

[30] Chunkai Z, Du Z, Zu Y. An Efficient Algorithm for Extracting High-Utility Hierarchical Sequential Patterns. Wireless Communications and Mobile Computing. 2020 07;2020:1-12.

[31] Gan W, Lin JCW, Fournier-Viger P, Chao HC, Tseng VS, Yu PS. A Survey of Utility-Oriented Pattern Mining. IEEE Transactions on Knowledge and Data Engineering. 2021;33(4):1306-27.