Sequential Pattern Mining: Review, Algorithmic, and Performance Analysis Survey

Dr. Poonam Yadav
D.A.V College of Engineering & Technology,
Kanina, Haryana 123027, India

Abstract*— *Generally, DATA mining is the concept extracting implicit, earlier unknown as well as potentially helpful information from databases. In that, Sequential pattern mining is the vital data mining issue, which discovers recurring subsequences in the sequence database. The major objective of sequential pattern mining is the pattern growth. In this work, 20 research works are reviewed to analyze the sequential pattern models. The analysis outcome is given under three categories like algorithmic analysis, performance measure and the attained best measure as well.

Keywords*—*Data mining; Pattern mining; Sequence database; performance measure; Association rule

Nomenclature

Acronym	Description
REs	Regular Expressions
UDDAG	Up Down Directed Acyclic Graph
LASSO	Logistic Regression along with the Selection operator
POI	Period of Interest
HUSP	High Utility Sequential Patterns
PSP	Positive Sequential Patterns
HUNSP	High Utility Negative Sequential Patterns
HUNSC	High Utility Negative Sequential Candidates
POSITING	Prediction mOdel based on SequentIal paTtern mINinG
PLMS-tree	Preorder Linked Multiple Supports tree
MSCP-growth	Multiple Supports – Conditional Pattern growth
SPM	Sequential Pattern Mining
SPIRIT-	Sequential Pattern mining with Regular Expression constraints
CTMSP-Mine	Cluster-based Temporal Mobile Sequential Pattern Mine

I. Introduction

In general, Data mining does the mining of implicit, earlier undefined, as well as potentially helpful data from databases. Number of approaches has been developed for extracting information. Among them, the most vital mining is extracting the sequential patterns. The approach of sequential pattern mining was initially developed for drawing huge count of works for pursuing certain applications. Even though, number of works has been performed in the field, the patterns (sequential) exposed in the conventional research works are either on 1 or 2-D patterns.

More importantly, the mining of SEQUENTIAL pattern is the significant data mining issue, which discovers recurrent subsequence in the database (sequence). The main model for this mining (sequential) [21] [22] [23] is the growth of pattern. Conventional pattern growth-related models (e.g., PrefixSpan) are on the basis of projected databases. At every recursion level, the identified patterns length is developed by 1, and the patterns are developed unidirectionally with the suffix direction. Giving a sequences set, where every sequence includes the elements list and every element includes items set. Moreover, the mining of sequential pattern is for identifying the total repeated subsequences, which means the subsequences that occur frequency is not minimum than min_support. Some of the sequence mining algorithms have been proposed under many studies, and among them, an author have generalized their determination of sequential patterns with the inclusion of sliding time window, time constraints, as well as user-defined taxonomy. An algorithm namely improved algorithm GSP (prior basis) was also developed. One another method has considered a inter-transaction association rule generalization. The mentioned generalization is basically rules whose left-hand as well as right-hand sides are episodes along time-interval restrictions. Further, certain studies have extended the scope from extracting sequential patterns for mining the periodic patterns (partial). Almost all the mentioned models to mine the sequential patterns are usually priori-like, i.e., on the basis of priori principle. This states that any of super-patterns cannot be recurrent, and that was on the basis of candidate generation-and test paradigm.

This paper gives a deep review on sequential pattern mining models. 20 papers are reviewed, and the papers are reviewed by means of the performance analysis, methods used and best values. The rest of the paper is arranged as follows: Section II reviews the literature work. Section III concludes the paper.

II. Analysis on Different Contributions

A. Related Works

In 2009, Garofalakis *et al.* [1] have stated that the discovering sequential patterns were the most vital issue in the process of data mining along the application domains host that comprises of telecommunications, medicine as well as World Wide Web. Further, the existing sequential pattern mining models grant users along single restricted approach. In this research work, the authors have developed the utility of REs that enables the user-controlled under incorporating the process of pattern mining. The authors have developed a family of new algorithms (namely SPIRIT) to mine the frequent sequential patterns, which also suit for user-specified RE conditions. Furthermore, the major differentiating fact between the developed approaches was enforced for pruning the patterns search space during evaluation. Finally, the solutions have granted valuable insights in the process of data mining.

In 2009, Perera *et al.* [2] have aimed for exploiting the data to do the supporting aspect of mirroring that exist helpful high-level data reviews about the group, along with the needed patterns. The major aim was to enabling the groups as well as their facilitators on seeing the related group's operation aspects and grants the feedback if all those highly related with both positive and negative outcomes. They have explored how the helpful mirror data could be mined through the theory-driven model and the clustering range and the mining of sequential pattern as well. They have extracted the patterns that differentiating the best from the poor groups and have got the imminent in the attained factors. The final outcomes have pointed the significance of leadership and the interaction of group. Also, they have identified the patterns that represented the best distinctive practices.

In 2008, Huang *et al.* [3] have developed a sequence index as the vital metric that was consequential because of its interpretability aspect by spatial statistics. The authors have proposed a new algorithm namely Slicing-STS-miner for tackling the algorithmic design limitations by the sequence index that would not defend the downward closure property. They have compared the developed algorithm over other STS-miner algorithm, which uses the property of weak monotone. Finally, the performance formulation via both synthetic as well as real-world data sets has shown the better performance of proposed work.

in 2005, Yu and Chen [4] have discovered the commonly attained sequential patterns from databases. Even though number of works has been already progressed, no model could identify the d-dimensional sequence data's sequential patterns. Without this ability, certain practical information might be unfeasible to mine. For instance, an stock-trading site might have a database (customer), in which every customer might visit a internet. In order to mine the sequential patterns, the authors have developed two effective algorithms.

In 2011, Lu *et al.* [5] have proposed a new algorithm termed CTMSP-Mine, for discovering the CTMSPs. Further, they have developed a prediction approach for predicting the consequent mobile behaviors. They have developed the user clusters via a new algorithm called CO-Smart-CAST and resemblance among users were formulated through the Location-Based Service Alignment (LBS-Alignment) and developed measure. At the same time, they have also developed a time segmentation model for finding the segmenting time intervals, in which the same mobile characteristics do presently. Finally, the investigational outcome has reviewed its better performance.

In 2010, Chen [6] had invented a new data structure, UDDAG, for effective mining of sequential pattern. Here, the proposed UDDAG permits the growth of bidirectional pattern with all the detected patterns ends. Hence, the identification of length-k pattern could be done in [log2k + 1] recursion levels, and that outcome in less recursion levels and rapid growth of pattern. If minSup is huge so that the pattern length (average) was nearer to 1, the UDDAG, as well as PrefixSpan, have attained same performance as the issue degrades into repeated item counting issue. Yet, UDDAG has attained better. In addition, the UDDAG special feature could enable its expansion on applications including large spaces search.

In 2009, Garg *et al.* [7] have presented a novel mathematical modelling application for healthcare, which granting vital data to the health service managers as well as policymakers for aiding them in identifying the sequential patterns that need attention for effective administration scarce healthcare resources as well as creating the efficient policies of healthcare management. The authors have presented a non-homogeneous Markov approach to find the patient pathways that has great probability and the findings of pathways that need greatest cost as well as time. To have a realistic approach, the authors have also considered the time-dependent covariates along with their effect on pathways. Additionally, the authors have also developed a new algorithm that was on the basis of branch as well as bound global optimization, and that could effectively mine the needed count of interest patient pathways. The proposed model was illustrated by historical information on geriatric patients from the London hospital's administrative database.

In 2016, Mert *et al.* [8] have commited on the location prediction of the supplementary activity of users of mobile phone. This research work has concentrated on three issues: prediction of location and the respective time of subsequent activity of user, location prediction of user's subsequent activity while changing the user location, and location as well as time prediction of user activity when the alteration of location of user happens. The authors have presented a model that was on the basis of sequential pattern extraction for all the three mentioned issues and has evaluated the success of the proposed approaches under real information. The results have proven the superiority of proposed model.

In 2017, Lim *et al.* [9] have considered a statistical process control approach to predict the last breakdown of the printed circuit board lot. This was on the basis of preserved event serials in the process of wire bonding. In order to assess the parameters more effectively, the authors have developed a two-stage processing. In this, the unrelated event subsequences were removed through the process of sequential mining, and next to this, they have developed a predictive approach with the residual events' subsequences through the bagged least absolute shrinkage as well as LASSO assessment that named as B-LASSO approach. Particularly, for solving the issues that occurred by unbalanced information, the developed B-LASSO have used the case-control sampling moderately. Finally, the performance was compared to other conventional approaches and has proven the superiority of proposed work.

In 2008, Huang *et al.* [10] has developed a common approach, in which the information in database might be inserted, static, or it can be removed. Additionally, they have presented a Pisa algorithm that stands for the progressive extraction of sequential patterns for discovering the sequential patterns in determined time POI. In fact, POI was a sliding window that constantly advances as days pass by. This has used a processing sequential tree for effective maintenance of newest information sequences, discovering the whole set of up-to-date sequential patterns, and also removing the outdated information as well as patterns. The sequential pattern tree height developed was surrounded via the POI length, so that efficiently restricting the needed memory space by Pisa, which was eventually least than the required memory by direct appending (DirApp). Finally, the investigational outcomes have proven the benefits of POI in terms of computational time and scalability.

In 2004, Pei *et al.* [11] have developed a sequential pattern-growth model that was on the basis of projection to have effective pattern extraction. Here, a sequence database was recursively anticipated through the exploration of local frequent fragments. Further, the authors have proposed an effective approach named PSP that has provided ordered enhancement and also minimized projected databases. In order to make better enhancement, the pseudoprojection approach was proposed in PrefixSpan. it was found that the PrefixSpan molded with pseudoprojection was more rapid between all the compared works. Apart from this, the extraction model could be extensive for mining the sequential patterns with the constraints like user-defined constraints.

In 2018, Xu *et al.* [12] have stated that the HUSP)extraction process was showing its vital role in various applications including smart campus as well as data analysis. Present HUSP mining algorithms, have only considered the PSP and no the negative one. Further, the negative sequential mining has been playing its significant role in many applications; however, the present algorithms have not given much contribution on mining the negative sequential pattern. Thus, the authors have developed a new algorithm called HUNSPM for mining HUNSP, which has the capability on solving the key issues of how to evaluate the usage of negative sequences, and how to produce high utility HUNSC more efficiently. At last, the performance of proposed model was proven.

In 2018, Maryam *et al.* [13] have stated that the prediction of applications' future workload was the most important step before doing the resource provisioning. The authors have developed a new prediction approach that was on the basis of sequential pattern mining, named POSITING. This has considered the correlation among various resources as well as extracts behavioral patterns. On the basis of mined patterns and the current application behavior, the resources' future demand was predicted. The major aim of the proposed model was the prediction of application workloads. Finally, the investigation was carried out using some real time workloads, and the results have proven the superiority of proposed model over other conventional methods with high accuracy rate.

In 2014, Shaw and Gopalan [14] have concerned with the identification of repeated trajectories of moving objects through the adoption of a new model that incorporates the clustering concept as well as the mining of sequential pattern. in this, the novel model has applied the threshold for obtaining the active clusters and has arranged those in descending order, which was on the basis of count of trajectories. This activity has reviewed the cluster patterns through the sequential pattern mining approach. Subsequently, the progression was continued till the linking of all the active clusters happen. Then, they have conducted the set of investigations through real-time datasets that have proven that the developed model was five times effective than the conventional methods. Further, examinations were done for identifying the effective threshold value.

In 2007, Kim *et al*. [15] have considered the issue in the extraction of sequential patterns along quantities. The authors have demonstrated that naive extensions over conventional models for patterns are ineffective since they might specify the search space more blindly. In order to handle the situation, the authors have developed the hash filtering as well as the quantity sampling models, which vitally enhance the naive extensions' performance. Investigation outcomes have shown that the when compared to naive extensions, the mentioned schemes reviews its better performances in terms of computational time as well as scalability.

In 2011, Kaneiwa and Yasuo [16] had stated that the Sequential pattern mining was more crucial and difficult task in various applications, for instance, analyzing the data behaviors under transactions as well as determining frequent patterns. Moreover, the task has become more complex if patterns were locally or implicitly concerned in the noisy data. This research work has developed a novel model for extracting the local patterns from sequences. With the use of rough set theory, they have described an algorithm to generate the decision rules. In order to concern the sequential data to rough set theory, they have specified the local patterns size, permitting a set of sequences that to be altered into a sequential data system. They have used the decision classes discernibility for establishing the assessment of decision rules in the sequential data system.

In 2018, Guangfei *et al.* [17] have analyzed the spatial-temporal features of PM2.5 pollution as well as have discovered the sequential patterns in cities in 3 significant zones. The sequential patterns have revealed the hidden associative relationships, which could grant proof for supporting united policy-making in various areas. The investigational outcome has revealed the vital heterogeneities between the underlying relationships, whereas some homogeneities present in certain seasons in the 3 regions. Further, the authors have discussed the relations with the monsoon systems.

In 2011, A.Sallaberry *et al*. [18] have focussed on mining of sequential pattern, and also have developed a novel visualization system for helping the analysis of end users' mined knowledge. This was also for highlighting the novelty as per the referenced biological documents' databases. The proposed system was on the basis of three visualization modalities: clouds, solar systems, as well as treemaps. The authors have viewed that the mentioned models were helpful to identify the associations along with the hierarchical relationships among patterns between relevant documents. Subsequently, the mined patterns from gene information via proposed system were assessed through 2 biology laboratories working under cancer disease and Alzheimer's disease.

In 2013, Hu *et al*. [19] have proposed a compact data structure, named PLMS-tree, to store and compressing the whole sequence database. On the basis of PLMS-tree, the authors have developed an effective algorithm, MSCP-growth, particularly for discovering the complete patterns set. Finally, the investigational outcomes have shown that the developed model could attain more preferable identifications than the MS-GSP and the existing SPM.

In 2008, Themis *et al*. [20] have used a SPM for recognizing the sequence-based fold. Further, the classification of protein in terms of fold recognition has played a significant role in computational analysis of protein, as it could contribute on the determination of protein function (unknown structures). Particularly, the authors have employed an effective SPM algorithm, cSPADE, to do the analysis process of the protein sequence. They have also used a classifier for classifying the proteins in suitable fold category. For the purpose of training as well as assessing, the authors have utilized the protein sequences from SCOP database. The investigational outcome has proven the betterments of proposed model over other methods with great accuracy rate.

B. Algorithmic Analysis

It is observed that various approaches and methods are developed for mining the sequential patterns. Fig 1 shows the algorithmic analysis of the current survey by reviewing 20 papers. In this, Family of novel algorithm named SPIRIT is proposed in [1]. The authors in [2] have used theory-driven approach. Slicing-STS-Miner is the model used in [3]. the algorithms AprioriMD algorithm, PrefixMD span algorithm are used in [4]. CTMSP-Mine is proposed in [5]. the new model named UDDAG is proposed in [6]. In [7], the authors have used non-homogeneous Markov model to mine the sequential pattern. Some process under sequential pattern mining-based techniques is used in [8]. The authors in [9] have proposed the bagged LASSO. Direct Appending is used in [10]. the authors in [11] have used projection-based, sequential pattern-growth approach. HUNSPm is developed in [12]. the new algorithm named POSITING is proposed in [13]. The authors in [14] used the k-means algorithm. SQUIRE: Sequential pattern mining with quantities is the new model that proposed in [15]. the authors in [16] have mined the sequential patterns using rough set theory. the authors in [17] have analysed various sequential pattern mining procedures. The visualization system is used in [18]. In [19], PLMS-tree is proposed. cSPADE is a method that is proposed in [20].

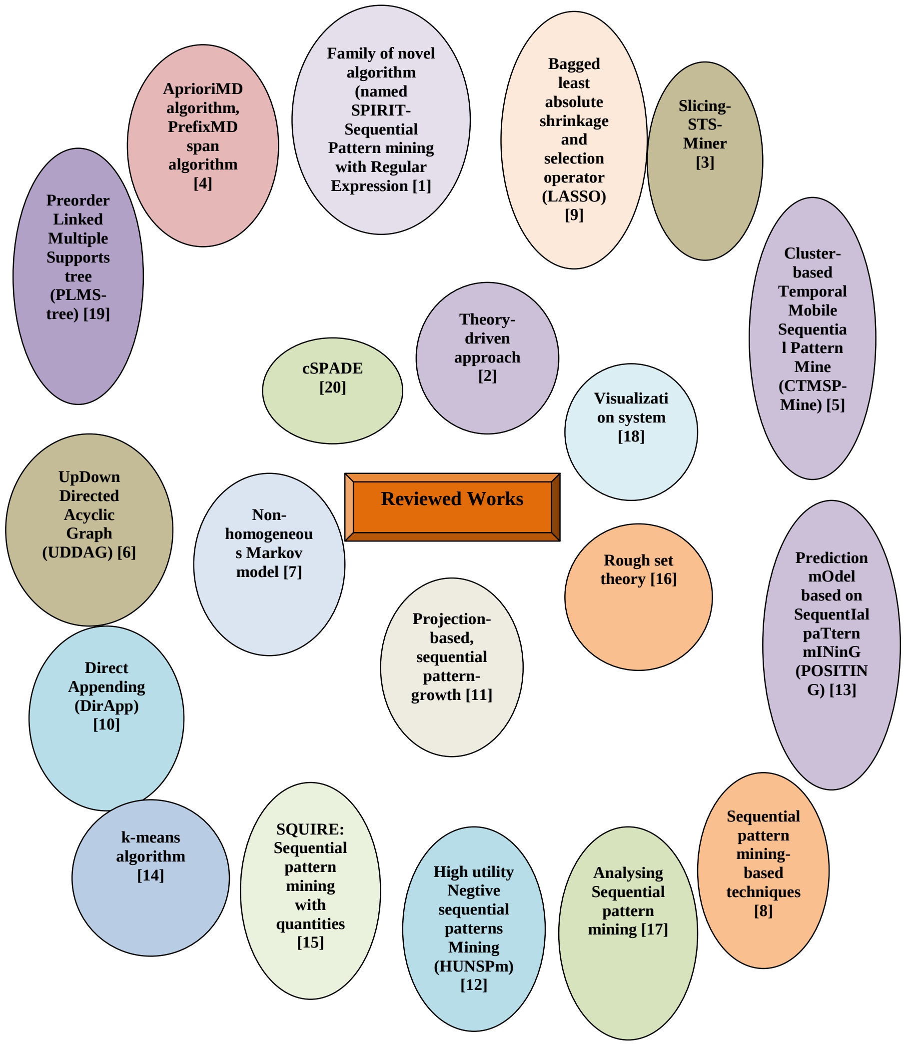

Fig. 1. Algorithmic analysis of reviewed works

C. Analysis on Performance Measures

In this revision, it is observed that the methods under sequential pattern models have used some measures like Running time, pattern size, precision, accuracy and so on. From the analysis (Table I), it is observed that 45% of contributions have measure the runtime. Only 5% of contributions have measured the pattern size. Precision measure is used by 20% of contributions. Similarly, only 5% of contributions have used the measures like recall and F-measure. 20% of contributions have used the memory usage as the measure to analyze the performance of proposed works. Similarly, 5% of contributions have worked on probability and expected cost. 10% of total contributions have measured the accuracy rate of proposed work. 5% of total contributions uses sensitivity measure. 35% of total contributions have used some other measures for analyzing the performance.

TABLE I. ANALYSIS ON PERFORMANCE MEASURES REVIEWED FROM VARIOUS WORKS

Citation	Running Time	Pattern Size	Precision	Recall	F-measure	Memory usage	Probability	Expected Cost	Accuracy	Sensitivity	Others
[1]	□										
[2]											
[3]	□	□									□
[4]	□										
[5]			□	□	□						
[6]						□					□
[7]							□	□			□
[8]									□		
[9]											□
[10]	□					□					□
[11]	□					□					□
[12]											□
[13]			□								
[14]	□										

[15]	□										
[16]	□										
[17]											
[18]											
[19]	□					□					
[20]									□	□	

D. Attained Best Performance measure

This section gives the attained best measure from all the reviewed works. Table II shows the analysis of the same. From this, it is observed that the best execution time is 900sec. the attained best pattern size is 5. The attained best precision rate is 0.58%. The best recall value is 0.7%. The best memory utilization is 10MB. Similarly, the best probability value is 0.001671. The best accuracy rate is 91.12%.

TABLE II. ATTAINED BEST PERFORMANCE

Measure	Best value	Citation
Execution time	900sec	[1] [3] [4] [10] [11] [14] [15] [16] [19]
pattern size	5	[3]
Precision	0.58%	[5] [13]
recall	0.7%	[5]
Memory	10MB	[6] [10] [11] [19]
probability	0.001671	[7]
Accuracy	91.12%	[8]
Number of HUNSP	1500	[12]
Sensitivity	0.98%	[20]

E. Analysis on Running time

This section gives a deep analysis on running time measures. From the Fig 2, it is reviewed that the 15% of contributions have attained the running time in the range of 1-200 sec. Then, 10% of contributions have attained the running time that in the range of 300-900 sec. Similarly, 15% of contributions have achieved the running time in the range of 1000-10,000 sec.

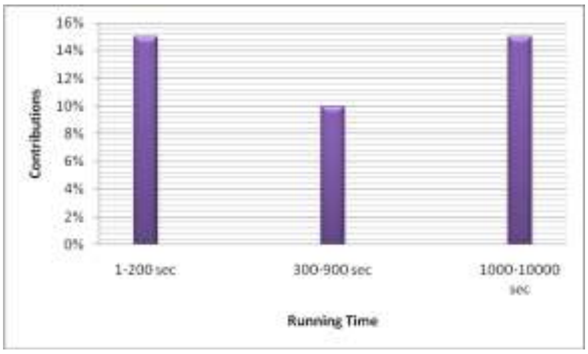

Fig. 2. Running time analysis

III. CONCLUSION

Sequential pattern mining was considered as the vital data mining issue that discovers recurring subsequence in the sequence database. This paper has reviewed 20 research works on sequential pattern models. The analysis outcome has reviewed the outcomes in terms of pictorial representation and tabulation. All the used performance measures are verified and tabulated. Then, the best values attained were also analyzed and progressed.

References

[1] M. Garofalakis, R. Rastogi and K. Shim, "Mining sequential patterns with regular expression constraints," in IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 3, pp. 530-552, May/Jun 2002.

[2] D. Perera, J. Kay, I. Koprinska, K. Yacef and O. R. Zaïane, "Clustering and Sequential Pattern Mining of Online Collaborative Learning Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 6, pp. 759-772, June 2009.

[3] Y. Huang, L. Zhang and P. Zhang, "A Framework for Mining Sequential Patterns from Spatio-Temporal Event Data Sets," in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 4, pp. 433-448, April 2008.

[4] Chung-Ching Yu and Yen-Liang Chen, "Mining sequential patterns from multidimensional sequence data," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 1, pp. 136-140, Jan. 2005.

[5] E. H. C. Lu, V. S. Tseng and P. S. Yu, "Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments," in IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, pp. 914-927, June 2011.

[6] J. Chen, "An UpDown Directed Acyclic Graph Approach for Sequential Pattern Mining," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 7, pp. 913-928, July 2010.

[7] L. Garg, S. McClean, B. Meenan and P. Millard, "Non-homogeneous Markov models for sequential pattern mining of healthcare data," in IMA Journal of Management Mathematics, vol. 20, no. 4, pp. 327-344, Oct. 2009.

[8] Mert Ozer, Ilkcan Keles, Hakki Toroslu , Pinar Karagoza and Hasan Davulcu, " Predicting the Location and Time of Mobile Phone Users by Using Sequential Pattern Mining Techniques", The computer journal, vol. 59, no.6, pp. 908–922, 2016.

[9] H. K. Lim, Y. Kim and M. K. Kim, "Failure Prediction Using Sequential Pattern Mining in the Wire Bonding Process," in IEEE Transactions on Semiconductor Manufacturing, vol. 30, no. 3, pp. 285-292, Aug. 2017

[10] J. W. Huang, C. Y. Tseng, J. C. Ou and M. S. Chen, "A General Model for Sequential Pattern Mining with a Progressive Database," in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 1153-1167, Sept. 2008.

[11] Jian Pei, Member, IEEE Computer Society, Jiawei Han, Senior Member, IEEE, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, Member, IEEE Computer Society, and Mei-Chun Hsu, " Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol. 16, no.10, 2004.

[12] T. Xu, T. Li and X. Dong, "Efficient High Utility Negative Sequential Patterns Mining in Smart Campus," in IEEE Access, vol. 6, pp. 23839-23847, 2018.

[13] Maryam Amiri, LeyliMohammad-Khanli and RaffaelaMirandola, " A sequential pattern mining model for application workload prediction in cloud environment", Journal of Network and Computer Applications, vol.105, pp. 21-62, March 2018.

[14] Arthur A.Shaw and N.P.Gopalan, " Finding frequent trajectories by clustering and sequential pattern mining", Journal of Traffic and Transportation Engineering , vol. 1, no. 6, pp.393-403, 2014.

[15] ChulyunKim, Jong-HwaLim, Raymond T.Ng and KyuseokShim, " SQUIRE: Sequential pattern mining with quantities", Journal of Systems and Software, vol. 80, no.10, pp. 1726-1745, October 2007.

[16] Ken Kaneiwa and Yasuo Kudo, " A sequential pattern mining algorithm using rough set theory", International Journal of Approximate Reasoning, vol. 52, no. 6, pp. 881-893, September 2011.

[17] Guangfei Yang, JianHuang and XiannengLi, " Mining sequential patterns of PM2.5 pollution in three zones in China", Journal of Cleaner Production, vol. 170, pp. 388-398, 2018.

[18] A.Sallaberry, N.Pecheur, S.Bringay, M.Roche and M.Teisseire, " Sequential patterns mining and gene sequence visualization to discover novelty from microarray data", Journal of Biomedical Informatics, vol. 44, no. 5, pp.760-774, October 2011.

[19] Ya-Han Hu, FanWu and Yi-JiunLiao, " An efficient tree-based algorithm for mining sequential patterns with multiple minimum supports", Journal of Systems and Software, vol. 86, no. 5, pp. 1224-1238, 2013.

[20] Themis P.Exarchos, CostasPapaloukas, ChristosLampros and Dimitrios I.Fotiadis, " Mining sequential patterns for protein fold recognition", Journal of Biomedical Informatics, vol. 41, no.1, pp.165-179, February 2008.

[21] WiardJorritsma, FokieCnossen, Rudi A.Dierckx, MatthijsOudkerk and Peter M.A.van Ooijen, " Pattern mining of user interaction logs for a post-deployment usability evaluation of a radiology PACS client", International Journal of Medical Informatics, vol. 85, no. 1, pp.36-42, January 2016.

[22] Mohammad Khanbabaeia, Farzad Movahedi Sobhani, MahmoodAlborzia and RezaRadfar, " Developing an integrated framework for using data mining techniques and ontology concepts for process improvement', Journal of Systems and Software, vol.137, Pages 78-95, March 2018.

[23] Chun-ShengWang, Ying-HoLiu and Kuo-ChungChu, " Closed inter-sequence pattern mining", Journal of Systems and Software, vol. 86, no.6, pp. 1603-1612, June 2013.