International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 28th nov 2021.
Link : https://ijiemr.org/downloads/Volume-10/Issue-11

DOI: 10.48047/IJIEMR/V10/I11/46

Title **EXPLORATORY VISUAL SEQUENCE MINING BASED ON PATTERN-GROWTH**

volume 10, Issue 11, Pages: 280-287

Paper Authors: **MISS. JURREYYAH FIRDAWS MOHAMMAD**

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

EXPLORATORY VISUAL SEQUENCE MINING BASED ON PATTERN-GROWTH

MISS. JURREYYAH FIRDAWS MOHAMMAD
UG Scholar, Dept of Computer science Engineering, BITS Pilani,Dubai.
Email:juriya.shayaan@gmail.com

ABSTRACT :

Sequential pattern mining finds applications in numerous diverging fields. Due to the problem's combinatorial nature, two main challenges arise. First, existing algorithms output large numbers of patterns many of which are uninteresting from a user's perspective. Second, as datasets grow, mining large numbers of patterns gets computationally expensive. There is, thus, a need for mining approaches that make it possible to focus the pattern search towards directions of interest. This work tackles this problem by combining interactive visualization with sequential pattern mining in order to create a "transparent box" execution model. We propose a novel approach to interactive visual sequence mining that allows the user to guide the execution of a pattern-growth algorithm at suitable points through a powerful visual interface. Our approach (1) introduces the possibility of using local constraints during the mining process, (2) allows stepwise visualization of patterns being mined, and (3) enables the user to steer the mining algorithm towards directions of interest. The use of local constraints significantly improves users' capability to progressively refine the search space without the need to restart computations. We exemplify our approach using two event sequence datasets; one composed of web page visits and another composed of individuals' activity sequences.

Keywords : Sequential pattern mining, interactive mining, visual data mining, mining with constraints.

I INTRODUCTION

Sequential pattern mining addresses the problem of detecting sequences of events as patterns in data . Identification and analysis of sequential patterns are of increasing importance in a range of top priority application domains such as electronic health record analysis, process control, cybersecurity and safety, autonomous systems and software, and aid in the understanding and debugging of machine learning systems. There are, however, two main challenges that need to be addressed before sequential pattern mining can be fully utilized. The first challenge is based on the vast number of possible patterns. State-ofthe-art algorithms may extract too many patterns, many of which may be of lesser significance or even irrelevant for the current analysis. This aspect makes it difficult for the user to grasp, and consequently use, the multitude of obtained patterns. Although tailored visualization techniques have been proposed helping the user to explore the large number of patterns produced by the mining algorithm, the effectiveness of the existing techniques needs to be significantly improved, both at the algorithm and visualization level. The second challenge is the computational complexity involved in pattern identification, as mining large number of patterns is computationally very expensive. One approach to tackling these problems, is to introduce constraints and promising results have been shown in many applications.

These two challenges are the motivation behind several interactive systems which allow the user to define constraints to increase the effectiveness and efficiency of the mining process. However, the actual mining algorithms in these systems then operate as a black box, and the user only gets to interact with the resulting patterns and not with the pattern generation. This paper builds on the idea of opening this black box and involving the expert in the mining process by embedding interactivity deeper in it, catering in this way for the possibility of the user to guide the execution of the algorithms at suitable points. To our knowledge, the possibility of changing and refining constraints while a particular sequence pattern is being built has not yet been considered, and it is an approach that addresses both challenges described above. To this end, we aim to investigate the possibility of breaking down existing algorithms into incremental steps making it possible to check point the mining process, display the current status and allow a user to intervene by imposing constraints that steer the algorithm in the direction of interesting patterns.

II. LITERATURE SURVEY

Sequential pattern mining, was developed in 1995 by R Agrawal and R Srikant, has been used in data mining research field with various applications. There are several sequential pattern mining algorithm which have been described in literature. Generally Sequential Pattern Mining Algorithms differ in two ways.

1) The process in which candidate sequences are generated and stored. The main objectives of algorithm are to minimize the set of candidate sequences.

2) The process in which support and frequency of candidate sequence is counted. Based on these two key criteria's sequential pattern mining can be divided into two parts: • Apriori Based. • Pattern Growth Based.

The Apriori and AprioriAll algorithms proposed by Agrawal and Srikant [2][4]. AprioriAll was the first generation of pattern mining algorithm, that is based on the apriori property and generate candidate sequences by using Apriori-generate join procedure. Apriori property says that all non-empty subset of frequent item set should also be frequent and this property is also called antimonotonic property. The limitation of this algorithms is multiple scan of database and creation of huge number of candidate sequence.

The SPADE (Sequential PAttern Discovery using Equivalence classes) [5] is based on vertical format pattern mining algorithm which was introduced by M Zaki, 2001. The algorithm uses vertical id-list database format and breakdown original search space into sub-lattices by using lattice-theoretic approach. The SPADE algorithm scan database three times. In 1st scan it construct frequent 1-sequences, In 2nd scan it construct frequent 2-sequences and in 3rd scan it construct all other frequent sequences. SPADE reduces input/output costs by minimizing database scans. It also reduces calculational cost by using efficient search schemes.

SPIRIT (Sequential Pattern mIning with Regular expressIon consTraints) [6] is sequential pattern mining algorithm with regular expression constraints. It uses some relaxed constraint which is very good with pruning technique. There are several versions of this algorithm, in which SPIRIT (V) (V for valid) perform best among all algorithm of SPIRIT family.

III SYSTEM ANALYSIS

EXISTING SYSTEM

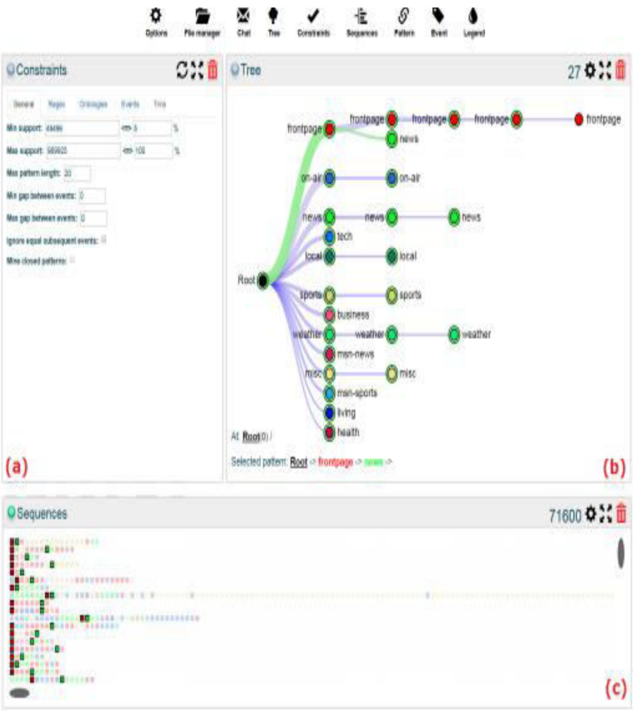
Sequential pattern mining finds applications in numerous diverging fields. Due to the problem’s combinatorial nature, two main challenges arise. First, existing algorithms output large numbers of patterns many of which are uninteresting from a user’s perspective. Second, as datasets grow, mining large numbers of patterns gets computationally expensive. There is, thus, a need for mining approaches that make it possible to focus the pattern search towards directions of interest. This work tackles this problem by combining interactive visualization with sequential pattern mining in order to create a “transparent box” execution model.

PROPOSED SYSTEM

We propose a novel approach to interactive visual sequence mining that allows the user to guide the execution of a pattern-growth algorithm at suitable points through a powerful visual interface. Our approach (1) introduces the possibility of using local constraints during the mining process, (2) allows stepwise visualization of patterns being mined, and (3) enables the user to steer the mining algorithm towards directions of interest. The use of local constraints significantly improves users’ capability to progressively refine the search space without the need to restart computations. We exemplify our approach using two event sequence datasets; one composed of web page visits and another composed of individuals’ activity sequences.

IV IMPLEMENTATION

Architecture:



Min support: [illegible] <=> 5 %
Max support: [illegible] <=> 100 %
Max pattern length: 20
Min gap between events: 0
Max gap between events: 0
Ignore equal subsequent events: □
Mine closed patterns: □

At: Root(0) /
Selected pattern: Root -> frontpage -> news ->

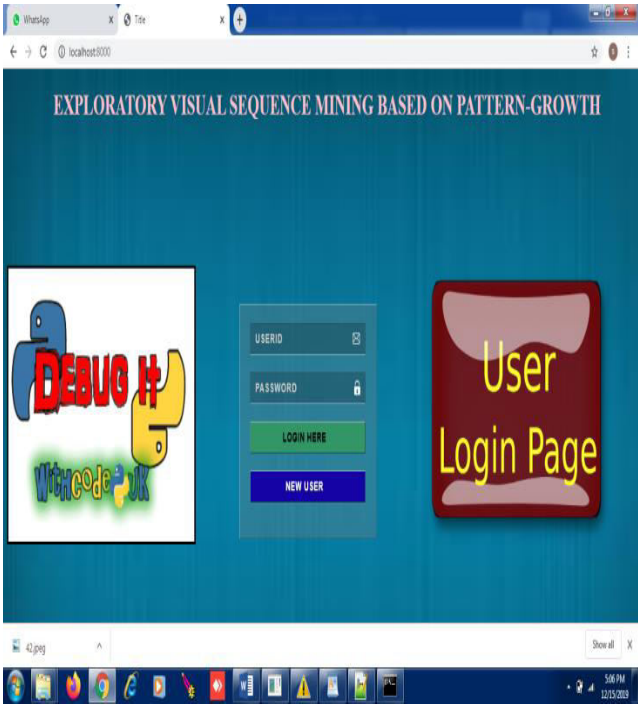
Fig-1: Architectures of the System Model

The motivation behind this work has been our vision of an approach to sequential pattern mining that deeply embeds interaction within the mining algorithm to facilitate exploratory mining. To this end, we have in this paper presented a novel interactive sequence mining approach based on the pattern-growth methodology, supporting local constraints, and implemented within ELOQUENCE. A key strength of our proposed approach is the introduction of local constraints in the mining process. An advantage of the use of local constraints is that it allows to incorporate expert knowledge on the fly, while searching for patterns. For instance, the user may know quite well that events in a set E usually occur after a sequence of events α. Thus, if α is revealed to be a pattern then the user can set event filter constrains to eliminate the events in E, when searching for α-supersequence patterns. In this way, the system does an amount of computation more proportional to the amount of new knowledge discovered and avoids displaying irrelevant patterns.

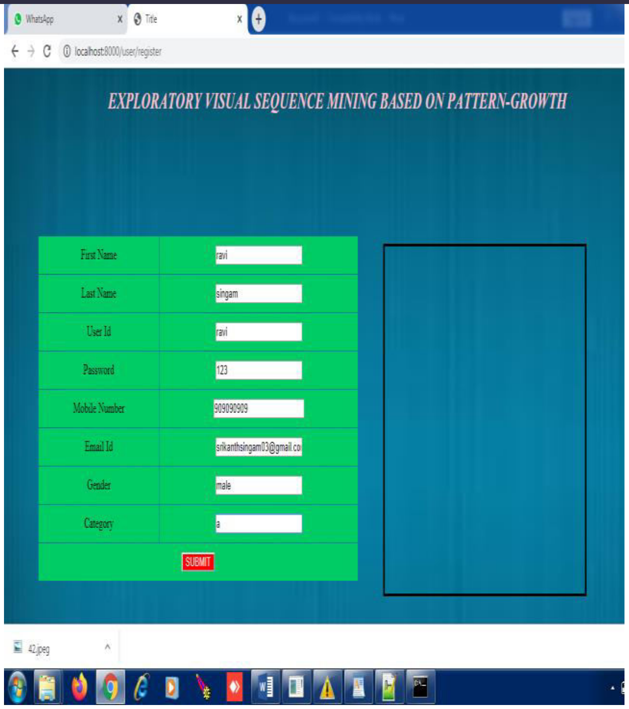
Another example, that illustrates the relevance of local constraints, is the fact that support constraints can be used to modify on the fly the support of patterns to be discovered. In our experiments with ELOQUENCE, we noticed that a common strategy is to set the initial support to a high value and expand the pattern tree a few levels. The choice of a high support leads to more common patterns and a smaller pattern tree. Upon inspection of the uncovered patterns in the pattern tree view, an interest in some of the patterns usually arises and there is the wish to further expand the tree with the corresponding supersequence patterns. Since the support tends to decrease as patterns get longer, it is common that some of the selected patterns cannot be further expanded due to the high support value set initially, preventing the user to continue the exploration in the direction the user is interested in. In some systems the user needs to re-start the mining algorithm with a lower value for the support. ELOQUENCE overcomes this limitation by simply allowing the user to modify the support constraint associated with a node (to lower the support) and continue finding supersequence patterns, without the need to redo computations.

Our proposed system embeds a “lazy execution” approach, in the sense that it is possible to select in the pattern view a pattern of length l > 0 and then request the system to build the α-supersequences of length l + n, where n > 0. In other words, the user can select a leaf-node and expand the sub-tree with root at that node by e.g. two levels at a time (n = 2). The advantages of this approach are twofold. First the user can request the system to display a few patterns at a time (as required by design goals G4 and G6). Second, it is a way to control the cluttering of the display resulting from expanding the tree many levels at once. Based on observation, we conclude that it is preferable to select a node in the tree, expand it by a few levels, and then inspect the new patterns uncovered. In this way, the user can get a preliminary insight about the patterns, often using the event sequence view, and then decide which patterns to expand next. Note that users may even prune a sub-tree from the pattern-tree, if the patterns in that tree are deemed as not interesting enough.

V RESULT AND DISCUSSION

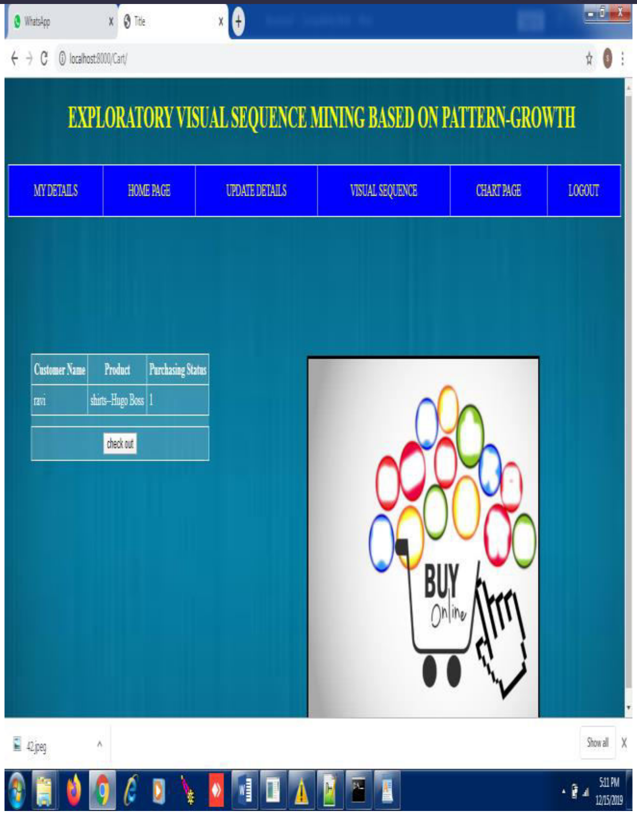

EXPLORATORY VISUAL SEQUENCE MINING BASED ON PATTERN-GROWTH

User Login Page



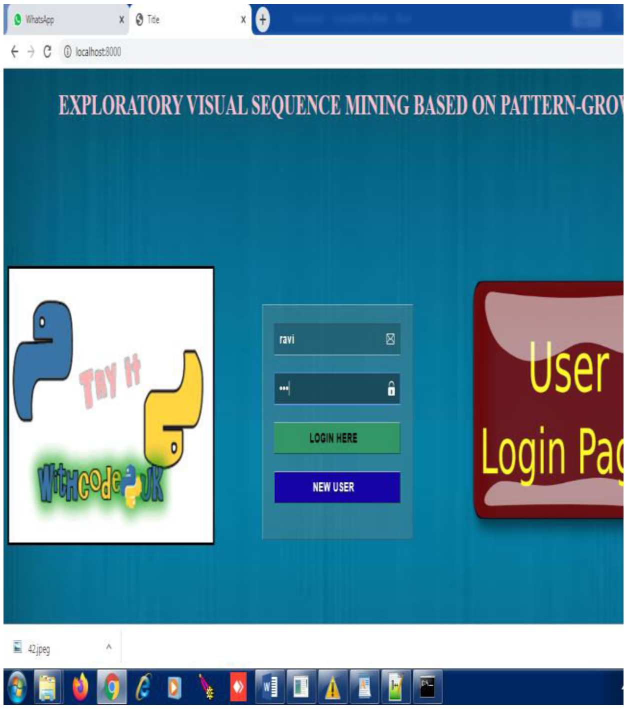
EXPLORATORY VISUAL SEQUENCE MINING BASED ON PATTERN-GROWTH

First Name	ravi
Last Name	singam
User Id	ravi
Password	123
Mobile Number	909090909
Email Id	srikanthsingam03@gmail.co
Gender	male
Category	a



EXPLORATORY VISUAL SEQUENCE MINING BASED ON PATTERN-GROWTH

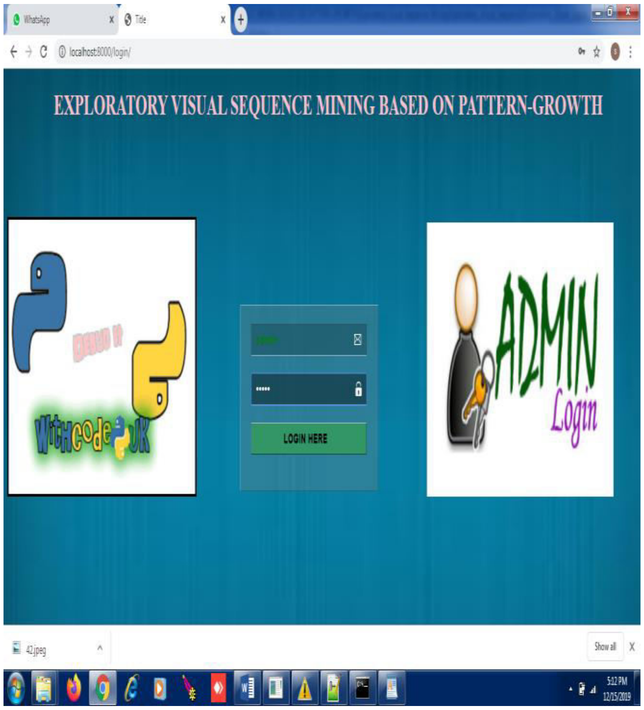
Customer Name	Product	Purchasing Status
ravi	shirts--Hugo Boss	1



EXPLORATORY VISUAL SEQUENCE MINING BASED ON PATTERN-GROV

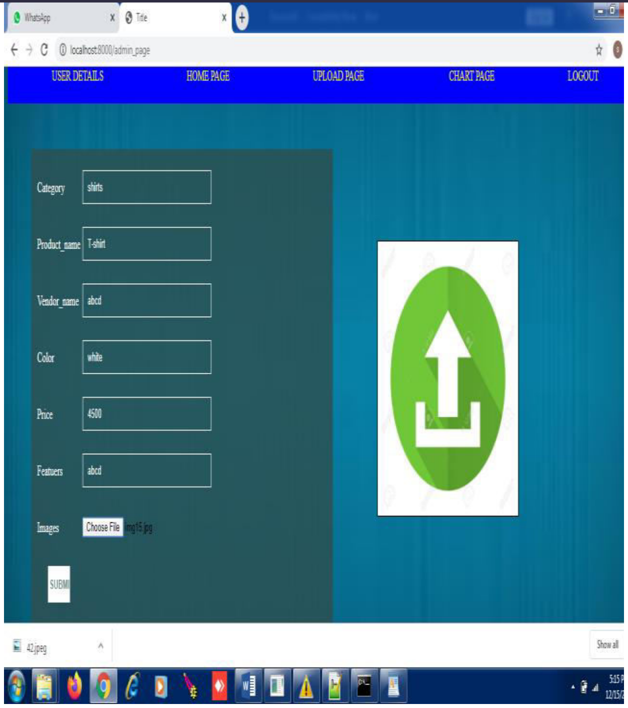
ravi

•••

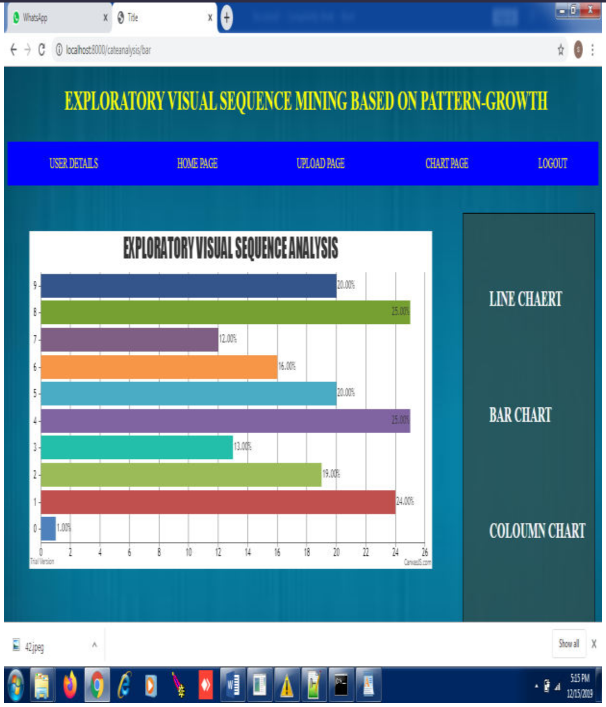


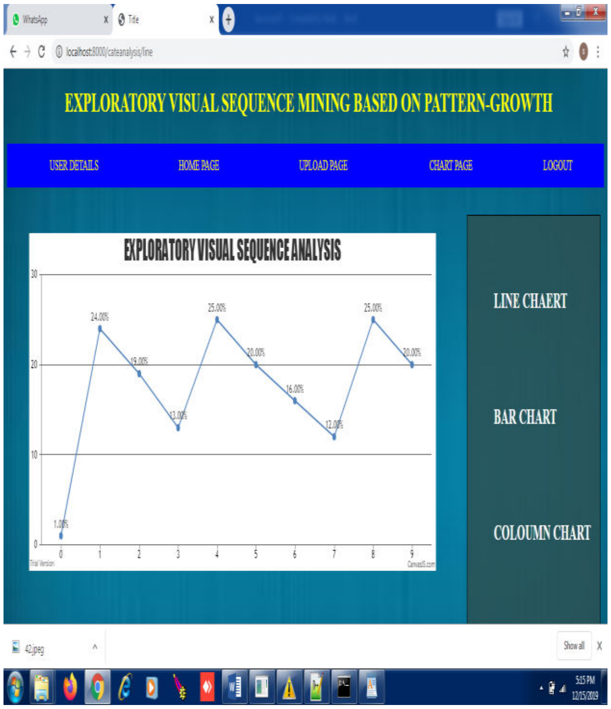
EXPLORATORY VISUAL SEQUENCE MINING BASED ON PATTERN-GROWTH

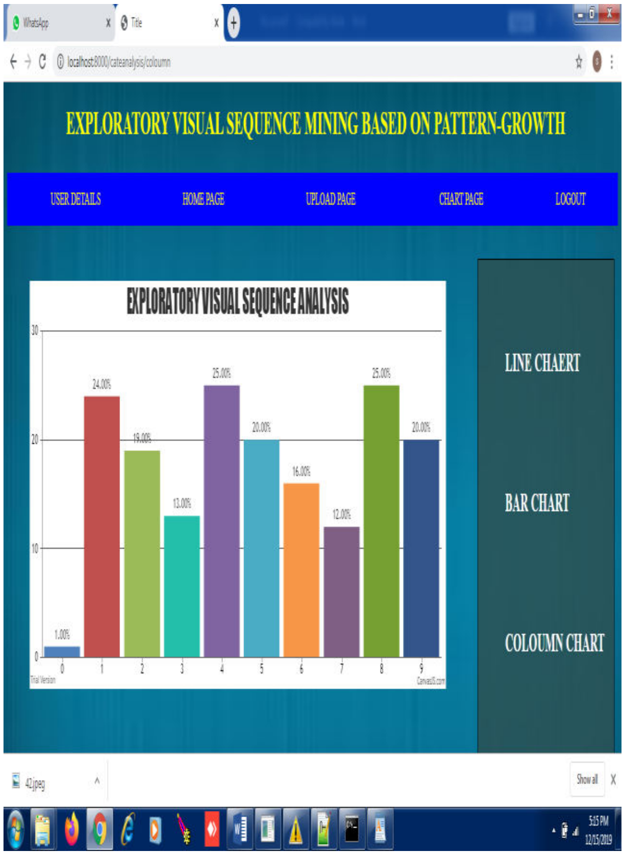
•••••



Category: shirts
Product_name: T-shirt
Vendor_name: abcd
Color: white
Price: 4500
Featuers: abcd
Images: img15.jpg







VI CONCLUSION AND FUTURE WORK

The main contribution of the proposed work is an interac- tive sequence mining approach that allows a user to progres- sively refine constraints while pattern sequences are being built, enhancing in this way user exploration and control over the search for interesting patterns. This contrasts with existing interactive sequential pattern mining systems that mostly offer the possibility of setting constraints at the start of the mining process, using then different visualization techniques to explore the resulting patterns. Consequently, the latter tend to treat the mining process as a black box while our approach and prototype system ELOQUENCE, attempts to open the box, reveal the process and allow a user to intervene and steer it Additional key strengths of ELOQUENCE are the fol lowing. First, it combines two visual views, pattern tree and event sequence view, providing in this way additional context to the mining process by revealing how a selected pattern appears in the data. Second, different types of constraints are supported such as ontology level or gap constraints, and data filters. The practical usefulness of these features is demonstrated by two example use cases. Several interesting problems merit further research. First, we would like to investigate how our proposed interactive “transparent box” approach can be incorporated in other sequence mining algorithms. It would also be interesting to closely examine how the pattern-growth approach can be extended to mine soft sequential patterns, and which type of constraints and visualization techniques could be used to guide the search for such patterns. In the current status of ELOQUENCE, pattern support is computed base on the first match of the pattern in a sequence. A future step would be to extend this to also take into account the number of times a pattern appears within a sequence. Furthermore, more research is required to find ways.

VII REFERENCES

[1] C. H. Mooney and J. F. Roddick, “Sequential Pattern Mining - Approaches and Algorithms,” ACM Computing Surveys, vol. 45, no. 2, 2013.

[2] K. Vrotsou, K. Ellegard, and M. Cooper, “Exploring Time Diaries ˚ Using Semi-Automated Activity Pattern Extraction,” electronic International Journal of Time Use Research, vol. 6, no. 1, pp. 1–25, 2009.

[3] A. Perer and F. Wang, “Frequence : Interactive Mining and Visualization of Temporal Frequent Event Sequences,” in Int’l Conf on Intelligent User Interfaces. Haifa, Israel: ACM, 2014, pp. 153–162.

[4] B. C. Kwon and A. Perer, “Peekquence : Visual Analytics for Event Sequence Data,” in KDD 2016 Workshop on Interactive Data Exploration and Analytics (IDEA’16), San Francisco, CA, USA, 2016, pp. 72–75.

[5] P. J. Polack, S.-T. Chen, M. Kahng, K. de Barbaro, M. Sharmin, R. Basole, and D. H. Chau, “Chronodes: Interactive Multi-focus Exploration of Event Sequences,” CoRR, vol. abs/1609.0, 2016.

[6] J. Han, J. Pei, and X. Yan, “Sequential pattern mining by patterngrowth: principle and extensions,” Studies in Fuzziness and Soft Computing, vol. 180, pp. 183–220, 2005.

[7] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman, “LifeLines: visualizing personal histories,” in CHI ’96:

Proc. of the SIGCHI conference on Human factors in computing systems. New York, NY, USA: ACM, 1996, pp. 221–227.

[8] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, and S. Murphy, "Aligning Temporal Data by Sentinel Events : Discovering Patterns in Electronic Health Records," CHI 2008 Proceedings · Health and Wellness, pp. 457–466, 2008.

[9] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith, "Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison," IEEE Transactions on Visualization and Computer Graphics, vol. 15, no. 6, pp. 1049–1056, 2009.

[10] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, "Temporal event sequence simplification," IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 2227–36, 2013.

[11] S. Guo, K. Xu, R. Zhao, D. Gotz, H. Zha, and N. Cao, "Eventthread: Visual summarization and stage analysis of event sequence data," IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 56–65, Jan 2018.

[12] J. A. Fails, A. Karlson, L. Shahamat, and B. Shneiderman, "A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories," in IEEE Symposium on Visual Analytics Science and Technology, 2006, pp. 167–174.

[13] K. Wongsuphasawat and D. Gotz, "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization," IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 12, pp. 2659–2668, 2012.

[14] D. Gotz and H. Stavropoulos, "Decisionflow: Visual analytics for high-dimensional temporal event sequence data," IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1783–1792, Dec 2014.

[15] M. Monroe, R. Lan, J. Morales del Olmo, B. Shneiderman, C. Plaisant, and J. Millstein, "The Challenges of Specifying Intervals and Absences in Temporal Queries : A Graphical Language Approach," in CHI 2013, 2013, pp. 2349–2358.