Order Dependency in Sequential Correlation

Kazi Fahim Ahmed Nasif
dept. of CSE
RUET
Rajshahi, Bangladesh
nasif.ruet@gmail.com

Md. Nurul Absur
dept. of ICE
BUP
Dhaka, Bangladesh
nurulabsur11@gmail.com

Md. Al Mamun
dept. of CSE
RUET
Rajshahi, Bangladesh
head.cse@vu.edu.bd

***Abstract*—Data analytics is a core part of the fast-moving world in the modern era, and its origin is data mining. SSequential pattern mining is a cutting-edge studies location of facts mining. An easy sample mining technique is inadequate to apprehend the connection among the objects in a dataset. A correlation degree may be extra beneficial to decide the dependency among the entities. We suggest a method in this paper that can evaluate the relationship's strengths and the degree of order. While assessing the ranking of dependency, we use one way and reverse connections. Moreover, To determine the order of dependency, we use one method and contrary relationships. Sequential correlations are widely applicable in the sector of diagnosing disease, optimize the internet, and develop active retail strategies.**

***Index Terms*—Sequential pattern, Frequent pattern, Order dependency, Sequence database, Sequential Correlation.**

I. INTRODUCTION

Data mining is a phenomenon that represents the procedure of extracting intriguing (implicit, antecedently unknown, non-nugatory, unorganized, and potentially subsidiary) information or patterns from a plethora of information sources and sizably voluminous information repositories like a relational database, XML repository, data warehouse, etc. The term data mining is a misnomer due to the fact the goal is to extract facts from uncooked and unorganized points and rework it into a suitable shape for diverse besides use, now no longer data extraction.

Correlation measures and sequential pattern mining are two of the most active research fields in the data mining community. For a conventional correlation measure in the transaction database, the results are all about items that are used together frequently. Simultaneously, the effects of sequential pattern mining are about which items are used in a particular order. Sequential patterns suggest the order of object use, at the same time as correlations constitute the degree of dependency among them.

Sequential patterns rarely constitute information; this is beneficial to the end-users. When the minimal support threshold is high, in general, most effective evident, not unusual place sense knowledge can be found. But whilst the minimum support is low, a wide variety of patterns will commonly be introduced. In this instance, maximum data is redundant, irrelevant, uninformative, or just random combinations of popular data objects. We need measures for correlations to determine sequences that are useful for end-users and have one-way order dependency.

It is essential to determine regularities from item set sequences; it is necessary to describe how they are correlated. Since large numbers of useless information are available in every science sector, it is crucial to find authentic correlation relationships among the relevant data objects. Mining correlations has been identified as important facts mining project for its many benefits over common mining styles to triumph over vain records difficulty. Another confusion concerning sequential sample mining is to discover the suitable minimal help threshold to revoke records loss. Because if we apply a high support threshold, we will get the most apparent sequences, so we will lose information. On the alternative hand, if we use a low aid threshold, we can accumulate a massive variety of patterns, and for this reason, we can face an issue in filtering out which designs are important to the involved end-user.

Our contributions in this paper are-

- In our proposed system, both the relationship's strength and the degree of order between variables can be measured.
- The construction of a null-invariant measure, SC Score, can measure the degree of order between the items in a sequence.
- We proposed an algorithm for using this measure to extract interesting pat-terns based on the SC Score threshold.

II. LITERATURE REVIEW

A. Mining Frequent Sequential Patterns

One of the foremost popular methods for mining frequent sequential patterns is PrefixSpan [1]. From a sequence database and also the minimum support threshold, sequential pattern mining is to seek out the entire set of sequential patterns within the database. The prefixSpan approach adopts divide and conquer. one in every of the foremost popular methods for mining frequent sequential patterns is PrefixSpan [1]. Construct a sequence database and the minimum support threshold; sequential pattern mining helps find the whole set of sequential patterns within the database. PrefixSpan approach adopts a divide and conquers, the pattern-growth principle as fellow: Sequence databases are recursively projected into a set of smaller projected databases based on the current sequential pattern(s), and sequential patterns are grown in

978-1-7281-6410-6/19/$31.00 ©2019 IEEE

each projected database by exploring only locally frequent fragments. Agrawal and Srikant first introduced the sequential pattern mining problem in [2] ”Given a group of sequences, where each sequence consists of an inventory of elements and every element consists of a group of things and given a user-specified minsupport threshold, sequential pattern mining is to search out all of the frequent subsequences, i.e., the subsequences whose occurrence frequency within the set of sequences isn't any but minsupport”.

B. Mining Frequent Correlated Sequential Patterns

One of the most complicated matters in the data mining sector is pattern mining with real data sets. A considerable number of co-occurrence patterns are usually generated, a majority of them either redundant or not informative. With this aim in mind, various impressive measures were examined to select the appropriate one(s), which can disclose sequential patterns' succinct relationships. In that case, the implementation of a new algorithm called PSBSpan. This successful algorithm is based on the pattern-growth methodology paradigm, which mines often correlated patterns of sequential. The basic idea of this approach is based on the likelihood that things will appear together as a sequence, rather than separately [3]. If the conditions are satisfied by the previous sentence, then a regular series is associated. The likelihood ratio assists in evaluating an upper bound prefix and suffix, which can be determined for series. With a new measure of 31 PrefixSpan is very similar to Mining Sequential Correlation. Mining Sequential Correlation adds that a specific frequent sequence is selected according to the correlation threshold. It will consider a recurring series at each step if its prefix/suffix upper-bound crosses a correlation threshold.

C. Similarity Measure for Sequential Patterns

The available measures are not adapted to the determination of specific characteristics of sequential patterns. The calculation of sequential patterns, a similarity measure, is introduced named S2MP. This method adopts the characteristics, properties, and the semantics of sequential patterns into account [4]. S2MP reflects an average of both ratings. These two scores are known as the mapping score, cumulatively. This score is given based on the similarity of two sequences, and an order score is given based on sequence object order and location.

D. CMRules

Associative rule mining is a commonly used and important problem of knowledge extraction involving the discovery of linkages between objects in a database. If the time information of a sequence database SD is ignored or deleted, we obtain a transaction database SD for each sequence database SD and its subsequent transaction database SD [5].

Each sequential rule

$$r : X => Y$$

of S has a corresponding association rule

$$r' : X => Y \ in \ S'$$

III. Related Works

Sequential pattern mining is difficult due to the need to consider several potential subsequence combinations. One of the algorithms frequently used for recurrent sequential patterns in mining is PrefixSpan [6]. Then we study an algorithm for PSBSpan sequential patterns correlated to mining [7] and then an algorithm for similarity measures between sequential patterns S2MP [8], which takes the characteristics and semantics of sequential patterns into account. CMRules [9] focuses on mining sequential rules common to multiple sequences and the final paper is Sequence Classification based on Interesting Models.

IV. Dataset

To portray the importance of sequential correlation, we created our real-life dataset. We collected the tech product purchase history of 30 users using the website [10]; we made a website with 16 items and requested the users to select the items they purchased in the past four years sequentially. We logged the sequence of items, and each sequence represents the purchase history of a customer. To keep things simple, we used items as the building block of a sequence instead of item sets. We used a minimized version of our custom dataset, containing ten distinct elements and ten sequences for demonstration purposes.

We also used some common databases which are available in real-life sources.

- Kosarak: Dataset Kosarak [11] is a vast online dataset. It includes 990,000 click-stream data from the log of an online news portal in Hungary. The dataset can be found online in its original format [12].
- BMS-Webview1: BMS-Webview1 dataset [13] is a sequence database that comprises click-stream data from an e-commerce website over several months.
- BMS-Webview2: BMSWebView2 dataset [14] is the second set of data used in the 2000 KDD-CUP competition. It contains 77,512 click-stream sequences of the data.
- Bible: Bible data-set [15] is a conversion of the Bible into a sequence database (each word is an item). It contains 36,369 sequences and 13905 distinct instances.
- Leviathan data-set: Leviathan data-set [16] is a conversion of Thomas Hobbes' novel Leviathan (1651) as a database of sequences (each word is an item). It includes 5834 sequences and 9025 unique individuals with an average of 33.8 items per string. The mean number of various things per series is 26.34.

V. Equations

Let F (A, B) be such a function that generates the frequency or total sequence count in the data-set where A was in place before B. A and B can be an item or an item identified here. So the Sequential Correlation score becomes:

$$SC(A, B) = \frac{F(A, B) - F(B, A)}{F(A, B) + F(B, A)}$$

This score will always have a value in the [-1, 1] range, as A, B are frequent products. means items are autonomous and means objects have an extreme reliance on order. A negative score stands for frequent reverse sequence.

VI. PROPOSED METHOD

There are many useful and efficient algorithms and approaches available regarding this type of scenario. But the optimization of the proper threshold value and practical calculation of sequential correlation are the successors of the proposed model of this paper. The proposed model can be divided into some significant steps. They are:

- Mine frequent sequences and Record the frequencies.
- Calculate sequential correlation score between items in the sequence.
- Score is always in range [-1,1].
- Predict their degree of independence or dependence based on score.

One sample example can make this steps understandable. Suppose A= Smart phone ,B= Earphone ,F(A, B)= 8 and F(B, A)= 2 then,

$$Sequential\ Correlation,\ SC(A,B) = \frac{8-2}{8+2} = 0.6$$

A sequential correlation score = 0 means the items are independent, —SC—= 1 means they have a very high degree of order. If the value is negative, that means the reverse sequence is more interesting. Below are some sample observations for different values of SC.

VII. OBSERVATION & RESULT

The projected algorithm performed differently in various real-life data-sets. These performance results are based on multiple relevant metrics.

A. AdL Dataset

This data-set is based on the information regarding the ADLs performed by two users daily in their own homes. This data collection contains two data instances, each corresponding to a different consumer and summarizing up to 35 days of completely labeled data.

Some Observations in terms of this dataset:

1) Dense Data-set.
2) At 3 percent support threshold and SC.Threshold=0.1, 94.12 percent of patterns are order dependent.
3) Significant change at SC.Threshold 0.1,0.2 and 0.3, so many weakly dependent patterns.

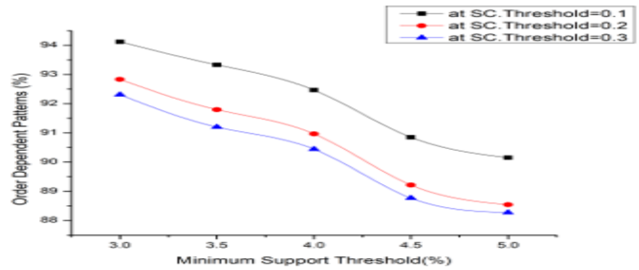

Fig. 1. Impact of SC. Threshold in ADL Dataset

B. Snake Dataset

This is a list of the proteins that bind DNA. The number of transactions is 163, and this dataset includes 22 different objects. Some Observations in terms of this dataset:

- Very dense Data-set.
- At 60 percent Support Threshold and SC.Threshold=0.1, 89.69 percent of patterns are order dependent.
- No change at SC.Threshold 0.1,0.2 and 0.5, either strong order dependency(SC 1) or none at all(SC=0).

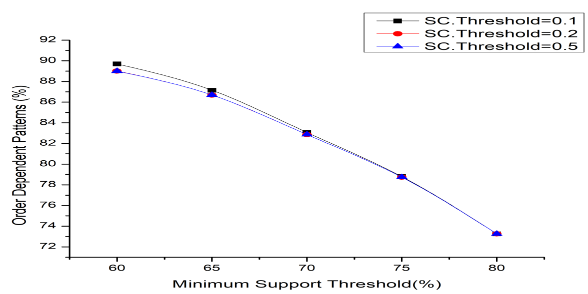

Fig. 2. Impact of SC. Threshold in Snake Dataset

C. MSNBC Dataset

This is Web page type dataset. The number of transactions is 9,89,818 and this dataset has a total of distinct items 17. This dataset is known for its instances of high precision.

Some Observations in terms of this dataset:

- At 1 percent Support Threshold and SC. Threshold=0.1, 39.13 percent of order-dependent patterns, and at SC.Threshold=0.5, only 25.91 percent of practices are Order dependent.
- No change at SC.Threshold 0.1,0.2 and 0.5, either strong order dependency(SC 1) or none at all(SC=0).

Outputs of these datasets indicate the accuracy and acceptability of the proposed model in this paper. In terms of usability and integrity, this method excelled previous approaches.

Comparing our work with the PrefixSpan algorithm in terms of runtime in the Leviathan dataset can be represented by the given comparison table.

This table reflects progressively the rise in the minimum assistance level, less the runtime. In our proposed approach

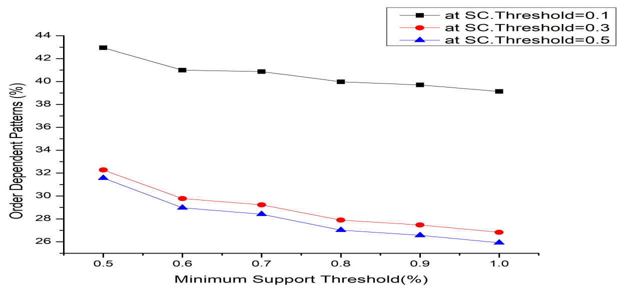

Fig. 3. Impact of SC. Threshold in MSNBC Dataset

TABLE I
COMPARISON OF RUNTIME WITH PREFIXSPAN ALGORITHM

Minimum Support Threshold (%)	PrefixSpan	Proposed method
1	17.603	0.601
2	6.723	0.189
3	3.03	0.092
4	2.178	0.056
5	2.621	0.035

the decrease in runtime is more important than the PrefixSpan algorithm.

Comparing our work with the PrefixSpan algorithm in terms of memory consumption in the Snake dataset can be represented by the given comparison table.

TABLE II
COMPARISON OF MEMORY CONSUMPTION WITH PREFIXSPAN ALGORITHM

Minimum Support Threshold (%)	PrefixSpan	Proposed method
55	80.4382	19.671
60	85.004	15.986
65	92.873	8.013
70	95.416	3.674
75	98.206	0.883

They are consuming memory for every unit minimum support threshold increases for the PrefixSpan algorithm. But our proposed model decreases memory consumption, which is efficient and desirable.

VIII. CONCLUSION

Sequential Correlation and SCScore introduce test and partition sequential patterns into sub-sets. A review of results reveals that repeated patterns are significantly reduced to a smaller number of exciting patterns. Runtime and memory consumption is well acceptable. This approach can be improved in the future to assess dependency among three or more objects.

REFERENCES

[1]] Jian Pei, Jiawei Han, Mortazavi-Asl, B., Pinto, H., Qiming Chen, Dayal, U., Mei-Chun Hsu. (n.d.). "PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth.," Proceedings 17$^{\text{th}}$ International Conference on Data Engineering.

[2] Agrawal, R., Srikant, R. (n.d.). "Mining sequential patterns., "Proceedings of the Eleventh International Conference on Data Engineering.

[3] Lin, C.X., Ji, M., Danilevsky, M., Han, J. "Efficient mining of correlated sequential patterns based on null hypothesis.," Proceedings of the 2012 International Workshop on Web-Scale Knowledge Representation, Retrieval and Reasoning, Web-KR2012. ACM, New York (2012).

[4] Saneifar, H., Bringay, S., Laurent, A., Teisseire, M. "S2MP: similarity measure for sequential patterns.," Proceedings of the 7$^{\text{th}}$ Australasian Data Mining Conference - Volume 87, AusDM 2008. Australian Computer Society Inc., Australia (2008).

[5] Fournier-Viger, P., Faghihi, U., Nkambou, R., Nguifo, E.M. "CMRules: an efficient algorithm for mining sequential rules common to several sequences.," FLAIRS Conference. AAAI Press (2010).

[6] Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H. "Pre xSpan: Mining Sequential Patterns Efficiently by Pre x-Projected Pattern Growth," ICDE '01 Proceedings of the 17th International Conference on Data Engineering, 215.

[7] Cindy Xide Lin,Ming Ji, Marina Danilevsky, Jiawei Han,,E cient. "Mining of Correlated Sequential Patterns Based on Null Hypothesis.," in Web-KR '12 Proceedings of the 2012 international workshop on Web-scale knowledge repre-sentation, retrieval and reasoning, pp. 17-24 , 2012.

[8] Hassan Saneifar , Sandra Bringay , Anne Laurent , Maguelonne Teisseire. "S2M P : similarity measure for sequential patterns.," Proceedings of the 7th Aus-tralasian Data Mining Conference, November 27-28, 2008, Glenelg, Australia.

[9] Philippe Fournier-Viger , Usef Faghihi , Roger Nkambou , Engelbert Me-phu Nguifo. ", CMRules: Mining sequential rules common to several sequences.," Knowledge-Based Systems, v.25 n.1, p.63-76, February, 2012.

[10] Custom Dataset Collection Available at: http://cart-2bfc0.firebaseapp.com [Accessed 16 Feb. 2019].

[11] Kosarak Dataset in SPMF Form Available at: http://www.philippe-fournier-viger. com/spmf/datasets/Kosarak_converted.txt [Accessed 16 Feb. 2019].

[12] Kosarak Dataset in Original Form Available at http://fimi.ua.ac.be/data/ [Accessed 16 Feb. 2019].

[13] BMS1 Dataset in SPMF Form Available at: http http://www.philippe-fournier-viger. com/spmf/datasets/BMS1_spmf [Accessed 16 Feb. 2019].

[14] BMS2 Dataset in SPMF Form Available at: http://www.philippe-fournier-viger. com/spmf/datasets/BMS2.txt [Accessed 16 Feb. 2019].

[15] Bible Dataset in SPMF Form Available at: http://www.philippe-fournier-viger. com/spmf/datasets/BIBLE.txt [Accessed 16 Feb. 2019].

[16] Leviathan Dataset in SPMF Form Available at: http://www.philippe-fournier-viger. com/spmf/datasets/BIBLE.txt [Accessed 16 Feb. 2019].