AN EFFICIENT METHOD TO DISCOVER RARE PATTERNS USING HYPER-LINKED DATA STRUCTURE

A Practice School Report submitted to
Manipal Academy of Higher Education
in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

In

Computer Science & Engineering

Submitted by

Mohammed Tariq
150905397

Under the guidance of

SHWETHA RAI
Assistant Professor - Senior Scale
Department of Computer Science & Engineering
Manipal Institute of Technology

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

May 2019

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Manipal
09-05-2019

CERTIFICATE

This is to certify that the project titled **AN EFFICIENT METHOD TO DISCOVER RARE PATTERNS USING HYPER-LINKED DATA STRUCTURE** is a record of the bonafide work done by **Mohammed Tariq** (*Reg. No. 150905397*) submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech.) in **COMPUTER SCIENCE & ENGINEERING** of Manipal Institute of Technology, Manipal, Karnataka, (A Constituent Institute of Manipal Academy of Higher Education), during the academic year 2019.

Shwetha Rai
Associate Professor-Senior Scale,
CSE Dept. M.I.T, MANIPAL

Prof. Dr. Ashalatha Nayak
HOD, CSE Dept.
M.I.T, MANIPAL

ACKNOWLEDGMENTS

I would like to take this opportunity to express my gratitude towards everyone who has supported me to complete this project successfully.

I offer my sincere appreciation for the learning opportunities provided by the Dr. D. Srikanth Rao, Director, MIT Manipal.

I express my thanks to Head of the Department Prof. Dr. Ashalatha Nayak for permitting me to use lab facilities available in department of computer science for my project work.

I really grateful to my project guide Mrs Shwetha Rai, for her continued support and encouragement throughout this process.

ABSTRACT

Rare pattern mining has emerged as a convincing field of research throughout the years. Experimental results from literature illustrate that tree-based methodologies are most efficient among the rare pattern mining strategies. Regardless of their importance and suggestion, tree-based methodologies become inefficient while dealing with sparse data and data with short patterns and also suffer from the limitation of memory. Hyper linked data structure proposed in the paper [1] is inefficient in terms of time and memory. In this study, an efficient rare pattern mining technique has been proposed that employs a hyper-linked data structure to overcome the shortcomings of hyper linked data structure proposed in the paper [1]. The hyper-linked data structure technique enables dynamic adjustment of links during the mining process that reduces the space overhead, reduces the number of scans and performs better with sparse datasets.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

This chapter discuss the term rare pattern mining and few key points related with rare pattern mining. Then this chapter will also discuss how rare pattern mining plays a vital role in solving the problems persist in the present day scenario and motivation for the selection of this project. The objectives and the expected results are stated.

1.1 Introduction to the area of work

Data mining, also called knowledge discovery in databases, in computer science, is the process of discovering interesting and useful patterns and relationships in huge amount of data [2]. The field combines tools from statistics and artificial intelligence with database management to analyse large digital collections, known as data sets. Data mining is a major area where the researchers carry out their research work. Pattern mining can be used to extract frequent and rare patterns from the large dataset. Rare Pattern mining is a technique used to extract rare patterns from the dataset.

1.2 Present day scenario

For a considerable period of time, pattern mining research was restricted only to the extraction of frequent patterns disregarding the mining of rare patterns. Rare patterns have proved to be of vital importance in a wide range of applications. Considering the significance of rare patterns, research on rare pattern mining is increasing rapidly and a considerable amount of work has already been carried out for the extraction of these momentous patterns. In recent years, Pattern mining technique has played an important role in solving many data mining tasks. Appreciable amount of work has been accomplished for the extraction of rare patterns. The applications of rare pattern mining can be found in different fields such as network anomaly detection, health care, fraud detection etc. It uses several techniques to mine the data such as image, videos, numeric values etc. from huge databases to search for a pattern that is useful in understanding the behaviour of the data in the area of health care, supermarket, stock market etc.

1.3 Motivation

Pattern mining is a major area where tremendous works have been done over the past years. Handling sparse datasets is a severe research issue that needs utmost attention. This work is therefore an attempt to resolve the issues in context of rare pattern mining. The experimental results of the Hyper-Linked Rare Pattern Mining (Hyper-Linked RPM), as claimed by the author, is faster compared to tree based approaches [1]. But the algorithm takes at least three database scans to discover the transactions that contain rare pattern. If the database is large then scanning the database will take a lot of time. An efficient method is required to overcome this problem.

1.4 Objectives

The basic goal. Therefore, project has following objectives.

- To implement the existing algorithm.
- To develop an algorithm which is more time efficient than existing algorithm
- To compare the proposed algorithm with the existing algorithm.

1.5 Target Specifications

In this project, the existing algorithm [1] is modified to such that it takes a smaller number of scans to extract rare patterns from the dataset. A comparison is made between the existing and proposed algorithm and the results are analysed to check whether the proposed algorithm shows an improvement in execution time over the existing algorithm.

1.6 Project work schedule

- *January 2019*
  - Literature survey.
- *February 2019*
  - Synopsis Submission.
  - Synopsis presentation.
- *March 2019*
  - Worked on Hyper linked Data structure.
- *April 2019*
  - Implementation of existing algorithm.
  - Implementation of proposed algorithm.
- *May 2019*
  - Comparison between existing and proposed algorithm.
  - Report Submission

1.7 Organization of the project report

This report contains five chapters' viz., introduction, background theory, methodology, result analysis and conclusion.

Chapter 1(Introduction)

This chapter will give a brief introduction to rare pattern mining and present day scenario with respect to rare pattern mining. Then the motivation, objectives are stated with a highlight on project schedule and organization of the report.

Chapter 2(Background theory)

This chapter will give a brief introduction to the project title and a bit more in brief about the prior works in the area of rare pattern mining and summarized outcome of literature review.

Chapter 3(Methodology)

This chapter will have a detailed explanation of existing algorithm and proposed algorithm along with the flowchart and implementation details of the same.

Chapter 4(Result analysis)

This chapter will discuss the results of the existing and proposed algorithm with the observed outputs of both algorithm and comparison of both in terms of time complexity.

Chapter 5(conclusions)

This chapter provides the brief summary of the work and overall conclusions of the project. It also discuss the future scope of the work.

CHAPTER 2
BACKGROUND THEORY

This chapter will give a brief introduction on the project and its title and a bit more in brief about the background theory and prior works done in the area of rare pattern mining.

2.1 Introduction to the project title

The project title is "An efficient method to discover Rare Patterns using Hyper Linked Data Structure". This project deals the extraction of rare patterns from the dataset by implementing new algorithm which is more efficient when compared to existing algorithm.

2.2 Literature review

Since its inception, an appreciable amount of work has been accomplished for the extraction of rare patterns. This section illustrates the previous works in the field of rare pattern mining. The techniques of rare pattern mining available in the literature either follow a level-wise approach or generate candidates like Apriori [3] or use efficient data structure and extract rare patterns without generating candidates like FP-Growth [4].

Liu et al. in [5] made the initial attempt using an Apriori based approach to generate the rare patterns by assigning a minimum support to each item individually. ARIMA [6] and AfRIM [7] on the other hand, carried a single support threshold for extracting the rare patterns. Considering the shortcomings of Apriori based approaches, few other techniques have adopted.

Tree based approach is most popular and efficient pattern mining approach referred as Rare Pattern Tree (RP-Tree) which was proposed in [8]. The algorithm uses two support thresholds and takes into consideration of those transactions that consists of at least one rare item. RP-Tree algorithm is further enhanced for better performance using multiple support thresholds in [9]. Tree based approaches achieve good compression in case of dense datasets as dense dataset contains many frequent items. Also, tree based approaches work well with data having long patterns but fails miserably when the data have short patterns.

In the paper [1], an efficient rare pattern mining approach has been proposed that generates appreciable results in case of sparse datasets and data with long patterns and also requires less memory compared to the eminent tree based rare pattern mining approaches. This paper presents the proposed rare pattern mining approach called Hyper-Linked Rare Pattern Mining (Hyper-Linked RPM) with the help of a suitable example for better understandability. The proposed approach follows a hyper-linked data structure [10] for extracting rare pattern. [1]Experimental results illustrate that it is faster and outperforms the pattern growth and level-wise approaches in many cases. This is because the proposed technique employs memory-based data structure.

2.3 Conclusions

Rare patterns have proved to be of vital importance in a wide range of applications. Considering the significance of rare patterns, researchers are increasing their work on rare pattern mining and a considerable amount of work has already been carried out for the extraction of these patterns. Different authors have proposed different techniques to extract patterns from the dataset. Each of the techniques discussed in the previous section has its own merits and drawbacks. In the paper [1], as claimed by the author, is faster compared to other approaches, it takes a greater number of database scans which in turn increases the time consumption. So new algorithm is proposed which is more time efficient than the existing algorithm proposed in the paper [1].

.

CHAPTER 3
METHODOLOGY

This chapter discuss the methodology of both algorithms and implementation details of same.

3.1 Existing Algorithm

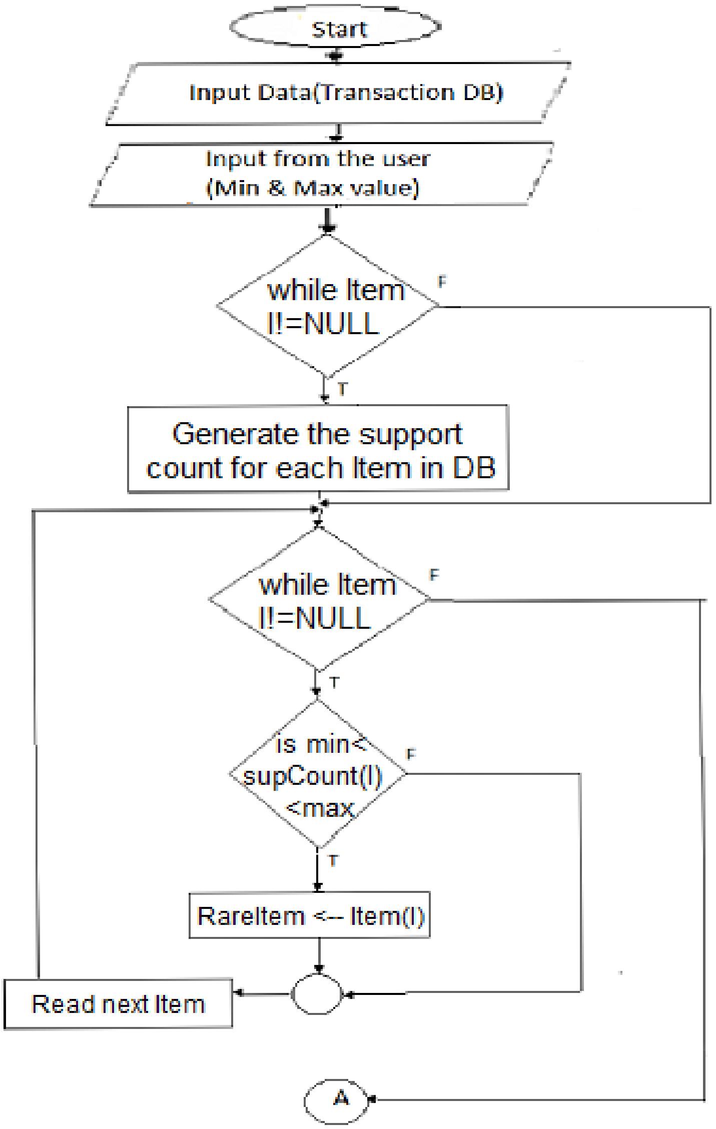


Fig 3.1 Existing Algorithm

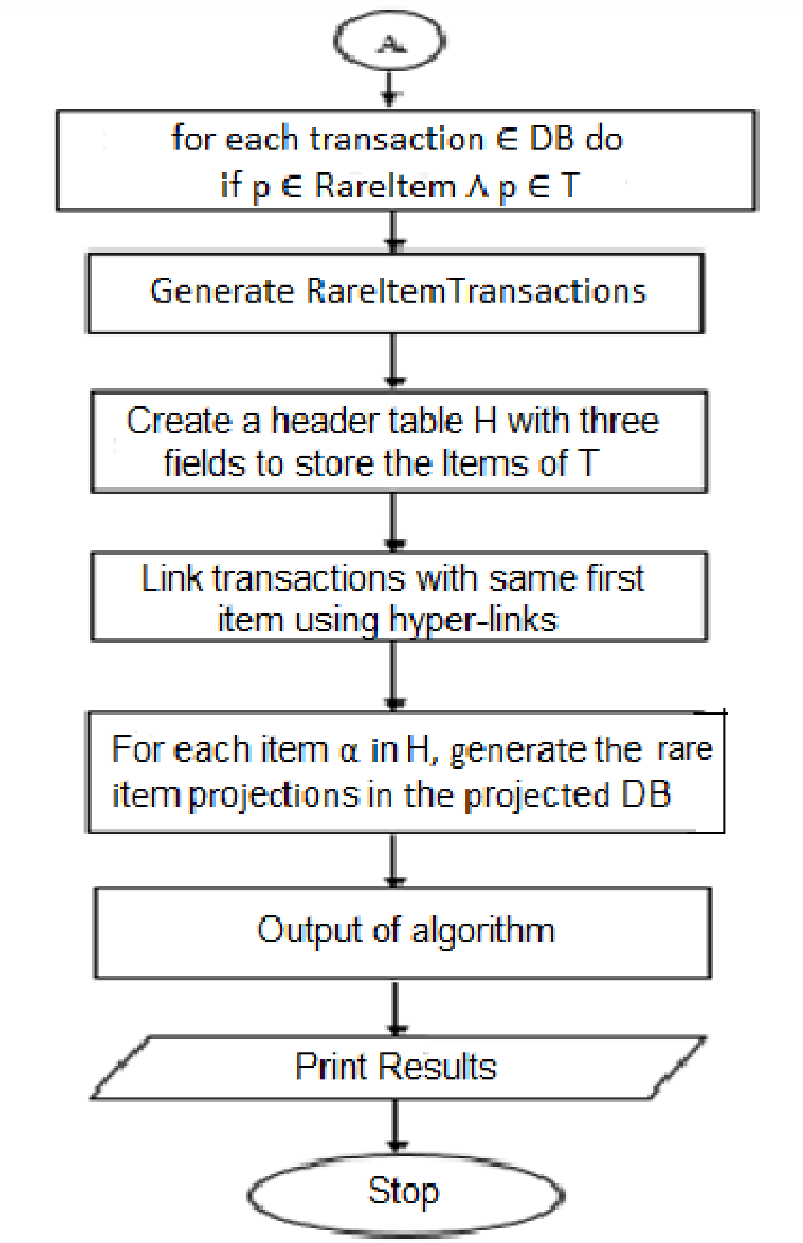

A
for each transaction ∈ DB do
if p ∈ RareItem ∧ p ∈ T
Generate RareItemTransactions
Create a header table H with three fields to store the Items of T
Link transactions with same first item using hyper-links
For each item α in H, generate the rare item projections in the projected DB
Output of algorithm
Print Results
Stop

Fig 3.1 Existing Algorithm continued

The existing algorithm showed in the figure 3.1 generates the support count or frequency of occurrence of every item during the initial scan of database. To distinguish between frequent and rare items, two support thresholds freqSup and rareSup have been used, which is taken from the user. The support count of rare itemset (SupCount) must lie between these two support thresholds, i.e. freqSup> SupCount(I) >rareSup. During the second scan, for building the hyper-linked data structure, it checks the support count of each item in the database. If the support count lies between two thresholds, such items will be referred as rare items and it will be transferred to RareItem table. During the third scan, only those transactions will be considered that involves at least one rare item. The reason being, transactions containing only frequent items have no contribution in the rare pattern mining process. The items of the reduced database will be stored in a header table H having three fields: item id, hyper-link and support count of items. It is to be noted that the support counts of the items in the header table will be their support counts in the original database. Support counts of items in the reduced database will not be considered. While loading the transactions into memory, transactions with the same first item are stored in linked list and linked together using hyperlinks. The heads of the linked lists will be represented by the items of the header table. For each item α in H, rare item projections will be generated in the α-projected database.

3.2 Proposed Algorithm

In the proposed algorithm showed in the figure 3.2, is a modified version of existing algorithm uses a header table with three fields: item id, support count and Boolean. The Boolean value consists of either zero, one or two which refers whether it is frequent or rare. During the initial scan it reads each item and increment its support count if it is present in the table. If it is not present in the table it will create an entry for item in header table and assigns its support count as 1 and Boolean as 0 which is novice. During each scan if the item is in header table and if it its support count lies between two support thresholds given by the user, i.e. freqSup> SupCount(I) >rareSup , then its Boolean will be 1 which is referred as Rare item, if not then Boolean is marked as 2 referring frequent item. By this we can reduce the number of database scan. Then rest of the algorithm follows same approach as mentioned in the existing algorithm.

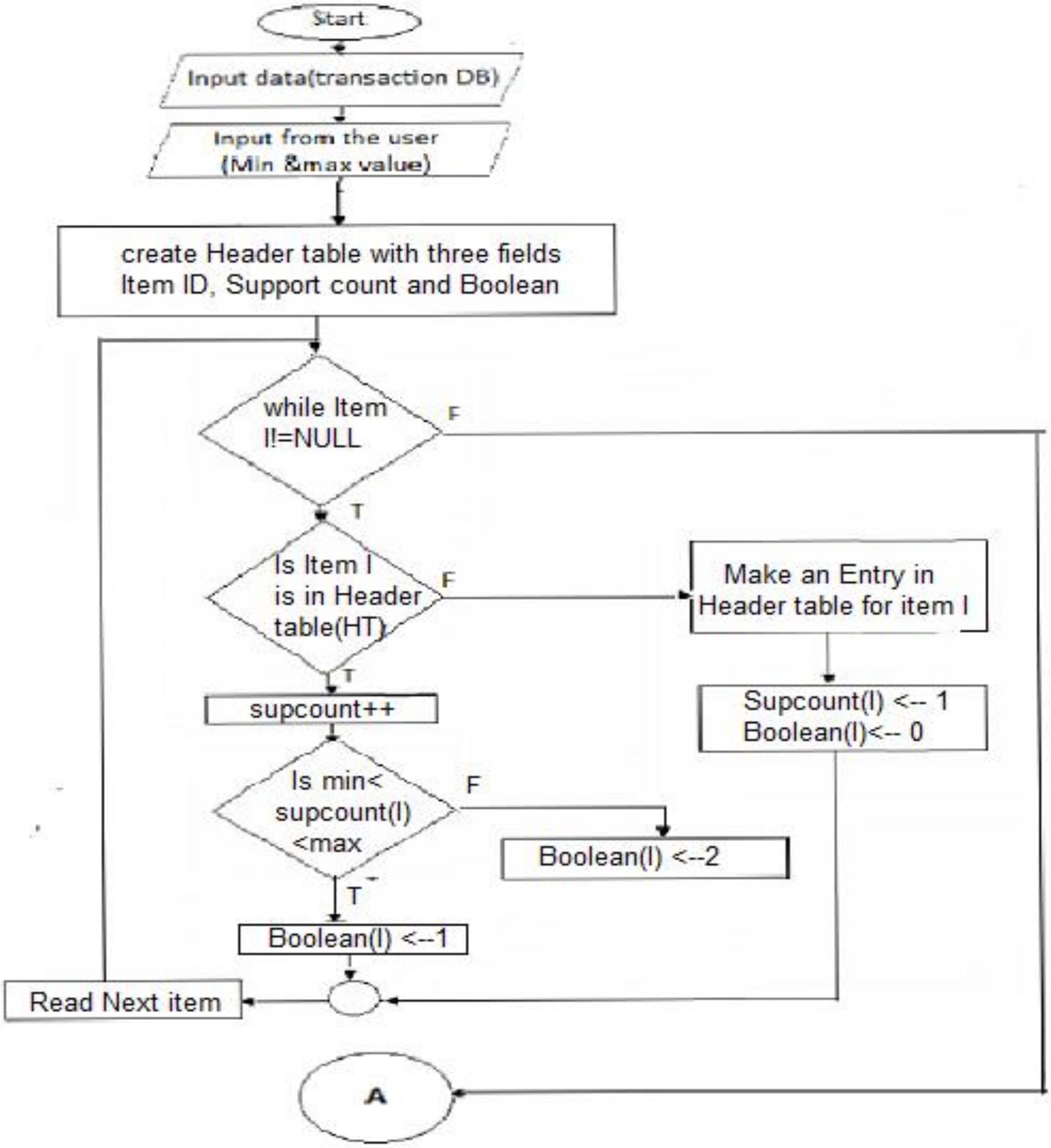


Fig 3.2 Proposed Algorithm

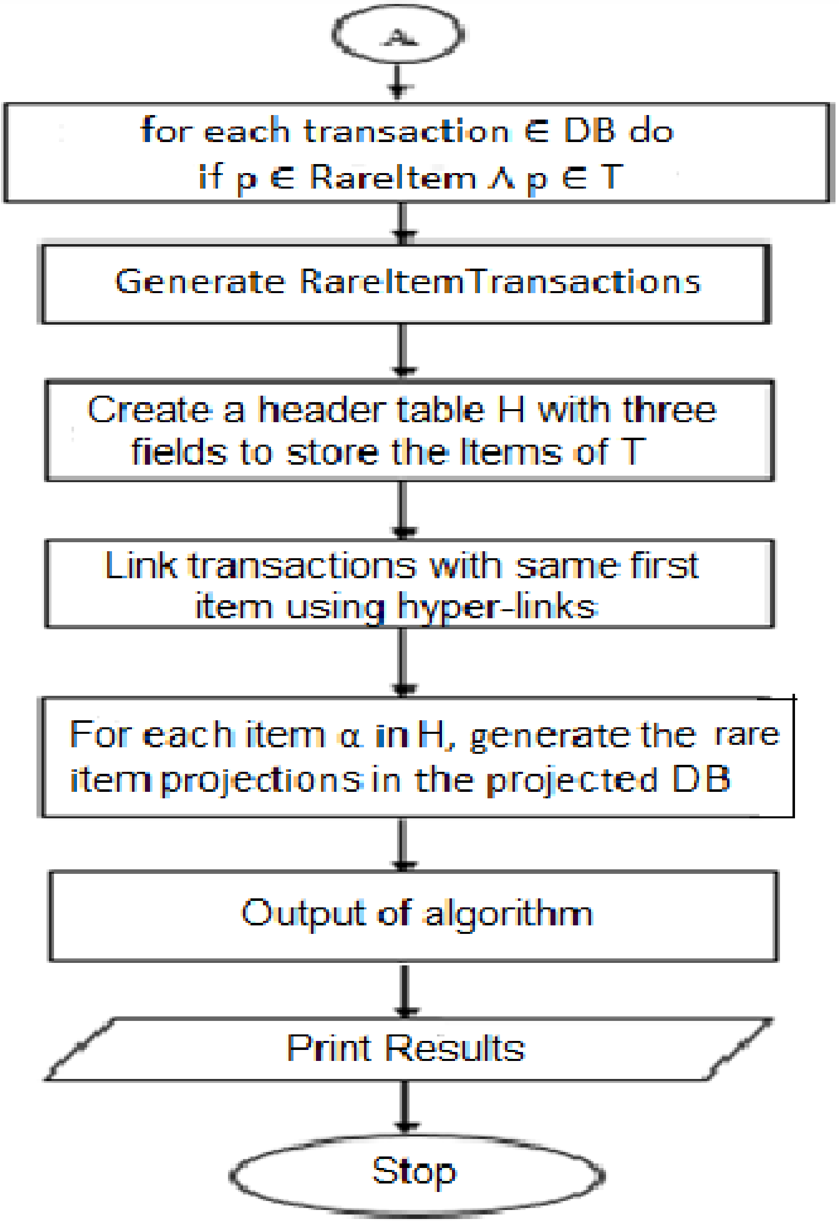

A
for each transaction ∈ DB do
if p ∈ RareItem ∧ p ∈ T
Generate RareItemTransactions
Create a header table H with three
fields to store the Items of T
Link transactions with same first
item using hyper-links
For each item α in H, generate the rare
item projections in the projected DB
Output of algorithm
Print Results
Stop

Fig 3.2 Proposed Algorithm continued

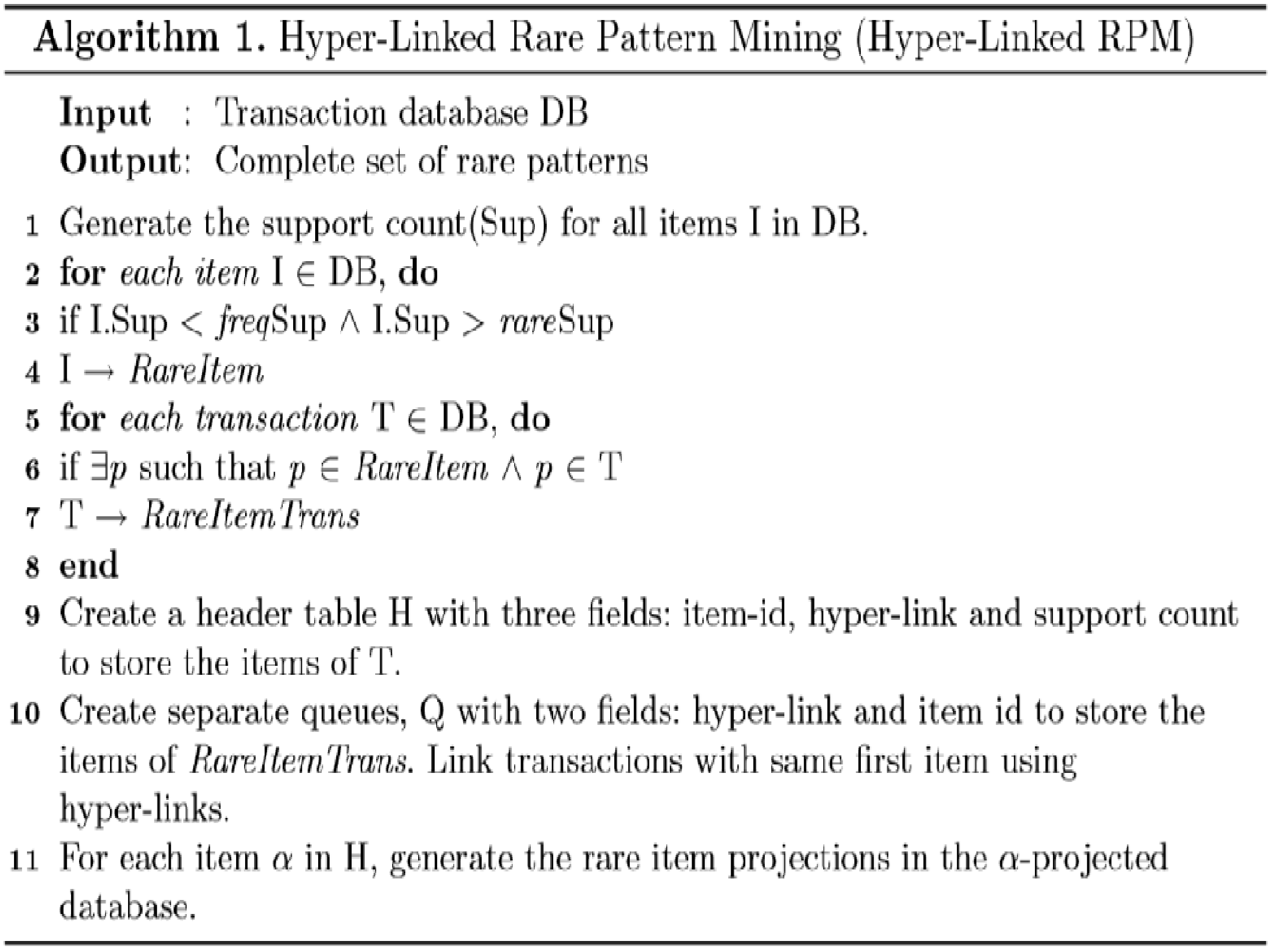
3.3 IMPLEMENTATION DETAILS

3.3.1 Tools used

- Visual Studio 2017 for C programing.

3.3.2 Implementation Description

Existing Algorithm:



Algorithm 1. Hyper-Linked Rare Pattern Mining (Hyper-Linked RPM)

```
Input  : Transaction database DB
Output : Complete set of rare patterns
Generate the support count(Sup) for all items I in DB.
for each item I ∈ DB, do
if I.Sup < freqSup ∧ I.Sup > rareSup
I → RareItem
for each transaction T ∈ DB, do
if ∃p such that p ∈ RareItem ∧ p ∈ T
T → RareItemTrans
end
Create a header table H with three fields: item-id, hyper-link and support count
  to store the items of T.
Create separate queues, Q with two fields: hyper-link and item id to store the
  items of RareItemTrans. Link transactions with same first item using
  hyper-links.
For each item α in H, generate the rare item projections in the α-projected
  database.
```

Proposed Algorithm:

An efficient method to discover rare patterns using Hyper-Linked Rare Pattern Mining (Hyper-Linked RPM)

```
Input: Transaction database
Output: Complete set of rare patterns
1. Create header table (HT) containing the following fields: item id, sup count, Boolean to indicate rare/ frequent.
2. For each item $I \in DB$, do
    If item I is in the HT
        Increment the support count of item I in the HT
        If the support count (I) > MinSup and support count (I) < Maxsup
            Mark Boolean (I) as 1(Rare)
        Else
            Mark Boolean (I) as 2 (Frequent)
    Else
    If item I not in the HT
        Create entry for item I in HT
        Support count is 1
        Boolean (I) is 0
End
3. For each transaction $T \in DB$, do
If $\exists p$ such that p is Rare $\wedge\, p \in T$
$T \rightarrow$ RareItemTrans
End
Storing the value of T in HT with three fields
Create queues with two fields: to store the items of Rare Item Transaction. Link transactions which has same first item using hyper-links.
  Generate the rare item projections in the $\alpha$-projected database for each item in HT.
```

3.4 Conclusions

The existing algorithm uses three database scans for the extraction of rare patterns from the dataset. During its initial scan it generates the support count and in its second scan it compares the support count with two threshold to generate the rare items. In the proposed algorithm instead of taking two scans for generation of support count and rare items, it is done in one scan by introducing one extra field Boolean which will identify the items whether it is rare or frequent in one scan. So with this methodology we are reducing a database scan which in turn reduces the time required for the extraction of rare items.

CHAPTER 4
RESULT ANALYSIS

To gauge the effectiveness of proposed approach, this chapter illustrates the performance comparison of both existing and proposed algorithm on the basis of time. This chapter also discuss the significance of result obtained from both algorithms.

4.1 Result analysis

To make the study more convincing, a comparison between existing and proposed algorithm is done with data set of 1000 transactions starting with 100 transaction. Total of 10 outputs were analysed with increment of 100 transaction each time and results were analysed on the basis of execution time of both the algorithm. Below table 4.2 shows he observed output of both the algorithms and a graphical analysis is provided in Fig 4.2.

Table 4.1: Observed results of both algorithm.

	100	200	300	400	500	600	700	800	900	1000
Execution time of existing algorithm in seconds	0.545	1.04	1.913	2.159	2.8	3.363	3.683	4.439	4.889	5.492
Execution time of proposed algorithm in seconds	0.427	0.83	1.339	1.6	2.214	2.446	2.811	3.348	3.811	4.093

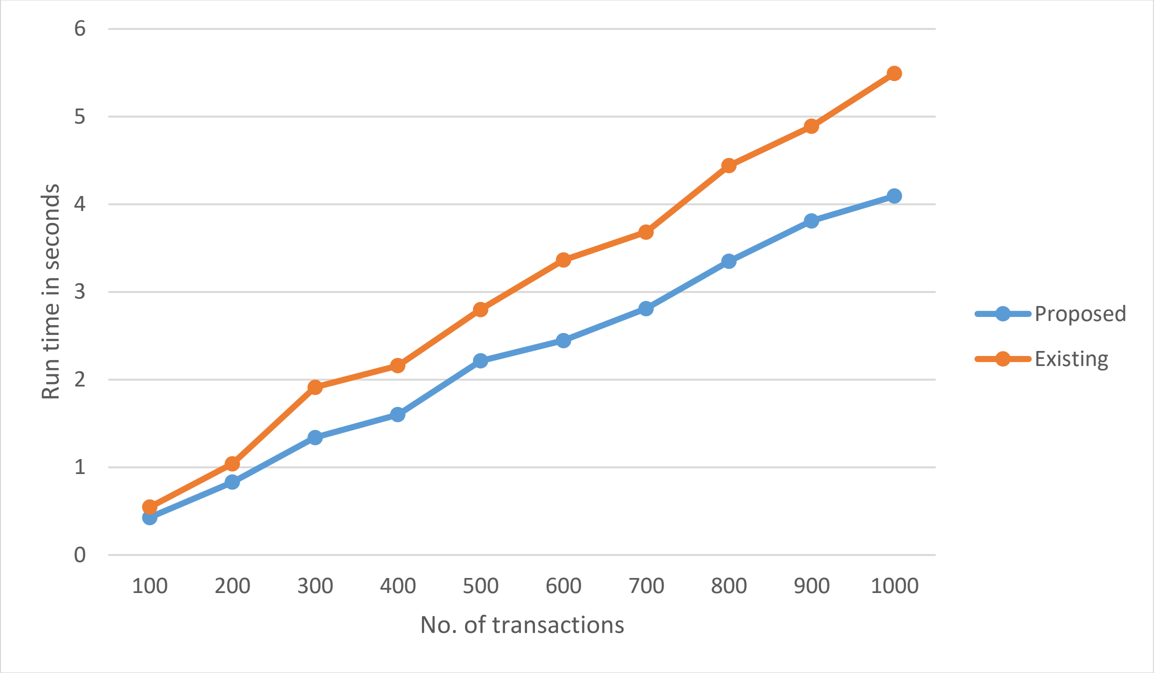


Fig 4.1 Comparison of both algorithms

4.2 Significance of the result obtained

The figure 4.2 shows that the proposed algorithm is faster and performs better than the existing algorithm as the number of transactions increases. We believe that the main reason for this big difference in performance is because of reduction of database scan in the proposed algorithm.

4.3 Conclusions

According to the result analysis the existing algorithm takes more time as number of transaction increases when compared to proposed algorithm. This is because proposed algorithm uses less number of database scans. Since it takes less time compare to existing algorithm, proposed algorithm is efficient in terms of time complexity.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 Brief summary of the work

Over the years there has been extensive research in the field of pattern mining. Pattern mining techniques have mainly considered the mining of only frequent patterns, neglecting the rare ones. Recent studies illustrate the significance of rare patterns in a wide range of application areas. As the number of research increasing in the field of rare pattern mining, there are lot of improvements in the techniques for mining rare patterns. Although the results of their work claims that it is faster, it has its own drawbacks.

In the paper [1] they have introduced a new technique for rare pattern mining which uses hyper linked data structure. The author of the paper [1] claims that it is faster and efficient it takes at least three database scans. So this paper present a new efficient algorithm for mining rare patterns using queue based hyper linked data structure which dynamically adjusts the links during the mining process that reduces the space overhead, reduces the number of scans and performs better with sparse datasets.

5.2 Conclusions

Rare pattern mining has established its significance in front of the data mining community. Hyper linked data structure approach is efficient among the rare pattern mining techniques. However, they takes more number of database scans for extraction of rare patterns. This paper, introduces a new algorithm for rare pattern mining that employs a queue based hyper-linked data structure. Performance evaluation given in Sect. 4 elicits the fact that the proposed method is more time efficient and faster than the existing algorithm proposed in the paper [1]. However, for dense datasets, other approaches outperforms the proposed approach.

5.3 Future scope of work.

Rare pattern mining is a relatively less explored area than frequent pattern mining. The growing urge for rare patterns in various domains indicates that the field of rare pattern mining is emerging extensively and there is much room for expansion.

The algorithm proposed in this project has reduced the number of database scans from three to two. As a future work, the number database scans can be reduced to one and store the data in a way that it doesn’t compromise with memory usage for sparse dataset.

REFERENCES

[1] Borah A., Nath B. (2017) Mining Rare Patterns Using Hyper-Linked Data Structure. In: Shankar B., Ghosh K., Mandal D., Ray S., Zhang D., Pal S. (eds) Pattern Recognition and Machine Intelligence. PReMI 2017. Lecture Notes in Computer Science, vol 10597. Springer, Cham

[2]. https://www.britannica.com/technology/data-mining

[3]. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proceedings of 20th International Conference on Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)

[4]. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Discov. 8(1), 53–87(2004)

[5]. Liu, B., Hsu, W., Ma, Y.: Mining association rules with multiple minimum supports. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 337–341. ACM (1999)

[6]. Szathmary, L., Napoli, A., Valtchev, P.: Towards rare itemset mining. In: 2007 19th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2007, vol. 1, pp. 305–312. IEEE (2007)

[7]. Adda, M., Wu, L., Feng, Y.: Rare itemset mining. In: 2007 Sixth International Conference on Machine Learning and Applications, ICMLA 2007, pp. 73–80. IEEE (2007)

[8]. Tsang, S., Koh, Y.S., Dobbie, G.: RP-Tree: rare pattern tree mining. In: Cuzzocrea, A., Dayal, U. (eds.) DaWaK 2011. LNCS, vol. 6862, pp. 277–288. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23544-3 21

[9]. Bhatt, U., Patel, P.: A novel approach for finding rare items based on multiple minimum support framework. Proc. Comput. Sci. 57, 1088–1095 (2015)

[10]. Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., Yang, D.: H-Mine: hyper-structure mining of frequent patterns in large databases. In: 2001 Proceedings IEEE International Conference on Data Mining, ICDM 2001, pp. 441–448. IEEE (2001)

PROJECT DETAILS

Student Details			
Student Name	**Mohammed Tariq**		
Register Number	150905397	Section / Roll No	A/30
Email Address	tariqsheik786@gmail.com	Phone No (M)	9686463519
Project Details			
Project Title	**An efficient method to discover rare pattern using Hyper linked data structure**		
Project Duration	4 months	Date of reporting	14-01-2019
Internal Guide Details			
Faculty Name	**Shwetha Rai**		
Full contact address with pin code	Dept of Computer Science & Engg, Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA		
Email address	shwetha.rai@manipal.edu		