**Rare Pattern Mining**

**Karl Friesen, Xuanming Su and Alan Bridgeman**

**Abstract.**

1. **Introduction**

Data mining is representing to discover previously unknown, but potentially useful information from one or more database repositories. As from its definition, a non-trivial extraction of implicit, previously unknown, and potentially useful information from data. Traditional data mining is mainly focused on frequent patterns, but lately, there are more researches that are focusing on rare patterns mining. Not only frequent patterns will be useful for multiple areas, but in some cases, rare patterns will also be useful (Szathmary, Napoli and Valtchev 2007).

* 1. **Background Information**

Data mining is a sub-area under subject of Database. A database is an organized collection of related data, usually stored on disk. Database is a combination of databases and database management systems. Frequent patterns are represented those frequent patterns appeared in a collection of datasets, as for the opposite area, rare patterns are those patterns which are not like frequent patterns. Rare patterns will not appear as frequent as frequent patterns within a database. Another word, rare patterns can also be discussed as exceptions with some interesting information (Huang, Koh and Dobble 2012). Perfectly rare itemsets are those itemsets with support lower than minimum support (Huang, Koh and Dobble 2012). The reason is represented some difficulties of data mining algorithms that why rare pattern mining get more attentions lately (Szathmary, Napoli and Valtchev 2007).

* 1. **Possible Real-Life Applications**

In area of biology and medicine, rare patterns mining can be found as a most suitable real-life applications (Szathmary, Napoli and Valtchev 2007). In some studies, some diseases can be associated with some life styles. For example, the cause of cardiovascular diseases can be found less connection from someone who is vegetarian. Even cardiovascular disease and vegetarian are both frequent itemset, but the pattern of cardiovascular disease and vegetarian are more rare than other patterns (Szathmary, Napoli and Valtchev 2007). In the studies of pharmacology, rare pattern mining can also have its contribution in detection of drug effects. On the other word, studies of adverse effects of a drug can be translated into studies of rare patterns mining (Szathmary, Napoli and Valtchev 2007). Rare pattern mining can also be used in studies of healthy cohorts. With some rare association rules that found from health records database, experts could associate some diseases with some rare conditions, like race and environmental factors (Szathmary, Napoli and Valtchev 2007).

1. **Related works**

As Huang et al mentioned in their paper, there are five algorithms, MS-Apriori, Rarity, ARIMA, AfRIM, and Apriori-Inverse (Huang, Koh and Dobble 2012). All these algorithms are based on technique from Apriori, which will require candidate generation and pruning.

* 1. **MS-Apriori**

As an extension of Apriori, it is using a bottom-up approach, which is similar to Apriori (Huang, Koh and Dobble 2012). By looking for rare itemsets, they can be retrieved as those itemsets with a low minimum support value. Then from the pruning step, rare itemsets can be stored as candidates to construct the rare itemsets and to produce the associate rules.

* 1. **Rarity**

Huang et al describe Rarity algorithm as start by looking for the longest transaction, then search for rare itemsets by using a top-down approach through the transaction sets. According to Huang et al, this approach will avoid search through lower layers, which will only contain frequent itemsets (Huang, Koh and Dobble 2012).

* 1. **ARIMA**

Szathmary et al proposed an approach of mining through database and retrieve rare patterns (Szathmary, Napoli and Valtchev 2007). They split the algorithm into two tasks, one is to generate all minimal rare generators by using an algorithm that named as MRG-Exp; the other task is to retrieve a complete set of rare itemsets which is called A Rare Itemset Miner Algorithm (A.K.A ARIMA).

* 1. **AfRIM**

AfRIM was proposed by Adda et al, Huang et al describe it as generate candidates by combining common k-itemset subsets with all rare k+1-itemset pairs (Huang, Koh and Dobble 2012).

* 1. **Apriori-Inverse**

The main purpose is for mining perfectly rare itemsets (Huang, Koh and Dobble 2012). As introduced in section 1.1, perfectly rare itemsets are those itemsets with support is lower than minimum support.

* 1. **Summary**

According to Huang et al, all the algorithms as describe above are descendant from Apriori algorithm, which will require expensive candidate generation and pruning process (Huang, Koh and Dobble 2012). There will be some issues of performance and memory usage. As Szathmary et al described, as the limitation of ARIMA would be the algorithm requires storing all rare itemsets that could be expensive for memory space ().

1. **Our approach**
   1. **Foundation of our approach**
   2. **Improvements from foundation**
2. **Experiment results**
   1. **Time Performance**
   2. **Memory Usage**
3. **Discussion**
4. **Conclusion and Future work**

# Bibliography

Huang, David, Yun Sing Koh, and Gillian Dobble. "Rare Pattern Mining on Data Streams." Edited by A. Cuzzocrea and U. Dayal. *DaWaK 2012*, 2012: 303-314.

Rina N., Patel, and Khushboo Trivedi. "An Efficient Rare Interesting Item Set Mining using Modified MCCFP-Growth." *International journal of Emerging Technologies and Engineering* 1, no. 4 (May 2014): 87-91.

Szathmary, Laszlo, Amedeo Napoli, and Petko Valtchev. "Towards Rare Itemset Mining." *19th IEEE International Conference on Tools with Artificial intelligence*, Oct 2007: 305-312.