

Temporal Knowledge Discovery with Infrequent Episodes*

Dan Li, Liying Jiang, Jitender S. Deogun
Dept. of Computer Science & Engineering
University of Nebraska – Lincoln, Lincoln, NE 68588-0115

Abstract

In this paper, we present an efficient algorithm which discovers rare episodes with a combination of bottom-up and top-down scanning schema. The information sharing between bottom-up and top-down scanings helps prune candidate episodes, and thus, efficiently find infrequent episodes that are interesting to user. We evaluate the performance of the algorithm using real-life weather databases. We observe from experimental results that our approach results in 30%-90% reduction in computation time and 25%-75% reduction in the number of candidates comparing with Apriori algorithm.

1 Introduction

Drought is a natural process of Great Plains landscapes and results in significant economic, social, and environmental impacts. Thus, through the National Science Foundation (NSF) Digital Government program, we are in the process of developing an advanced Geospatial Decision Support System (GDSS) to improve the quality and accessibility of temperature and precipitation data for drought assessment and drought risk management [4]. A common question in risk analysis is “How are events related in time?” Data mining algorithms have the potential to identify these relationships. The basis of all relationship detection by data mining is essentially association rule mining [1]. Frequent itemsets play an important role in many mining tasks that try to find interesting patterns from databases. However, there exist applications where the items (or events) occur infrequently, but their occurrences provide important information. For example, consider the following scenario, “If El Niño occurs, with 0.01% possibility, drought event will occur”, although the occurrence of drought event seems infrequent, the link between El Niño and this hazard still cannot be ignored. In this paper, we develop and evaluate efficient algorithms to facilitate knowledge discovery with infrequent events.

2 Mining Infrequent Episodes

Although our goal is to discover infrequent episodes from time series databases, the existing algorithms for finding the *frequent set*, i.e., the set of all frequent itemsets, can help us meet the goal. Most of these algorithms follow the basic idea in the Apriori algorithm [1], and work in a *bottom-up, breadth-first* manner [2, 3, 5]. The algorithms start from frequent 1-itemsets at the bottom and then extend one level up in every pass until no more frequent itemsets can be generated. The principle of these algorithms are based on the following property: *If an event set is infrequent, all its super-event sets must be infrequent*. This is called *upwards closed* property of event sets. Different from previous work, we are interested in event sets such that their supports not only meet the lower bound requirement, but also satisfy an upper bound. These requirements force us to develop a new algorithm which can efficiently discover infrequent episodes that are interesting-to-user. A *downwards closed* property of event sets is discovered and applied to our algorithm: *If an event set is frequent, all its sub-event sets must be frequent*. With these two properties, our algorithm works in a hybrid way which combines *bottom-up* and *top-down* methods.

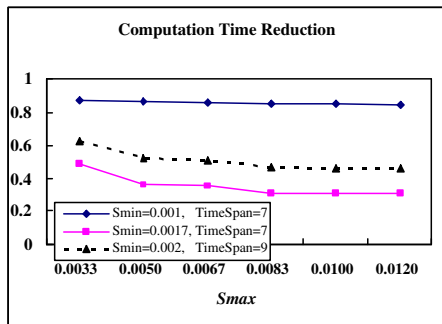
We develop a two-way (bottom-up and top-down) scanning algorithm *INFER* (INFrequent Episode generatoR) for discovering infrequent episodes which satisfy pre-specified support thresholds. In this algorithm, the most important work is to efficiently generate candidate sets for the next pass. To facilitate the process of candidates generation, the frequent episode set discovered by top-down search is used to generate candidate event sets for bottom-up search. This is based on downwards closed property discussed above. Accordingly, based on upwards closed property, the weak episode set discovered by bottom-up search is used to generate candidate event sets for top-down search. The information sharing between bottom-up and top-down scanings helps reduce the number of candidates for both scanning processes.

3 Experiments & Analysis

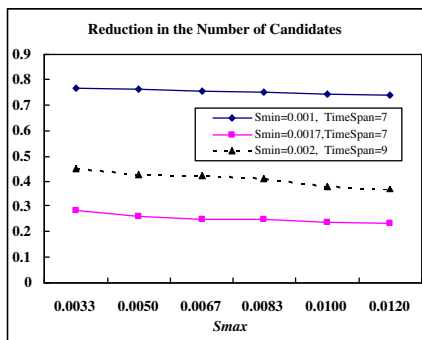
Experiments are designed to find infrequent episodes (e.g., drought) from weather related databases, and dis-

*This research was supported in part by NSF Digital Government Grant No. EIA-0091530, USDA RMA Grant NO. 02IE08310228, and NSF EPSCOR, Grant No. EPS-0091900.

cover relationships between rare weather events and environmental indices. Data is collected at the automated weather station in Clay Center, NE, from 1950-1999.



(a) Reduction in Computation Time.



(b) Reduction in the Number of Candidates.

Figure 1. Reduction in Computation Time and the Number of Candidates.

Figure 1 (a) summarizes the computation time comparison between *INFER* and *Apriori*. The value of time reduction is the ratio of time difference between the two algorithms to the computation time for *Apriori*, i.e., $\text{time-reduction} = (T_{\text{Apriori}} - T_{\text{INFER}}) / T_{\text{Apriori}}$. We test on three different minimum support thresholds S_{min} . For each S_{min} , we observe that the time reduction increases as S_{max} decreases. When we fix the maximum support threshold, the time reduction is also inversely proportional to S_{min} if the maximum time span does not change. The reduction in the computation time is even more obvious when the two support thresholds (S_{min} and S_{max}) are lower. In Algorithm *INFER*, the information sharing between bottom-up and top-down searching schemas helps prune more candidates in each pass, thus making the algorithm more efficient. The algorithm is much more efficient, especially when the support thresholds are lower, although in this case more preliminary candidates are generated, *INFER* can sharply reduce the number by filtering out more candidates. These results show that *INFER* algorithm is robust for infrequent episodes discovery. Overall, our algorithm results in 30%-90% reduction in computation time comparing with *Apriori*.

Figure 1 (b) shows the reduction in the number of candidates generated from *Apriori* to *INFER*. We can see that it presents the same performance as we have discussed for the reduction in computation time. Overall, our algorithm results in 25%-75% reduction in the number of candidates comparing with *Apriori*. However, one thing worth to say is that *INFER* notably outperforms *Apriori* when two support thresholds are small, this further shows the advantages of our algorithm for discovering infrequent episodes.

4 Conclusion

Knowledge discovery in temporal databases has been an active field of study. Many existing algorithms focus on temporal association rule mining. However, little work has been done on discovering *infrequent* episodes, and it is not easy to discover association rules with low *support* but high *confidence* by existing algorithms. This paper presents a new approach for discovering infrequent (consequent) episodes in time series databases.

We develop Algorithm *INFER* to discover infrequent episodes. It is based on two closure properties of event sets. It combines the bottom-up and top-down searching schemas to efficiently discover candidate episodes. The information sharing between the two searching methods helps prune candidates, and makes the entire algorithm more efficient. The performance of our algorithm is tested on real-life database. We observe that our approach results in 30%-90% reduction in computation time, 25%-75% reduction in the number of candidates comparing with *Apriori* algorithm, and the reduction is inversely proportional to the two support thresholds (S_{min} and S_{max}). Thus, our algorithm is reasonable and suitable for discovering infrequent episodes.

References

- [1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD 1993 Intl. Conf. on Management of Data*, pages 207–216, Washington D.C., 1993.
- [2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings. 20th International Conference on Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 1994.
- [3] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proc. of 1995 Int'l Conf. on Very Large Data Bases (VLDB'95)*, Zürich, Switzerland, September 1995, pages 420–431, 1995.
- [4] S. Harms, D. Li, J. Deogun, and T. Tadesse. Efficient rule discovery in a geo-spatial decision support system. In *Proc. of the 2002 Natl. Conf. on Digital Gov. Res.*, pages 235–241, Los Angeles, California, USA, May 2002.
- [5] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.