

Mining itemsets – an approach to longitudinal and incremental association rule mining

C. Mooney and J. F. Roddick

School of Informatics and Engineering, Flinders University of South Australia, Adelaide, Australia.

Abstract

With the improvements in data warehousing and database technology, the amount of data being collected has grown at a remarkable rate. The field of data mining, which enables the extraction of interesting or relevant information from such data, has also grown at a similar rate. Methods now exist that enable the extraction of rules from varied data sources from which users are able to draw inferences about the underlying data.

This paper surveys and extends a new area of data mining that has recently emerged – *Rule Mining*. Rule mining uses the results of previous mining sessions as input to a second mining process that produces rules with very different semantics which can be used to extend the inferences made from the underlying data. The paper first presents an overview of the work in this area. We then present an efficient method of longitudinal association rule mining that uses previously constructed frequent itemsets (rather than either the (larger) source datasets or the (larger) resultant rulesets) within a rule production engine. We also show how this method can be used for longitudinal association rule mining.

1 Introduction

During the past decade the collection and storage of data has continued to increase at exponentially. The technology associated with warehousing and

databases in general has also improved and this has given rise to the increasing volume and complexity of data now being able to be stored by organisations [6]. This phenomenon has in turn generated a need for techniques and applications able to extract interesting or relevant information for those who require it. The field of data mining has grown to fill this need and algorithms have been developed to mine many types of data in a variety of domains.

Traditionally, such methods have been employed on static datasets, often producing a large number of rules, which must be filtered for their interestingness, relevance and meaningfulness by a combination of heuristics and manual intervention by experts in the area of the data being mined. This has spawned many areas of research including significant discussion into the determination of what is interesting [3, 11, 13, 15, 18, 20, 23–25]. At the same time, as the generation of data is ongoing, techniques for the incremental updating and maintenance of rule sets from large datasets have also been developed [7, 8, 17, 27]. Although these two areas do not have a direct relevance in determining a workable generic framework for rule mining, some of the underlying ideas have proved useful in the discovery and interpretation of rules that may be discovered from rule mining.

There has also been an increasing interest in the mining of temporal and spatio-temporal data, as is evidenced by the body of work discussed by Roddick and Spiliopoulou [19], and some of these techniques we believe would be amendable to the work discussed here and be able to provide enhanced explanations of behaviour.

There are a number of motivating factors for the development of extensions of this type. First, the cost, in time, of mining the raw data every time there is some change is considerable. Second, the new rules that are produced reflect the current state of the enterprise, and since there is often no contingency to retain the previous rules, no inference can be drawn about the changing nature of the underlying data. In contrast, a rule set produced as a result of rule mining is potentially able to be used to draw a different set of inferences - that of trends across the ordered dataset. This opens up a productive avenue for data mining giving it another direction through which knowledge may be elicited.

The remainder of this paper discusses some recent developments in rule mining including, briefly, some experiments undertaken by the authors. The paper is organised as follows. The next section will outline the context in which rule mining is set and discuss previous research in the area. We then report on some recent research in this area which mines itemsets to facilitate longitudinal association rule mining.

2 Previous Research

2.1 Context

Rule mining can be applied to many types of rule, provided an applicable measure of difference between rule sets can be established, and therefore is

not, in general, restricted to any particular area of data mining provided it has ordered (for example a temporal) component or one can be inferred.

In order to provide a context to the research outlined later in the paper we focus on rule mining in the context of association rules. This does not infer that rule mining in other areas of data mining is less applicable. Indeed the results of rule mining from any rule types should be able to be used by different algorithms to gain different insight into the underlying data. One example might be to cluster the results of second-order association rules; another may be to employ an association mining technique across the results of a clustering session to detect if there exists an association between two clusters over time. However, it is quickly apparent that any framework for mining meta-rules must have some method of determining whether a particular rule has changed [26]. This is important as the decision on the nature of the first order rules will determine the type of information that can be inferred from second and subsequent higher order rules. For our work here we also need an ordering dimension such as time or space.

2.2 Terminology

The term *rule mining* has been variously termed *meta-mining*, *higher-order mining* and *second-phase mining* (including in our previous work). It has been noted that the term data mining is itself erroneous¹. However, given that strictly speaking, conventional data mining algorithms normally produce zero-order rules with rule mining algorithms producing either first order or temporal zero-order rules, we distinguish between the sub-fields as follows:

Data Mining. Algorithmic approaches to mining rules from data.

Rule Mining. A subset of data mining in which rules are mined from the results of previous mining operations. The other terms will not be used.

Meta-Rule. The results of a rule mining operation in cases where they describe aspects of the input rules rather than the data (in the same way that database meta-data describes data).

Knowledge Discovery. An overarching framework within which data (including rule) mining algorithms may be used to extract new knowledge.

The term meta-rule has also been used to describe the form or pattern to which a discovered rule must conform. In the work of Fu and Buchanan [12] this notion was used to guide the mining process to produce rules only of the type specified by the meta-rule. This was expressed in the case of association rules by templates used to guide the discovery of rules in such a way that only rules of some general form are reported. At first glance, as is the case in many areas of research, it seems that the same term has been used to describe

¹ If other exploratory disciplines such as gold mining are taken as exemplars, the field should more properly be called rule mining (ie. mining for rules).

two different concepts. However, their use of the term is consistent with our use. They use it as an input template; we use it to describe the result².

2.3 Previous Work

Researchers have been studying temporal data mining algorithms as a means of extracting knowledge from temporal or ordered data for a while. One recent area is in the discovery of temporal patterns or relationships in a time series, often using Allen's natural temporal relationships [2], to formulate the results of a mining session [4, 5, 9, 14]. Another area has been concerned with the nature of rule changes during incremental mining and in comparing the results of different mining sessions [27].

Directly related work in this area has been conducted by Feldman et al. [10] in which frequent itemsets are used for incremental association rule maintenance and Abraham and Roddick [1] which describes a methodology for processing rulesets and the rule mining process in general. In the latter work, a meta-rule set is described as a mining operation over two or more rulesets discovered over the same data, of the same rule type, but at different times. This aggregation process then yields rules that can be categorised into four types – *New*, *Expired*, *Unchanged* and *Changed*. The first three of these can be relatively easily separated from the resultant ruleset and examined using existing techniques depending on the interest placed on them by the user. The last, changed rules, are not so easily determinable. The nature of change is determined by some predefined difference measure; ie. any change criteria must be specific for the rulesets under consideration, yet general enough to be mapped to each subsequent level of meta-rule. In general two sets of criteria need to be supplied:

1. The minimum amount of allowable change before a rule is considered to have changed.
2. The maximum amount of allowable change before a rule is considered to be a different rule.

As an example, one possible change for association rules is a greater than some threshold alteration in the support and confidence of an association rule. A change that is likely to be considered as a mutation to a new rule might be changes in antecedent or consequent. For example, if the minimum support change, $\Delta_\sigma = 5\%$ then for the following rules:

$$t_1 \quad A, B \rightarrow C, D \quad \sigma(35\%) \quad \gamma(87\%) \quad (1)$$

$$t_2 \quad A, B \rightarrow C, D \quad \sigma(37\%) \quad \gamma(91\%) \quad (2)$$

$$t_3 \quad A, B \rightarrow C, D \quad \sigma(43\%) \quad \gamma(92\%) \quad (3)$$

$$t_4 \quad A, B \rightarrow C, E \quad \sigma(33\%) \quad \gamma(86\%) \quad (4)$$

² Note that the term meta-rule when applied to rule mining was first used in earlier work [1] to describe rules that express differences in two or more rulesets.

t_1 and t_2 would be considered unchanged, t_3 would constitute the same rule (compared with either t_1 or t_2) but changed while t_4 would be considered a different rule. This issue is further explored in [26] in which the notion of ruleset *closure* as determined by the criteria under which a rule remains the same despite changes in its statistics and its contents is proposed. The framework was set in the context of mining as a temporal sequence and takes the position that two rules are the same across different mining sessions if the contents remain the same. This was articulated as the *Rule Invariance Statement* (RIS), and used it as a basis of a framework for studying meta-rules as a temporal sequence of comparable mining results [26].

Previous research in this area also arises from the fields of expert systems and artificial intelligence. For example, Schoenauer and Sebag [21,22] discuss a reduction operator which is applied to examples extracted by discovery algorithms to produce behavioural rules, and Chakrabarti *et al* [6] who consider the evolution of rules, albeit in a fairly restrictive manner. However, while there has been a relatively small amount of direct research, a number of authors have investigated topics dealing with related issues. For example, a variety of temporal mining routines have been developed that attempt to construct rules that process either an explicit temporal or spatial component or are able to handle temporal or spatial data. These include temporal series miners and sequence miners and temporal and spatial extensions to association, clustering and characterisation rule algorithms [16,19].

Rule mining has a number of desirable characteristics, which include combining mining strategies through the modular combination of components, providing, in a natural way, for the development of alternative explanations in describing facts about data, particularly those describing changes over time, location or some other dimension and comparatively faster execution time due to reduced volumes of data. Rule mining demands the clear and unambiguous interpretation of rules obtained. That is, the semantics of the resultant rules must be carefully determined. Informally, the use of different combinations of data and rule mining algorithms will produce different interpretations.

3 Mining Frequent Itemsets

The rest of this paper discusses research that fits broadly within the framework discussed above and renders tractable otherwise difficult applications. The approach we take here is to mine sequences of frequent itemsets, which provides a number of opportunities including longitudinal and incremental data mining³.

Mining frequent itemsets provides a number of advantages. First, frequent itemsets, if stored with their support, contain all the information required to

³ Space precludes a full discussion of all aspects, however our website, <http://kdm.first.flinders.edu.au>, provides further details, including further details of the experimental results.

dynamically generate, at any arbitrary confidence, association rules without a further scan of the data. Second, the generation of frequent itemsets can be done autonomously and offline at times of low system usage. Third, the storage requirements of frequent itemsets is much reduced and thus enables a lower support threshold to be used for storage purposes (although as we will see, the reporting threshold may be higher). The process is as follows:

1. **Itemset Generation.** As the parameter setting for frequent itemset generation is a one-time decision, this can be done autonomously and thus the process can operate either offline and/or at times of low system usage. Moreover, it can operate incrementally, with new itemsets being added to the itemset collection at any time⁴.
2. **Itemset Analysis.** In this phase one or more of the following analyses can be performed.
 - (a) **Incremental Rule Maintenance.** The latest itemset is taken as an amendment to the current ruleset. Space precludes a discussion of this technique, but see [10].
 - (b) **Longitudinal Analysis.** The frequent itemsets are taken as an ordered sequence. Details are discussed in section 3.1.
 - (c) **Outlier Analysis.** Outlier datasets, ie. datasets that do not conform to patterns indicated by other datasets, can be identified. This is especially useful if the datasets represents spatially disparate datasets, such as, for example, sales locations. This may be combined with longitudinal analysis to show the market leaders and followers.
 - (d) **Rule Structure Analysis.** RSA enables the structure of the rules to be interrogated and can yield insights into the changes in the cohesion of data values (such as the number of items making up the longer itemsets) or the use of hierarchies (if a multi-level routine was used). This process is not described in this paper.

3.1 Longitudinal Analysis

Given that rules characterise a dataset and that these can be created from frequent itemsets, the purpose of our experiments was to determine whether interesting longitudinal rules could be found more efficiently using ordered frequent item sets (generated individually from multiple random data sets), than from using the source data sets directly. We also suspected that short-lived but strong rules were having the effect of masking weaker but longer-lived rules and we anticipated this method could assist in finding these.

We therefore ran two sets of experiments. The first fed the results of itemset generation runs back as source data into the association rule generation algorithm. The second, as a control, concatenated the source datasets and attempted to find useful associations from the concatenated dataset.

⁴ Similarly, obsolete frequent itemsets can be removed from those analysed in the second phase.

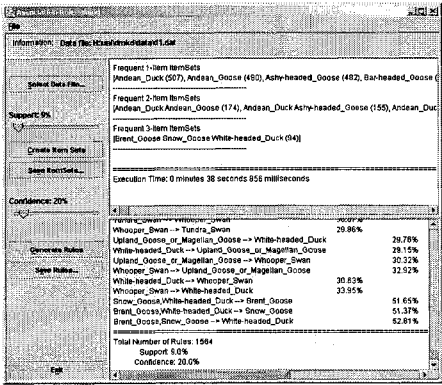


Fig. 1. Screen capture of the mining interface

Our experiments were conducted using the *Flinders Interactive Knowledge Discovery System*, FIDO. This system provided a flexible, plug-compatible system for running and visualising the results of data mining algorithms. In particular, they were conducted on an AMD 1.2 GHz Athlon with 768MB of Ram, using an Apriori-like algorithm implemented using version 1.4 of Java as shown in Figure 1. Five transaction sets were constructed as follows:

1. 10 transactions each with an average of 20 items per transaction. This group was seeded at a weight of 20%.
2. 100 transactions each with an average of 20 items per transaction. This group was seeded at a weight of 2%.
3. 1,000 thousand transactions each with an average of 20 items per transaction. This group was seeded at a weight of 1%
4. 10,000 transactions each with an average of 20 items per transaction. This group was not seeded.
5. 50,000 transactions each with an average of 20 items per transaction. This group was not seeded.

The items for each transaction were selected at random from a population of 148 values (we used species of duck) and each data set was generated using a random generator available within FIDO. Ten datasets were generated for each transaction set.

Each of the ten data sets that made up a transaction set were processed using an implementation of the Apriori algorithm available as part of FIDO. Frequent itemsets that had the potential to produce rules, (ie. those of two or more items) were saved and used as input to the next stage. Table 1 shows the average times to generate the item sets for this phase along with the support level used and the resultant average number of item sets produced.

Significantly, for each of the seeded datasets, the seeded rules were reported despite the noise level inherent in the datasets being higher.

The next phase of the experiment involved the concatenation of the ten saved item set files for each transaction set and processing them using the

Trans Set	Records per Transaction	Ave. Time (Secs)	Support (%)	Ave. Itemsets
1	200	0.04	40	150
2	2,000	1.782	10	854
3	20,000	37.56	2	2235
4	200,000	353.98	2	179
5	1,000,000	1791.05	2	4

Table 1. Results of processing seeded datasets in *rule* mining mode.

same Apriori algorithm as was used in the first phase. The support levels for this pass were set between 20% and 80% because we were interested in finding item sets that were frequent in the majority of the initial sets. Table 2 shows the results of this process. Note also that the seeded transactions were inconsistently reported. They were either not reported at all or, at best, they were buried in a large number of itemsets.

Trans Set	Records	Resultant Itemsets	Ave. Time (Secs)	Support (%)
1	1498	7	0.231	70
2	8541	2	0.987	80
3	22325	40	3	70
4	7233	1	0.921	60
5	34	4	0.01	20

Table 2. Results of processing seeded datasets in *data* mining mode.

4 Results

The experiments showed that the identification of useful rules was easier using the itemset mining method. The datasets created were smaller and the time taken to perform the analysis was shorter. Moreover, as we suspected, some seeded rules were not found using the ‘big bang’ approach as they were lost in the background noise.

Moreover, while some of the same results can be obtained by processing the entire transaction file, if further transactions are added this will result in the entire file having to be processed again. This could take considerable time even if methods of incremental mining were used. Our method involves only the processing of the new transactions and adding them to the frequent item set file and reprocessing that - a considerably lower cost in time.

Although this research has not been undertaken in this form before, these results reinforce the related research undertaken by Feldman et al. [10] in incremental rule mining which suggests the utility of such a method could also be useful in incremental association rule mining.

5 Future Work

We have shown that the methodology of mining itemsets is sound and that potential rules can be extracted using this method. Apart from the further

investigation of the areas discussed in Section 3 such as outlier detection and rule structure analysis, one important area of future research will be in the semantic deconstruction of the mined itemsets to investigate further the meaning of the rules being discovered. For example, while the semantics of a static rule can be easily understood, the meaning of higher order rules can require some additional thought on the part of users. For example, depending on the situation, the temporal rule

$$A - before \rightarrow B \quad (q) \quad (5)$$

can be interpreted as event A occurs before B (with some qualification, q). Or some event including the objects of type A occurs before events including those of type B . Or that rules containing A occur before the same rules contain B , and so on. This needs some investigation.

References

- [1] T. Abraham and J.F. Roddick. Discovering meta-rules in mining temporal and spatio-temporal data. In J. Fong, editor, *8th Int. Database Workshop, Data Mining, Data Warehousing and Client/Server Databases (IDW'97)*, pages 30–41, Hong Kong, 1997. Springer-Verlag.
- [2] J.F. Allen. Maintaining knowledge about temporal intervals. *CACM*, 26(11):832–843, 1983.
- [3] R.J. Bayardo Jr and R. Agrawal. Mining the most interesting rules. In S. Chaudhuri and D. Madigan, editors, *5th Int. Conf. on Knowledge Discovery and Data Mining*, pages 145–154, San Diego, CA, 1999. ACM Press.
- [4] C. Bettini, X. S. Wang, and S. Jajodia. Mining temporal relationships with multiple granularities in time sequences. *Data Eng. Bull.*, 21(1):32–38, 1998.
- [5] C. Bettini, X.S. Wang, S. Jajodia, and J.-L. Lin. Discovering frequent event patterns with multiple granularities in time sequences. *IEEE Trans Knowl. Data Eng.*, 10(2):222–237, 1998.
- [6] S. Chakrabarti, S. Sarawagi, and B. Dom. Mining surprising patterns using temporal description length. In Ashish Gupta, Oded Shmueli, and Jennifer Widom, editors, *24th Int. Conf. on Very Large databases VLDB'98*, pages 606–617, New York, NY, 1998. Morgan Kaufmann.
- [7] D.W. Cheung, J. Han, V.T. Ng, and C.Y. Wong. Maintenance of discovered association rules in large databases: an incremental updating technique. In Stanley Y. W. Su, editor, *Int. Conf. on Data Engineering (ICDE'96)*, pages 106–114, New Orleans, Louisiana, 1996. IEEE Computer Society.
- [8] D.W. Cheung, S.D. Lee, and B. Kao. A general incremental technique for maintaining discovered association rules. In *5th DASFAA*, Melbourne, 1997.
- [9] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. Rule discovery from time series. In *4th Int. Conf. on Knowledge Discovery and Data Mining (KDD-98)*. AAAI Press, 1998.
- [10] R. Feldman, Y. Aumann, A. Amir, and H. Mannila. Efficient algorithms for discovering frequent sets in incremental databases. In *Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97)*, in cooperation with *ACM SIGMOD'97*, Tucson, Arizona, 1997.

102 *Data Mining III*

- [11] A.A. Freitas. On rule interestingness measures. *Knowledge Based Systems*, 12(5-6):309–315, 1999.
- [12] L. Fu and B.G. Buchanan. Enhancing performance of expert systems by automated discovery of meta-rules. In *1st Conf. on AI Applications*, pages 107–115. IEEE Comput. Soc. Press, Silver Spring, MD, 1984.
- [13] R.J. Hilderman and H.J. Hamilton. Heuristic measures of interestingness. In J.M. Zytkow and J. Rauch, editors, *3rd European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, volume 1704 of *Lecture Notes in AI*, pages 232–241, Prague, 1999. Springer.
- [14] F. Hppner. Discovery of temporal patterns - learning rules about the qualitative behaviour of time series, online at <http://www.citeseer.nj.nec.com/446576.html>, 2001.
- [15] W. Klosgen. Efficient discovery of interesting statements in databases. *J. Intell. Inf. Syst.*, 4:53–69, 1995.
- [16] H. Miller and J. Han, editors. *Geographic Data Mining and Knowledge Discovery*. Research Monographs in GIS. Taylor and Francis, 2001.
- [17] C.P. Rainsford, M.K. Mohania, and J.F. Roddick. A temporal windowing approach to the incremental maintenance of association rules. In *8th Int. Database Workshop, Data Mining, Data Warehousing and Client/Server Databases (IDW'97)*, pages 78–94, Hong Kong, 1997. Springer Verlag.
- [18] J.F. Roddick and S. Rice. What's interesting about cricket? - on thresholds and anticipation in discovered rules. *SIGKDD Explorations*, 3(1):1–5, 2001.
- [19] J.F. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Trans. Knowl. Data Eng.*, 14(4), 2002.
- [20] S. Sahar. Interestingness via what is not interesting. In S. Chaudhuri and D. Madigan, editors, *5th Int. Conf. on Knowledge Discovery and Data Mining*, pages 332–336, San Diego, CA, 1999. ACM Press.
- [21] M. Schoenauer and M. Sebag. Incremental learning of rules and meta-rules. In *Seventh Int. Conf. on Machine Learning*, pages 49–57. Morgan Kaufmann, Palo Alto, CA, 1990.
- [22] M. Sebag and M. Schoenauer. Discovering meta-rules from rules and examples. In *11th Int. Conf. on Expert Systems and their Applications*, volume 1, pages 205–216, Nanterre, France, 1991.
- [23] D. Shah, L.V.S. Lakshmanan, K. Ramamritham, and S. Sudarshan. Interestingness and pruning of mined patterns. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, Philadelphia, 1999.
- [24] A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *1st Int. Conf. on Knowledge Discovery and Data Mining (KDD-95)*, pages 275–281, Montreal, Quebec, 1995. AAAI Press, Menlo Park, CA.
- [25] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems? *IEEE TKDE*, 8(6):970–974, 1996.
- [26] M. Spiliopoulou and J.F. Roddick. Higher order mining: Modelling and mining the results of knowledge discovery. In N. Ebecken and C.A. Brebbia, editors, *Data Mining II - Proc. Second Int. Conf. on Data Mining Methods and Databases*, pages 309–320. WIT Press, Cambridge, UK, 2000.
- [27] K. Wang and J. Tan. Incremental discovery of sequential patterns. In *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Quebec, 1996.