
ELSEVIER

Data & Knowledge Engineering 17 (1995) 245–262

DATA & KNOWLEDGE ENGINEERING

Set-oriented data mining in relational databases

Maurice Houtsma[a,*], Arun Swami[b,†]

[a]*University of Twente, Enschede, The Netherlands*
[b]*IBM Almaden Research Center, San Jose, CA, USA*

Received 25 July 1994; revised 14 March 1995; accepted 28 July 1995

Abstract

Data mining is an important real-life application for businesses. It is critical to find efficient ways of mining large data sets. In order to benefit from the experience with relational databases, a set-oriented approach to mining data is needed. In such an approach, the data mining operations are expressed in terms of relational or set-oriented operations. Query optimization technology can then be used for efficient processing.
In this paper, we describe set-oriented algorithms for mining association rules. Such algorithms imply performing multiple joins and thus may appear to be inherently less efficient than special-purpose algorithms. We develop new algorithms that can be expressed as SQL queries, and discuss optimization of these algorithms. After analytical evaluation, an algorithm named **SETM** emerges as the algorithm of choice. Algorithm **SETM** uses only simple database primitives, viz., sorting and merge-scan join. Algorithm **SETM** is simple, fast, and stable over the range of parameter values. It is easily parallelized and we suggest several additional optimizations. The set-oriented nature of Algorithm **SETM** makes it possible to develop extensions easily and its performance makes it feasible to build interactive data mining tools for large databases.

Keywords: Data mining; Optimization; Set-oriented algorithms

1. Introduction

The competitiveness of companies is becoming increasingly dependent on the quality of their decision making. Hence, it is no wonder that companies often try to learn from past transactions and decisions in order to improve the quality of decisions taken in the present or future. In order to support this process, large amounts of data are collected and stored during business operations. Later, these data are analyzed for relevant information. This process is called *data mining* [3, 18, 23, 5] or *knowledge discovery in databases* [13, 21, 15, 17]. Data

* Communicating author. M. Houtsma's research was made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences; his current affiliation is Telematics Research Centre, P.O. Box 217, 7500 AE Enschede, The Netherlands.
† Current affiliation is Silicon Graphics Computer Systems, 2011 N. Shoreline Blvd., Mountain View, CA 94043.

SSDI: 0169-023X(95)00024-0

mining is relevant to many different types of businesses. As examples, retail stores obtain profiles from customers and their buying patterns and supermarkets analyze their sales and the effect of advertising on sales. Such "target marketing" [6] is becoming increasingly important.

Different aspects of data mining have been explored in the literature. In *classification*, data units (tuples) are grouped together based on some common characteristics, and rules are generated to describe this grouping. This has been done both in the context of AI [22] and in the context of databases [2, 13, 5]. Work has been done to search for similar sequences or time series [1]. In finding *association rules*, one tries to discover frequently occurring patterns within data units [20, 4]. In this paper, we focus on the problem of finding association rules. There has been a lot of work in rule discovery that is related but not directly applicable, for example, [8, 9, 10, 12, 16, 19, 22].

Business applications deal with an uncontrolled real world, where many rules will overlap in their components and uncertainty is common [20]. Examples of rules could be: "Most sales transactions in which bread and butter are purchased, also include milk," or "Customers with kids are more likely to buy a particular brand of cereal if it includes baseball cards." Work on finding these kinds of rules has been done in AI for some specific applications (see [21] for an overview). Although the work done in AI is usually very general, the computational complexity of the proposed algorithms is high, and the algorithms are feasible only for small data sets [17]. Performance is a problem with these algorithms for the kind of applications we consider, which involve mining large databases. In [18] a small example is described of generating rules from data, but the emphasis is more on architectural issues than on performance and large data sets. In [4], the problem of rule discovery is addressed in a database context. The paper describes an algorithm for rule discovery on a large data set. However, the algorithm in [4] still has a tuple-oriented flavor (tuples are represented as strings, and the algorithm consists of string manipulation operations) and is rather complex.

Our focus is on achieving good performance in mining association rules on large datasets in relational databases. This differentiates our work from much of the work in AI. Problems of optimization of discovered rules, subsumption, etc. are beyond the scope of this paper. The perspective is one of analysis on existing data, i.e. the data is static. Though we use retailing transactions as an example application throughout the paper, the work here is applicable to mining of association rules from any domain. In an illustrative example, we assume that the items are distributed almost uniformly. However, our method does not require any such assumptions.

We address rule discovery in database systems from a set-oriented perspective. The motivations for a new approach to this problem are several. A set-oriented approach allows a clearer expression of what needs to be done as opposed to specifying exactly how the operations are carried out. The declarative nature of this approach allows consideration of a variety of ways to optimize the required operations. This means that the ample experience that has been gained in optimizing relational queries can directly be applied here. Eventually, it should be possible to integrate rule discovery completely with the database system. This would facilitate the use of the large amounts of data that are currently stored on relational databases. The relational query optimizer can then determine the most efficient way to obtain the desired results. Finally, our set-oriented approach has a small number of well-defined, simple concepts and operations. This allows easy extensibility to handling additional kinds of mining, e.g. relating association rules to customer classes.

The structure of this paper is as follows. In Section 2 we define the problem of set-oriented data mining and give an initial sketch of our approach. In Section 3 we present an initial set-oriented algorithm expressed in SQL and analyze its performance. In Section 4 we present a second set-oriented algorithm expressed in SQL and analyze its performance. In Section 4.4 we describe the latter algorithm (called Algorithm **SETM**) in terms of simple database operations: sorting and merge-scan join. We also illustrate the algorithm by means of a small example. Section 5 explains how rules are generated and how rules with multiple items in the consequent can be obtained. In Section 6 we describe several experiments we did with an implementation of our algorithm on a large data set. Section 7 describes several possible optimizations to Algorithm **SETM** that could further improve the response time. We also discuss different ways to parallelize the algorithm. Finally, Section 8 presents our conclusions.

2. Moving towards set-oriented mining

Consider the problem of finding association rules in sales data. Typically, a retail store records information for each customer transaction, where a customer transaction involves the purchase of a variable number of items. We can store this information in a relational database system using a table with the following schema: *SALES_DATA(trans_id, item)*. For each customer transaction that takes place, tuples corresponding to the items sold are inserted in *SALES_DATA*.

In order to find association rules, we need to scan transactions for reoccurring patterns that occur often enough to be of interest (this is made more precise later). We use the term *pattern* to capture the concept of *itemset* introduced in [4]. This is more in line with existing terminology [21]. A *pattern* can be defined as follows. If items A, B, and C frequently occur together in a single customer transaction, this means that the pattern ABC occurs often. This observation might allow us to conclude (among other rules) the association rule $A \wedge B \Rightarrow C$ [1]. Here, AB is called the *antecedent* of the rule and C is called the *consequent* of the rule. Usually, some constraints need to be met before we conclude that an association rule holds. As in [4], we define *support* for a pattern to be the ratio of customer transactions supporting that pattern to the total number of customer transactions. Also, the *confidence factor* for a rule obtained from a pattern is defined as the ratio of the support for the pattern to the support for the antecedent of the rule. For the rule $A \wedge B \Rightarrow C$, this would be $|ABC|/|AB|$, where $|ABC|$ denotes the support for pattern ABC.

We are interested only in association rules where the support for the pattern(s) involved in the rule is greater than some minimum value called *minimum support*. We also require that qualifying rules have a confidence factor greater than some value. In our example, patterns can be generated in a straightforward fashion by repeated joins with *SALES_DATA*. For instance, generating all patterns of exactly two items, is expressed by the following SQL query:

```
SELECT r1.trans_id, r1.item, r2.item
FROM SALES_DATA r1, SALES_DATA r2
```

[1] *Causality* is not necessarily implied. Also, the ordering of A and B in the antecedent of the rule is arbitrary.

```
WHERE r1.trans_id = r2.trans_id AND
      r1.item <> r2.item
```

For each pair of items (x, y), we count the number of transaction-ids in order to find the number of transactions supporting this pattern. All patterns of exactly three items can now be obtained by joining the result of the previous step again with *SALES_DATA* and so on. The order in which items appear is not relevant right now; (x, y) is equivalent to (y, x) since both pairs have the same support. Order only becomes important when generating the rules because confidence factors can be different for different orders.

This strategy is elaborated in Sections 3 and 4 and expressed in terms of a set-oriented query language, viz., SQL. The first expression that is generated naturally leads to nested-loop based joins. A rough analysis of its expected performance indicates that such an implementation would perform very poorly. Consequently, we generate an equivalent expression in SQL that naturally leads to sort-merge based joins. A first analysis shows it to be very promising, and we pursue this implementation in the remainder of the paper.

We include the discussion of both SQL-expressions of our strategy, because we wish to emphasize the methodology that we used in this research. Taking a set-oriented approach does not immediately lead to great results but it clearly helps in getting a good understanding of the problem. In our case, by first having studied the nested-loop strategy, we were able to develop the strategy based on sort-merge joins fairly easily, by taking into consideration the ways a relational query optimizer deals with these types of (complex) queries.

3. Strategy leading to nested-loop joins

We discuss a first formulation of our set-oriented data mining strategy. It naturally leads to nested-loopbased joins. We express the algorithm in SQL and then analyze its expected performance.

3.1. Formulation

All customer transactions are stored in the relation *SALES_DATA*. From this relation, we first generate the counts for each item x, i.e. the number of transactions that support item x. We check that the minimum support requirement is met. The output is stored in relation C_1, with schema (*item*, *count*).

```
INSERT INTO C1
SELECT r1.item, COUNT(*)
FROM SALES_DATA r1
GROUP BY r1.item
HAVING COUNT(*) >= :min_support
```

The next step is to generate all patterns (x, y) and check if they meet the minimum support

criterion. For a specific item A, this is easy to express. For example, all patterns (A, y) can be generated using a self-join of $SALES\_DATA$, as in the following SQL-expression.

```
SELECT r1.item, r2.item, COUNT(*)
FROM SALES_DATA r1, SALES_DATA r2
WHERE r1.trans_id = r2.trans_id AND
      r1.item = 'A' AND
      r2.item <> 'A'
GROUP BY r1.item, r2.item
HAVING COUNT(*) >= :min_support
```

This kind of expression only generates patterns with a specific item in the first position. The expression has to be generalized in order to generate arbitrary patterns. As stated earlier, the order of the items in a pattern is not relevant at the time of generation. The order is important only in the final rule generation process. We take advantage of this fact by generating patterns with the items in lexicographical order. For instance, we generate AB, but we do not generate BA. We generalize over all values of *item* having minimum support, by using the following SQL expression to generate *all* lexicographically ordered patterns of length k ($k > 1$) having minimum support (in case $k = 1$, the query presented before for C_1 is used).

```
INSERT INTO Ck
SELECT r1.item, ..., rk.item, COUNT(*)
FROM Ck-1 c, SALES_DATA r1, SALES_DATA r2, ..., SALES_DATA rk
WHERE r1.trans_id = r2.trans_id = ... = rk.trans_id AND
      r1.item = c.item1 AND
      r2.item = c.item2 AND
      :
      rk-1.item = c.itemk-1 AND
      rk.item > rk-1.item
GROUP BY r1.item, ..., rk.item
HAVING COUNT(*) >= :min_support
```

Relation C_k has schema ($item_1$, $item_2$, ..., $item_k$, *count*). All feasible rules are found by consecutively generating all qualifying patterns from length 1 to k until $C_{k+1} = \{\}$. Since the items in the patterns are lexicographically ordered, a single inequality test in the SQL query is sufficient.

3.2. Analysis

Before pursuing this strategy any further, we first perform a rough analysis of its expected performance. Let us consider how a relational query optimizer could optimize the final SQL expression in Section 3.1. If the joins are performed using the nested-loop join method, we need indexes to allow efficient evaluation of the join. It seems we need two indexes on table

SALES_DATA: an index on (*item*, *trans_id*) and another index on (*trans_id*). Given these indexes, the query can be evaluated as follows:

(1) Take a tuple c from C_{k-1}, and use the index on (*item*, *trans_id*) for r_1 to get qualifying tuples with $r_1.item = c.item_1$.
(2) For each of these tuples, use the index on (*item*, *trans_id*) for r_2 to get tuples that satisfy $r_2.item = c.item_2$ and $r_2.trans\_id = r_1.trans\_id$.
(3) Similarly for relations $r_3, \ldots, r_{k-1}$.
(4) Finally, use the index on (*trans_id*) for r_k to compute $r_k.trans\_id = r_{k-1}.trans\_id$ and check the remaining condition $r_k.item > r_{k-1}.item$.
(5) The qualifying tuples are sorted on the item values and the count is used to check the minimum support constraint.

Let us consider a hypothetical retailing database to characterize the performance of this strategy. There are 1000 different items that can be sold. The data consists of 200,000 customer transactions. The average number of items sold in a transaction is 10. Thus, the relation *SALES_DATA* contains about 2 million tuples. To make the analysis tractable, we assume that the items have approximately equal probability of being sold (in the actual data set, the items are not sold with equal probability). Hence, the chance of an item appearing in a particular transaction is 1%. We will assume the following characteristics for the database system. Page size is 4 Kbytes, and each item and transaction id is represented using 4 bytes (item values are represented by integers). Hence, each initial tuple consists of 8 bytes.

Consider the B^+-tree index on (*item*, *trans_id*). Since all the data is contained in the index, we do not need a pointer in the leaf page entries. Assuming little overhead, we can store up to 500 entries in each leaf page. The number of leaf pages in the B^+-tree index on (*item*, *trans_id*) is $2{,}000{,}000/500 \approx 4{,}000$. Assuming 4 bytes for a pointer, an index entry in the non-leaf pages has a size of 12 bytes. Assuming very little overhead, we can store about 333 key-value/pointer pairs on a non-leaf index page. The following inequality holds for the number of levels L of the index tree: $333^L \geq 1{,}000{,}000 > 333^{L-1}$; hence, $L = 3$. The number of non-leaf pages in this index is $(1 + 4{,}000/333) = 14$. Similar calculations for the index on (*trans_id*) show that the number of leaf pages is 2,000 and the number of non-leaf pages is 5. Since the number of non-leaf pages is small, we can assume that they reside in memory. Hence, accessing non-leaf pages does not require page fetches from disk.

Let the minimum support desired be 1000 transactions, i.e. 0.5% of the total number of transactions. On the average, each item appears in about 1% of the transactions. Assuming uniform probabilities, all items qualify as having minimum support. Therefore, the cardinality of C_1 will be 1000, i.e. each item is present.

To obtain C_2, we take each tuple c from C_1 and access the index on (*item*, *trans_id*). This requires $1\% \times 4{,}000$ leaf page fetches, i.e. ≈ 40 page fetches. The result consists of about 2,000 transaction-ids (1%). For each of the resulting transaction-ids we now have to access the index on (*trans_id*) resulting in 1 page fetch.

From this, we may conclude that the first step alone will require about $1000 \times (40 + 2000 \times 1) \approx 2{,}000{,}000$ page fetches. Most of these page fetches are random. A random page fetch costs about 20 ms. Hence, the time for the first step alone is $\approx 40{,}000$ seconds, which is more than 11 hours!

Clearly, an implementation based on nested-loop joins is very inefficient. However, one

could consider a different implementation for the same basic pattern finding strategy, viz., sort-merge joins. We will consider this strategy in the next section.

4. Strategy leading to sort-merge joins

We now discuss the second formulation of our set-oriented data mining strategy, based on using sort-merge joins. We again express the algorithm in SQL and then analyze its expected performance.

4.1. Formulation

In the previous implementation we would generate intermediate relations R_i(*trans_id*, $item_1, \ldots, item_i$), extract support information from these relations, and then discard them. But what if, after each step, we saved the last R_i that was generated? Furthermore, let us save R_i sorted on (*trans_id*, $item_1, \ldots, item_i$). We could then generate all lexicographically ordered patterns of length k using the following expression:

```
INSERT INTO R'_k
SELECT p.trans_id, p.item_1, ..., p.item_{k-1}, q.item
FROM R_{k-1} p, SALES_DATA q
WHERE q.trans_id = p.trans_id AND
      q.item > p.item_{k-1}
```

After generating all lexicographically ordered patterns of length k in R'_k, we now have to generate counts for those patterns in R'_k that meet the minimum support constraint. This can be done as follows:

```
INSERT INTO C_k
SELECT p.item_1, ..., p.item_k, COUNT(*)
FROM R'_k p
GROUP BY p.item_1, ..., p.item_k
HAVING COUNT(*) >= :min_support
```

Before we go on to generate patterns of length $k + 1$, we first have to select the tuples from R'_k that should be extended, viz., those tuples that meet the minimum support constraint. We also wish the resulting relation to be sorted on (*trans_id*, $item_1, \ldots, item_k$). This is done as follows:

```
INSERT INTO R_k
SELECT p.trans_id, p.item_1, ..., p.item_k
FROM R'_k p, C_k q
WHERE p.item_1 = q.item_1 AND
      :
```

$p.item_k = q.item_k$
ORDER BY $p.trans\_id, p.item_1, \ldots, p.item_k$

We can now repeat this process, until at some point $R_k = \emptyset$. Note that the sorting we did in the last step is not really required. It does, however, enable an efficient execution plan if the sort order of the relations is tracked across iterations.

4.2. Example

We illustrate this strategy by means of an example. The example database consists of 10 transactions where each transaction has 3 items. We require a minimum support of 30%, i.e. 3 transactions. The desired confidence factor is 70%. The customer transactions are shown in Fig. 1. For brevity, we have presented the transactions as non-normalized tuples. The algorithm, however, uses the tuple format described before; a subset of this corresponding relation is shown too. The contents of the count relation C_1 are also shown in Fig. 1. As in the algorithm we use R'_k and R_k to distinguish the R relations before and after elimination of patterns that do not meet the minimum support count. In the first iteration, R_2 is generated and sorted on items and C_2 is generated from R_2. The contents of R'_2, R_2 and C_2 are as shown in Fig. 2. In the next iteration, R_3 is generated and sorted on items and C_3 is generated from R_3. The contents of R'_2, R_2 and C_2 are as shown in Fig. 3. The next iteration will not generate any new tuples, and the algorithm terminates.

4.3. Analysis

In this section the performance of the sort-merge strategy is analyzed using the same data set as for the nested-loop strategy.

The I/O complexity of the sort-merge strategy can easily be expressed, by a formula derived as follows. Let $\|R_k\|$ denote the number of pages used to store the relation in iteration k. In the worst case, applying the minimum support constraints does not eliminate any tuples from R_k. Assume that no patterns of length n have the minimum support, i.e. the relation R_n is empty. By then, we have made $(n-1)$ passes, this means $(n-1)$ merge-scans requiring

tx_id	item	item	item
10	A	B	C
20	A	B	D
30	A	B	C
40	B	C	D
50	A	C	G
60	A	D	G
70	A	E	H
80	D	E	F
90	D	E	F
99	D	E	F

tx_id	item
10	A
10	B
10	C
20	A
20	B
20	D
30	A
30	B
30	C
...	...

item	cnt
A	6
B	4
C	4
D	6
E	4
F	3

Fig. 1. Customer transactions, corresponding relation, and relation C_1.

tx_id	item$_1$	item$_2$
10	A	B
10	A	C
10	B	C
20	A	B
20	A	D
20	B	D
30	A	B
30	A	C
30	B	C
...	...	...

item$_1$	item$_2$	cnt
A	B	3
A	C	3
B	C	3
D	E	3
D	F	3
E	F	3

tx_id	item$_1$	item$_2$
10	A	B
10	A	C
10	B	C
20	A	B
30	A	B
30	A	C
30	B	C
40	B	C
50	A	C
...	...	...

Fig. 2. Relations R'_2, C_2, and R_2.

$(n-1)\|R_1\| + \Sigma_{i=1}^{n-1} \|R_i\|$ page accesses. The number of page accesses to store the result of these merge-scans is $\Sigma_{i=2}^{n-1} R_i$. After each merge-scan, the output is read in again, sorted, and written out to disk; this requires $2\,\Sigma_{i=2}^{n-1} \|R_i\|$ page accesses. (We assume R_1 to be sorted, and the sort operations to take place in pipelining mode; see Section 7 for details.) In each step, C_i will be small enough to keep in memory as it is the result of an aggregation query. Hence, no page accesses are required for storing or retrieving C_i. Therefore, the total number of page accesses is bounded by:

$$n\,\|R_1\| + 4 \sum_{i=2}^{n-1} \|R_i\|$$

tx_id	item$_1$	item$_2$	item$_3$
10	A	B	C
30	A	B	C
20	A	B	D
40	B	C	D
80	D	E	F
90	D	E	F
99	D	E	F

item$_1$	item$_2$	item$_3$	cnt
D	E	F	3

tx_id	item$_1$	item$_2$	item$_3$
80	D	E	F
90	D	E	F
99	D	E	F

Fig. 3. Relation R'_3, C_3 and R_3.

```
k := 1;
sort R_1 on item;
C_1 := generate counts from R_1;
repeat
        k := k + 1;
        sort R_{k-1} on trans_id, item_1, ..., item_{k-1};
        R'_k := merge-scan R_{k-1}, R_1;
        sort R'_k on item_1, ..., item_k;
        C_k := generate counts from R'_k;
        R_k := filter R'_k to retain supported patterns;
until R_k = {}
```

Fig. 4. Outline of algorithm **SETM**.

Let us calculate the time to generate C_2 as we did for the nested-loops strategy. Let R_3 be empty. Using the same numbers as in Section 3.2, the cardinality of R_i is given by $\binom{10}{i} \times 200{,}000$. The size of a tuple from R_i is $(i+1) \times 4$ bytes. This gives us the following: $\|R_1\| = 4{,}000$ and $\|R_2\| = 27{,}000$. The number of page accesses is thus:

$$3 \times 4{,}000 + 4 \times 27{,}000 = 120{,}000$$

Reading and writing all the R_i relations can be done in a sequential fashion. We estimate the time for each page access as 10 ms. Hence, the total time spent on I/O operations is 1200 seconds or 10 minutes. In comparison, the nested-loop strategy required more than 11 hours.

This rough analysis shows that the implementation based on sort-merge joins will be much more efficient than the algorithm based on using nested-loop join with indexes. We will therefore proceed with further experimental evaluation of the algorithm based on sort-merge joins.

4.4. Algorithm SETM

The sort-merge strategy is described in pseudocode in Fig. 4. We refer to it as Algorithm **SETM**. The algorithm consists of a single loop, in which two sort operations and one merge-scan are performed. The first sort is needed to implement the merge-scan join that follows it. The second sort is used in order to generate the support counts efficiently. Generating the counts involves a simple sequential scan over R_k. Deleting the tuples from R_k that do not meet the minimum support, involves simple table look-ups on relation C_k. The C_k relations are of interest to us for rule generation. Note that we have not included in this algorithm the optimizations mentioned in Section 4.3.

5. Rule generation

We have omitted so far any discussion of how the rules are generated from the count relations. The rule generation algorithm is straightforward. For any pattern of length k, we consider all possible combinations of $k-1$ items in the antecedent. The remaining item not

used in the combinations is in the consequent. For each combination of antecedent and consequent, we check if the confidence factor meets or exceeds the minimum confidence factor desired. If the confidence factor is high enough, the rule is written to output. In order to check the confidence factor, we need the count for the current pattern (available in the current count relation C_i) and the count for the pattern comprising the antecedent (available by lookup in a previous count relation C_{i-1}).

5.1. Examples

Let us consider the example from Section 4.4. The minimum support is 30% (3 transactions) and the minimum confidence factor is 70%. After relation C_2 is obtained, the rules obtained are shown below. Rules have been written in the form $X \Rightarrow I$, $[c, s]$, where X is the list of items in the antecedent of the rule, I is the item in the consequent of the rule, s is the support expressed as a percentage and c is the confidence factor. Let us see how we obtain the rule $B \Rightarrow A$. The pattern AB is supported since its support is 3 and the minimum support desired is 3. The ratio $|AB|/|B| = 3/4 = 75\%$ which is greater than the minimum confidence factor of 70%. The ratio $|AB|/|A| = 3/6 = 50\%$ which is less than the minimum confidence factor of 70%. Hence, we do not obtain the rule $A \Rightarrow B$.

```
B ==> A, [75.0%, 30.0%]
C ==> A, [75.0%, 30.0%]
B ==> C, [75.0%, 30.0%]
C ==> B, [75.0%, 30.0%]
E ==> D, [75.0%, 30.0%]
F ==> D, [100.0%, 30.0%]
E ==> F, [75.0%, 30.0%]
F ==> E, [100.0%, 30.0%]
```

After the second iteration, relation C_3 is available. The rules generated from C_3 are:

```
D E ==> F, [30.0%, 100.00%]
D F ==> E, [30.0%, 100.00%]
E F ==> D, [30.0%, 100.00%]
```

In [4], a data set was used that consists of sales data obtained from a large retailing company with a total of 46,873 customer transactions. At the level of granularity in the original data set, each transaction contains the department numbers from which a customer bought an item in a visit. Hence, the algorithm finds if there is an association between departments in the customer purchasing behavior. There are a total of 63 departments. The rules obtained for this data set for a minimum support of 1% and a minimum confidence factor of 50% are given below (these rules also appeared in [4]).

```
(Home Laundry Appliances) ==> (Maintenance Agreement Sales), [66.6%, 1.25%]
(Auto Accessories) ==> (Automotive Services), [79.5%, 11.81%]
(Automotive Services) ==> (Auto Accessories), [71.6%, 11.81%]
(Children's Hardlines) ==> (Infants and Children's wear), [66.2%, 4.24%]
(Men's Furnishing) ==> (Men's Sportswear), [54.9%, 5.21%]
(Tires) ==> (Automotive Services), [98.8%, 5.79%]
(Auto Accessories) (Tires) ==> (Automotive Services), [98.3%, 1.47%]
```

5.2. *Multiple items in consequent*

We may now observe, that the generation of rules from the count relations that are generated by program **SETM** allows a very easy generalization of the kinds of rules that can be generated. For patterns of size k, it is easy to consider combinations of $k-2$ items in the antecedent with the remaining two items in the consequent. This allows us to generate rules with two items in the consequent. Similarly, by considering combinations of $k-q$ items in the antecedent, we obtain rules with q items in the consequent. Here are some of the rules we obtained for the retailing data set for a minimum support of 0.01% and a confidence factor of 80%.

```
(Junior Apparel) (Men's Sportswear) (Tires) ==>
        (Women's Sportswear) (Automotive Services), [80.0%, 0.01%]
(Junior Apparel) (Men's Sportswear) (Tires) ==>
        (Auto Accessories) (Automotive Services), [80.0%, 0.01%]
(Women's Sportswear) (Men's Furnishing) (Tires) ==>
        (Men's Sportswear) (Automotive Services), [80.0%, 0.01%]
(Boy's Clothing) (Children's Hardlines) (Automotive Services) ==>
        (Auto Accessories) (Infants and Children's wear), [80.0%, 0.01%]
(Boy's Clothing) (Girl's Clothing) (Tires) ==>
        (Infants and Children's wear) (Automotive Services),
        [100.0%, 0.01%]
```

The generation of rules with multiple items in the consequent was facilitated by the set-oriented nature of Algorithm **SETM**.

5.3. *Filtering of rules*

We observe that some of the rules generated may not carry very valuable information. For example, consider the following rules obtained for the retailing data set with a minimum support of 1%.

```
(Tires) ==> (Automotive Services), [98.8%, 5.79%]
(Auto Accessories) (Tires) ==> (Automotive Services), [98.3%, 1.47%]
```

The second rule has the same consequent as the first rule and is more specific since it has an

additional item (`Auto Accessories`) in the antecedent. However, the confidence factor for the second rule is smaller than that for the first rule. Thus, the inclusion of a new item has not increased our confidence. In many cases, this might be good reason to filter out the second rule before it is output. Such filtering is straightforward to implement and can often reduce substantially the number of rules to be considered. For example, when filtering was used in the retailing data set for a minimum support of 0.1% and a confidence factor of 50%, the number of rules without filtering is 52. When filtering is used, the number of rules drops to 18!

6. Experiments

In previous sections we have described the new algorithm and given some analysis to show that we expect it to be efficient. We implemented the algorithm to run in main memory and read a file of transactions. The execution times given are for running the algorithm on the IBM Risc/System 6000 350 with a clock speed of 41.1 MHz. The experiments were conducted using the retailing data set described in Section 5.

6.1. Variation of relation sizes

We first study how the size of the R_i (trans_id and items) relation varies with each iteration of algorithm **SETM**. In Fig. 5 we show the variation in the size (in Kbytes) of R_i with iteration i for the retailing data set. Curves are shown for different values of minimum support, where minimum support is varied from 0.1% to 5%. The maximum size of the rules is 3, hence in all cases $|R_4| = 0$ (with $|R_i|$ denoting the cardinality of R_i). Also, the starting relations are the same and hence $|R_1| = 115{,}568$ in all cases.

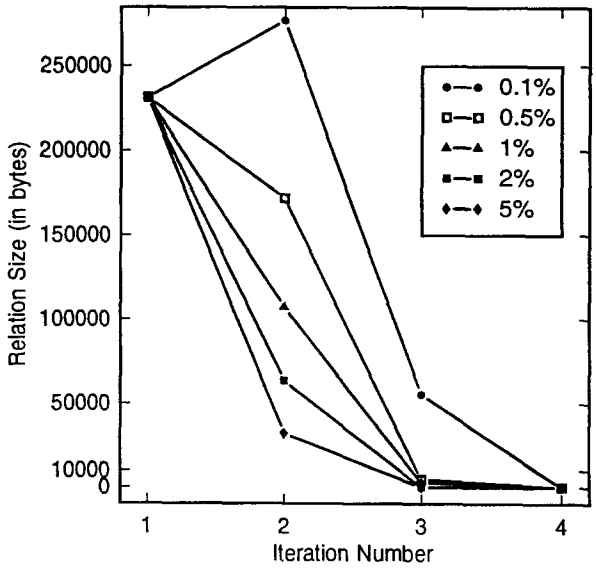

Fig. 5. Size of relation R_i.

If the minimum support is small enough ($\leq 0.1\%$), the size of relation R_i can first increase and then decrease. But the general trend is that the size relation R_i decreases. For large values of minimum support, $|R_i|$ decreases quite rapidly from the first iteration to the second. This sharp decrease is delayed somewhat for the smaller values of minimum support. Hence, using small values of minimum support allows us to obtain more rules. In general, it also allows us to obtain rules with more items in the antecedent. For example, if the minimum support is reduced to 0.05%, we obtain rules with 3 items in the antecedent.

We expect the C_i (count) relations to be small enough to fit in memory. We now study how the cardinality ($|C_i|$) of these relations varies with iteration number. Fig. 6 shows curves for different values of minimum support. The values of $|C_i|$ measure the number of item combinations that could garner enough support. We observe that for small values of minimum support the value of $|C_i|$ increases initially before decreasing with later iterations. Since $|C_i|$ is a measure of how many rules can possibly be generated, we again see the importance of handling small values of minimum support in a timely fashion. The maximum size of the rules is 3, hence in all cases $|C_4| = 0$. Also, the starting relations are the same and hence $|C_1| = 59$ for all minimum support values.

6.2. Execution times

We now measure the execution times of our set-oriented algorithm **SETM** for various values of the minimum support. As we saw in Section 6.1, the interesting values of minimum support are the small values. For large values of minimum support (greater than, say, 5%) very few rules are obtained, and the rules tend to have small antecedents and be uninteresting. Hence, we vary the minimum support from 0.1% to 5%. The execution times are shown in Table 1.

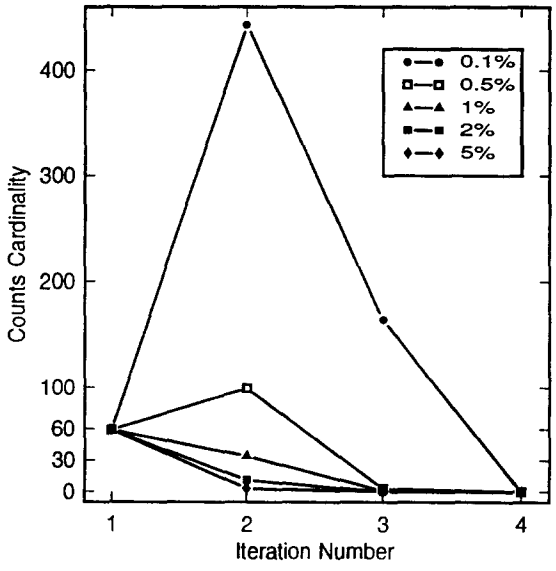

Fig. 6. Cardinality of C_i.

Table 1

Minimum support (%)	Execution time (seconds)
0.1	6.90
0.5	5.30
1	4.64
2	4.22
5	3.97

We see that algorithm **SETM** is very stable. The execution time varies from ≈7 secs for 0.1% minimum support to ≈4 secs for 5% minimum support. By contrast, the algorithm described in [4] had a running time of 2 minutes 53 seconds for a minimum support of 0.1%.

6.3. *Scaling with transaction file size*

In order to see how Algorithm **SETM** scales with the transaction file size, we replicated the transaction file a number of times and ran the algorithm on the scaled files. Since all the transactions are replicated, the support counts as a fraction of number of transactions do not change. Hence, the rules obtained are identical with those obtained when the file is not replicated.

In Fig. 7, we show the effect on execution time of Algorithm **SETM** when the transaction file is replicated 1 to 5 times. We show three different curves for minimum support values of 0.1%, 0.5% and 1%. We observe that Algorithm **SETM** scales linearly with the transaction file

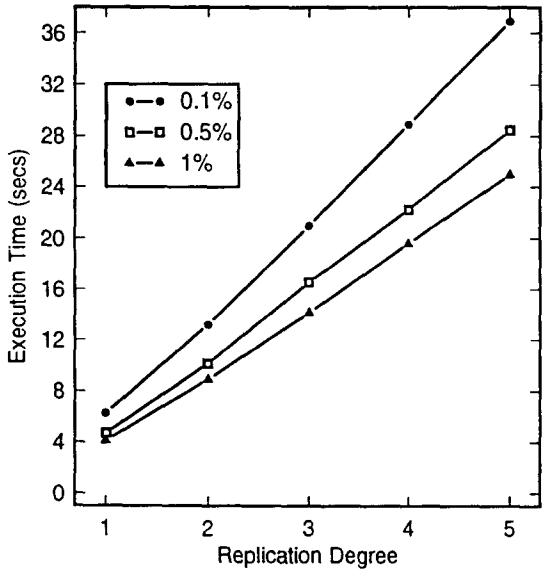

Fig. 7. Execution time of **SETM** for different replications of the transaction file.

size. This is important when the algorithm is used to mine association rules for large databases.

7. Reducing response time

We see that the running time of algorithm **SETM** (a few seconds) is good enough to facilitate interactive rule discovery. We now describe some optimizations that can be used if further reductions in response time are desired.

Pipelining between the different steps can be used to great advantage in algorithm **SETM**, especially when the steps involve access to external storage. The output of the step that applies the minimum support constraint can be piped to the step that sorts the resulting relation in (*tid*, *items*) order. This optimization saves writing and reading relation R_i once for each $i > 2$. The output of the final merge of the step that sorts on (*tid*, *items*) can then be piped to the step that joins relation R_i with the initial relation. This optimization saves writing and reading relation R_i once for each $i \geq 2$. The result of the join can be piped to the step that sorts the join result by items, saving a read and write of the join result for all the joins. The output of the final merge of the step that sorts on items can be piped to the step that applies the minimum support constraint, saving another read and write of the join result for all the joins.

An attractive feature of algorithm **SETM** is that it offers opportunities for parallelism that can be implemented easily. Clearly, all the steps that involve sorting can be parallelised using any of the existing parallel sorting techniques (see, for example, [11, 24]). Applying the minimum support constraints can easily be parallelised. The only requirement on the distribution function is that all the tuples containing a particular pattern are sent to the same processor. Both hash and range functions for distributing tuples meet this requirement. The join step can be parallelised by replicating the initial relation among the processors and fragmenting the relation R_k. We can use any distribution function including the round robin distribution function.

A straightforward scheme for parallelizing algorithm SETM is as follows. The transactions are distributed in round robin fashion amongst the processors. Each processor performs a single iteration of the algorithm on its local data in order to generate the counts obtained from the local transactions. One processor acts as a coordinator. The local count relations are sent to the coordinator. The coordinator combines the local count relations to form the global count relation and applies the minimum support constraint. The global count relation is then broadcast to all the processors. After receiving the global count relation, the processors can continue their processing and generate the local counts for the next iteration and so on, until the global count relation is empty. The coordinator stores all the global count relations, which can then be processed in the usual way to generate the rules.

We thus see that there are many opportunities for reducing the response time for algorithm **SETM** even further. The techniques used are primarily pipelining and use of parallelism. Such techniques are easily used because the algorithm **SETM** is set-oriented and can be applied orthogonally. The basic simplicity and ease of implementation of the original algorithm is still retained.

8. Conclusions

In this paper, we have investigated a set-oriented approach to mining association rules. The algorithm is straightforward – basic steps are sorting and merge scan join – and could be implemented easily in a relational database system. The algorithm is easily extended to generate rules with multiple items in the consequent.

Not only is the algorithm simple, its performance is remarkable. It exhibits very stable behavior, with execution time almost insensitive to the chosen minimum support. For a real-life data set, execution times are on the order of 4–7 seconds. We further indicated several optimizations for disk-based environments. We have shown how the algorithm can be parallelized easily and outlined such a parallel implementation. The simple and clean form of our algorithm makes it easily extensible and facilitates integration into a (interactive) data mining system.

It may be noted that the approach we have taken is similar to work on set-oriented processing of recursive queries (see e.g. [7]), in particular work on transitive closures. Indeed, we have expressed our algorithms in recursive SQL as supported by the Starburst system [14].

We are currently investigating additional uses of algorithm **SETM**. For example, it is possible to use algorithm **SETM** not only for finding association rules, but also for classification purposes. Given a relation *CUSTOMER*(*cid*, *age*, *income*, *elevel*) for which the attributes can be discretized, we may 'denormalize' the relation to get tuples of the form (*cid*, *age-group-encoding*), (*cid*, *income-group-encoding*) and (*cid*, *elevel-group-encoding*) and execute algorithm **SETM** on the set of tuples thus obtained. This will result in a meaningful classification: if the support for a pattern is large enough, we have identified a relevant class. Finally, the next plausible step is to refine rule generation by generating rules per class of customer.

Acknowledgements

We thank Rakesh Agrawal and Bill Cody for their comments on the paper.

References

[1] R. Agrawal, C. Faloutsos and A. Swami, Efficient similarity search in sequence databases, in: *Proc. Fourth Int. Conf. on Foundations of Data Organization and Algorithms* (Springer-Verlag, Berlin, Oct. 1993) 69–84. *Lecture Notes in Computer Science*, V303.

[2] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer and A. Swami, An interval classifier for database mining applications, in: *Proc. Eighteenth Int. Conf. on Very Large Data Bases*, Vancouver (August 1992) 560–573.

[3] R. Agrawal, T. Imielinski and A. Swami, Database mining: A performance perspective, *IEEE Trans. Knowledge and Data Engineering* 5(6) (Dec. 1993) 914–925. Special issue on Learning and Discovery in Knowledge-Based Databases.

[4] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, in: *Proc. ACM-SIGMOD Int. Conf. on Management of Data*, Washington, DC (June 1993) 207–216.

[5] T.M. Anwar, H.W. Beck and S.B. Navathe, Knowledge mining by imprecise querying: A classification-based approach, in: *IEEE 8th Int. Conf. on Data Engineering*, Phoenix, Arizona (1992).
[6] D. Shepard Associates, ed., *The New Direct Marketing* (Business One Irwin, Homewood, IL, 1990).
[7] F. Cacace, S. Ceri and M.A.W. Houtsma, A survey of parallel execution strategies for transitive closure and logic programs, *Distributed and Parallel Databases* 1(3) (Oct. 1993) 337–382.
[8] J. Catlett, Megainduction: A test flight, in: *8th Int. Conf. on Machine Learning* (Morgan Kaufman, June 1991).
[9] P. Cheeseman, Autoclass: A bayesian classification system, in: *5th Int. Conf. on Machine Learning* (Morgan Kaufman, June 1988).
[10] G. Cooper and E. Herskovits, A bayesian method for the induction of probabilistic networks from data, in: *Machine Learning* (1992).
[11] D.J. DeWitt, J.F. Naughton and D.A. Schneider, Parallel sorting on a shared-nothing architecture using probabilistic splitting, in: *Proc. 1st Int. Conf. on Parallel and Distributed Information Systems* (Dec. 1991) 280–291.
[12] D.H. Fischer, Knowledge acquisition via incremental conceptual clustering, in: *Machine Learning* (1987).
[13] J. Han, Y. Cai and N. Cercone, Knowledge discovery in databases: An attribute-oriented approach, in: *Proc. Eighteenth Int. Conf. on Very Large Data Bases*, Vancouver (Aug. 1992) 547–559.
[14] M. Houtsma and A. Swami, Set-oriented mining for association rules, Technical report, IBM Research Division, Oct. 1993, IBM Research Report RJ 9567.
[15] R. Krishnamurthy and T. Imielinski, Practitioner problems in need of database research: Research directions in knowledge discovery, *ACM-SIGMOD Record* 20(3) (Sep. 1991) 76–78.
[16] P. Langley, H. Simon, G. Bradshaw and J. Zytkow, eds., *Scientific Discovery: Computational Explorations of the Creative Process* (MIT Press, 1987).
[17] D.J. Lubinsky, Discovery from databases: A review of AI and statistical techniques, in: *IJCAI-89 Workshop on Knowledge Discovery in Databases* (1989) 204–218.
[18] R.S. Michalski, L. Kerschberg, K.A. Kaufman and J.S. Ribeiro, Mining for knowledge in databases: The INLEN architecture, initial implementation, and first results, *J. Intelligent Information Systems* 1 (1992) 85–113.
[19] J. Pearl, ed., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufman, 1992).
[20] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules, in: *Knowledge Discovery in Databases* (AAAI/MIT Press, 1991) 229–248.
[21] G. Piatetsky-Shapiro, ed., *Knowledge Discovery in Datbases* (AAAI/MIT Press, 1991).
[22] J.R. Quinlan, Induction of decision trees, *Machine Learning* 1 (1986) 81–106.
[23] S. Tsur, Data dredging, *IEEE Database Engineering Bull* 13(4) (Dec. 1990) 58–63.
[24] H.C. Young and A. Swami, A family of round-robin partitioned parallel external sort algorithms, Technical report, IBM Research Division , Nov. 1992, RJ 9014.

Maurice A.W. Houtsma is a senior member scientific staff at the Telematics Research Centre, where he is responsible for coordination of research into multimedia, computer supported co-operative work, heterogeneous systems, and public information systems. From 1992 to 1994 he held a position as research fellow of the Royal Netherlands Academy of Arts and Sciences, at the University of Twente. In that position he studied topics like distributed and parallel databases, replicated data, data mining, and geographical information systems. He spent several months as visiting researcher at Stanford University and at IBM Almaden Research Center, to co-operate with other researchers on those topics. His Ph.D. Thesis appeared in 1989 at the University of Twente, on data model and query processing in data and knowledge base management systems.

Arun Swami received his B.Tech. degree in Computer Science and Engineering from the Indian Institute of Technology, Bombay (1983), his M.S. degree in Computer Science (1985) and his Ph.D. degree in Computer Science from Stanford University (1989). From June 1989 to July 1994, he was a Research Staff Member at the IBM Almaden Research Center. He is currently at Silicon Graphics Computer Systems. Arun's technical interests include query optimization, random sampling, and parallel processing in database systems. He is now doing research and development in the emerging area of database mining. He has published extensively in these different areas of research and has patents and patent applications for work in these areas.