Generation of High-confidence and Large Association Rules using Galois Lattice

Fathi. ELLOUMI[(1)],

and Mohamed M. GAMMOUDI[(2)]

1) Insititut Supérieur de Gestion, Bouchoucha Le Bardo- 2000 Tunis, Tunisia.

E-mail: fethi.elloumi@fst.rnu.tn

2) University of Tunis, Department of Computer Science, Campus Universitaire, Le Belvédère, 1060, Tunis, Tunisia Phone/Fax: (+216)1 716691

ABSTRACT

In this paper we introduce a method for generation of high-confidence association rules from a large transaction's database. We represent this database by a binary relation where the domain is the set of transactions and the range is the set of items. From this binary relation we construct a Galois Lattice. We show how the hierarchical structure of Galois Lattice allows generation of high-confidence association rules which associate the maximum of items.

KEYWORDS : High-Confidence Association Rule, Galois Lattice, Maximal Rectangle.

1 INTRODUCTION

The problem of discovering association rules between items in a large database sales transaction of a supermarket was introduced in [1]. These rules are interesting because they allow the head of the supermarket to take the good decisions such for example what to put on sale and help to develop a marketing program.

In this paper we are interested with rules that associate the maximum of items, called large rules, and have high confidence values close to 1. Such rules are useful and can be used to organize the supermarket and place better the items in the shelves in order to maximize the profit. We develop an algorithm that uses the hierarchical structure of Galois lattice, the lattice of large and maximal itemsets to generate this kind of rules.

This paper is organized as follows. In section 2, we review the related work. Section 3 presents the mathematical background of the rectangle concept and Galois lattice. Section 4 shows how we generate large and maximal itemsets. The algorithm for generation of high confidence and large rules is introduced in section 5. Finally, section 6 concludes this paper and points out future issues.

2 RELATED WORK

The problem of discovering association rules can be decomposed into two steps [1] :

Step1: Find all sets of items (itemsets) that have transaction support greater than minimum support. The support for an itemset is the ratio of the number of transactions that contains the itemset to the total number of transactions. These itemsets are called frequent itemsets.

Step2 : For each frequent itemset E, find its all not empty subsets S and output rules of the form

S => E - S if the ratio of support(E) to support(S) is greater than a threshold called minimum confidence.

Step 1 is more difficult than step2 and many algorithms were proposed to find frequent

itemsets. APRIORI [2] is a well-known one . Its basic idea is that if an itemset is frequent so are all its sub-itemsets. Therefore the construction of sets of n items candidates to be frequent is generated by joining frequent sets of n-1 items and pruning those whose subsets of n-1 items are not frequents.

Other algorithms [6], [14], [9] have followed the same idea and tried to outperform Apriori-algorithm in reducing the time of execution, the number of passes made over the database and the memory consumption.

The apriori-like algorithms are inadequate for finding frequent itemsets with high number of items (large itemsets) because they produce an exponential number of candidates itemsets [3]. To find a frequent itemset of n items, Apriori produces 2^n frequent sub-itemsets. This exponential complexity restricts apriori-like algorithms to find only frequent itemsets with low number of items. To address this problem, many recent algorithms have been proposed such as of Bayardo [3] who find the maximal frequent itemsets or the close algorithm of [13] which find the frequent closed itemsets. These algorithms produce efficiently large frequent itemsets.

In our work we use the notion of maximal itemsets through the concept of maximal rectangles [8]. Many algorithms [5] ,[11] enable to produce the lattice of maximal rectangles. We insist on exploitation of the lattice structure to show how it serve to generate high confidence and large association rules.

3. RECTANGLE AND GALOIS LATTICE

A binary relation R on set E is defined as a subset of the Cartesian product E x E. We denote by e R e' the fact that an argument e (or input) of E is linked with an image e' (or output) of E by R. Among the relations on a set E, we can mention the *identity relation* **I** and the *null relation* **Ø** (we shall use this symbol to denote the null set Ø). If S is a subset of E, then we denote by **I**(S) = {(e,e) | e ∈ S} the image set of S by **I**. We can associate the following subsets of E and F with a given binary relation R:
the *image* set of e is defined by
e. R = {e' ∈ E | e R e'};
the *antecedents* of e' are defined by
R. e' = {e ∈ E | e R e'};
the *domain* of R is defined by
Dom(R)= {e | ∃ e' ∈ E : e R e'};
the *range* of R is defined by
Range(R) = {é | ∃ e ∈ E : e R e'}.

3.1 Rectangle

Given a set S, we let a rectangle on S be a pair of sets A and B of S, which we denote by (A,B). Except when A= ∅ or B = ∅ there is a one to one correspondence between r_1 and $r_2,r_3..r_n$. A is the domain of the rectangle (A,B) and B is its range.

Remarks:
The correspondence between rectangles (A_i,B_i) and the associated rectangular relations A_i x B_i is a bijective one, except when $A_i = \emptyset$ or $B_i = \emptyset$. For instance, the rectangles $(\emptyset, B_1)$ and $(A_1, \emptyset)$ are both associated with the null rectangular relation ∅. The main reason for making a distinction between rectangles and rectangular relations is that the concept of a rectangle enables us to obtain a lattice structure.

3.2 Maximal Rectangle

Let (A,B) be a rectangle of a given relation R defined on S. The rectangle (A,B) is said to be ***maximal*** if whenever A x B ⊆ A' x B' ⊆ R, then A = A' and B = B' [8].
Remark :
The notion of maximal rectangle is not new, it is found with different names such as : Complete couple in [10], or Concept in [15].

3.3 Partial Order Relation

The relation defined below on the set of maximal rectangles of a binary relation R is a partial order relation:
∀ (A_1, B_1) and (A_2, B_2) two maximal rectangles of R, $(A_1, B_1) \leq\leq (A_2, B_2)$ <=> $A_1 \subseteq A_2$ and $B_2 \subseteq B_1$. The proof of this proposition can be found in [7].

3.4 Galois Lattice

Let R be a finite binary relation defined on E and F there is a unique Galois Lattice

corresponding to R [11]. Each element of the lattice must be a maximal rectangle as defined in 3.2, noted (X,Y), composed of a set $X \in P(E)$ and a set $Y \in P(F)$.
Let *f* and *g* defined as following :

$f = \{ (X,Y) / \forall x \in X, \forall y \in Y (x , y) \in R\}$
$g = f^{-1} = \{(Y ,X) / (X,Y) \in f \}$

where the couple of functions (*f,g*) is said to be a Galois Connection between P(E) and P(F) and the Galois lattice (*GL*, $\leq\leq$) for the binary relation is the set of all maximal rectangles with the partial order relation defined in 3.3. The partial Order is used to generate the graph of the Galois lattice which is called Hasse Diagram. Galois lattice has a supremum (E,$\varnothing$) and an infimum ($\varnothing$,F), for more details on Lattice Galois the reader should consult [12].

4. GENERATION OF LARGE AND MAXIMAL ITEMSETS

Let T be the set of transactions and A be the set of items. We begin by transforming the transactional database to a binary relation database R such as $R \subseteq T \times A$. Then we construct he lattice of maximal rectangles (G, $\leq\leq$) where G is the set of maximal rectangles of R.
The domain of a maximal rectangle is a set of transactions ant its range is an itemset. From definition of a maximal rectangle we can assert that the range of a maximal rectangle corresponds to a large itemset (in number of items) with the maximal number of transactions that contains it.
.

Definition

The range of a maximal rectangle is said a large and maximal itemset. The support of this itemset is the ratio of the cardinal of domain of the maximal rectangle to the cardinal of domain of relation R.
Remark:
In the rest of this paper, we frequently use the terms support of maximal rectangle instead of support of large and maximal itemset which mean the same signification.

Properties

1- A large and maximal itemset is said frequent if its support is greater than minimum support.
2- Each subset of a large and maximal itemset is frequent but not necessarily large and maximal.
3- The frequent, large and maximal sub-itemsets of a frequent large and maximal itemset can be deduced directly from the lattice (G, $\leq\leq$).

Example
This example is extracted from [13] : Let be the following transaction's database :

TID (+)	LIST OF ITEMS
T1	A C D
T2	B C E
T3	A B C E
T4	B E
T5	A B C E

(+) TID: Transaction Identifier

R is the binary relation that corresponds to the previous database and is represented by the following matrice

	A	B	C	D	E
T1	1		1	1	
T2		1	1		1
T3	1	1	1		1
T4		1			1
T5	1	1	1		1

We represent below the lattice of maximal rectangle of relation R (see figure 1).

If we assume that minimum support is 2 (2/5) then we are concerned only with the following maximal rectangles:

REC0 = T1,T2,T3,T5	x	C
REC1 = T2,T3,T4,T5	x	B, E
REC2 = T1,T3,T5	x	A, C
REC3 = T2,T3,T5	x	B, C, E
REC4 = T3,T5	x	A, B, C, E

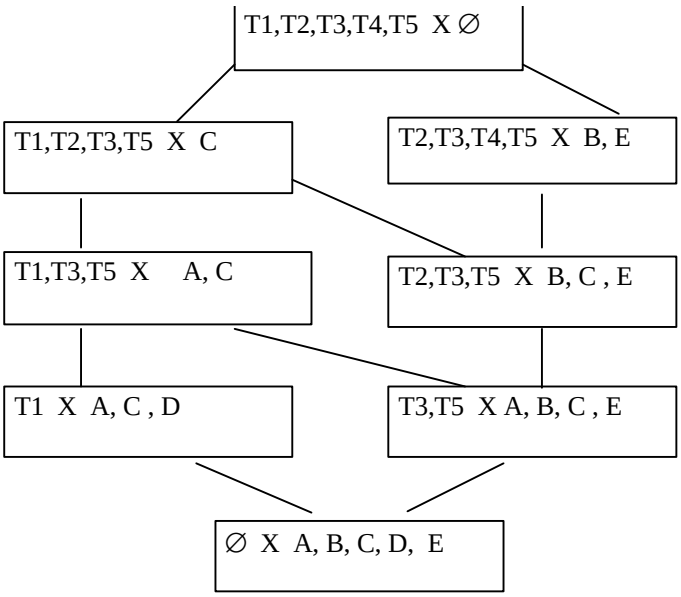

Figure1: lattice of maximal rectangles of R

From these rectangles we can deduce that the itemsets { C }, { B, E }, { A, C }, { B, C, E } and { A, B, C, E} are frequent, large and maximal. Each subset of { A, B, C, E} is frequent. Using the lattice we can see that REC4 ≤≤ REC2 and REC4 ≤≤ REC3 so the range of REc2 and REC3, which are { A, C }, { B, C, E }, are subset of { A, B, C, E} that are frequent but also large and maximal.

A directed advantage follows from the lattice structure is that we have an idea of the different levels or values of supports. We can help the user to choose the minimum support contrary to done actually that consist to assume a minimum support and change it when finding bad results.
We can go ahead , using this structure, to choose the best minimum supports for generating rules that contain user-specified items by selecting the specific rectangles. We can also help the user to reformulate its query when searching such rules by using Galois connection [7].

5. ALGORITHM FOR GENERATION OF HIGH-CONFIDENCE AND LARGE RULES

We use the lattice of maximal rectangles to generate high-confidence and large rules. We set a minimum support so we are concerned with only frequent large itemsets or maximal rectangles which supports (supports of its ranges) are greater than the minimum.
In fact a maximal rectangle REC assures a large itemset that enable generation of large rules. Next by using the partial order relation ≤ ≤ we find for each maximal rectangle REC all the maximal rectangles REC_i that are immediately higher than REC (REC ≤≤ REC_i). The range of each REC_i , is a sub-itemset of the range of REC which is frequent, large and maximal, will be used as premise for our rules. The cardinal of the range of each REC_i is the nearest of the cardinal of range of REC and so guarantee confidence value high and close to 1. In our work, we don't fix a minimum confidence.
We present below the main notations that we use in our algorithm. Let be

T	the set of transactions
A	the set of items
R	a binary relation / R ⊆ T x A
G	the set of maximal rectangles of R
(G, ≤≤)	the lattice of maximal rectangles of R
REC	a maximal rectangle
card(E)	the cardinal of set E
supmin	minimum support
Sup(E)	support of set E
Sup(REC)	support of REC / Sup(REC) = Sup(range(REC))

The lattice is (G, ≤≤) is the entry of our algorithm given in the following:

Algorithm
Begin
For each rectangle REC with sup(REC) >= supmin and card(range(REC)) >= 2
do
If REC ≤≤ (T x ∅)
then
/* CASE 1
All items of range(REC) are strongly linked. We output rules with confidence = 1 */
∀ S⊆ range(REC); S ⇒ range(REC)-S with confidence =1
/* because card(S) = card(range(REC)) */
Else
/* CASE 2
we output rules with high-confidence values */
For each rectangle REC_i that is immediately higher than REC (REC ≤≤ REC_i).
Do
/* output the rules: */

range(REC_i) ⇒ range(REC) - range(REC_i)
with Confidence = sup(REC) / sup(REC_i)

/* the value of confidence is the highest because range(REC_i) is the largest maximal itemset that can be a premise */

end do

/* CASE 3
we output other rules with confidence = 1 */

For each itemset S ⊆ range(REC) and S not ⊆ every range(REC_i)

/* rectangle REC_i is immediately higher than REC (REC ≤≤ REC_i).*/
Do
We output
S ⇒ range(REC) - S with confidence =1

/* because card(S) = card(range(REC)) */
end do
End if
End do
End

Using the same example stated above (see figure 1) and If we assume that minimum support is 2 (2/5) then we are concerned only with the following maximal rectangles:

REC0 = T1,T2,T3,T5	x	C
REC1 = T2,T3,T4,T5	x	B, E
REC2 = T1,T3,T5	x	A, C
REC3 = T2,T3,T5	x	B, C, E
REC4 = T3,T5	x	A, B, C, E

The rectangle REC0 is not retained because it contains only one item.

Applying our algorithm, we give below the rules relatives to each rectangle.

1) The rules relatives to REC1

We have REC1 << T1,T2,T3,T4,T5 X ∅

Case1 (see algorithm)
B ⇒ E with confidence = 1
E ⇒ B with confidence = 1

2) The rules relatives to REC2 :

we have REC2 << REC0

Case2
C ⇒ A with confidence = 3/4
Case3
A ⇒ C with confidence = 1

3) The rules relatives to REC3

we have REC3 << REC0 and REC3 << REC1

Case 2
C ⇒ BE with confidence = 3/4
BE ⇒ C with confidence = 3/4
Case3
BC ⇒ E with confidence = 1
CE ⇒ B with confidence = 1

4) The rules relatives to REC4

we have REC4 << REC2 and REC4 << REC3

Case2
AC ⇒ BE with confidence = 2/3
BCE ⇒ A with confidence = 2/3
Case3
AB ⇒ CE with confidence = 1
AE ⇒ BC with confidence = 1
ABC ⇒ E with confidence = 1
ABE ⇒ C with confidence = 1
ACE ⇒ B with confidence = 1
So for a minimum support equal to 2/5 our algorithm generates the following rules which are sorted on confidence.

B ⇒ E	with confidence = 1
E ⇒ B	with confidence = 1
A ⇒ C	with confidence = 1
BC ⇒ E	with confidence = 1
CE ⇒ B	with confidence = 1
AB ⇒ CE	with confidence = 1
AE ⇒ BC	with confidence = 1
ABC ⇒ E	with confidence = 1
ABE ⇒ C	with confidence = 1
ACE ⇒ B	with confidence = 1
C ⇒ A	with confidence = 3/4
C ⇒ BE	with confidence = 3/4
BE ⇒ C	with confidence = 3/4
AC ⇒ BE	with confidence = 2/3
BCE ⇒ A	with confidence = 2/3

6. CONCLUSION

This paper has shown how we can exploit the lattice of maximal rectangles to generate high confidence and large association rules. We also see that the lattice structure can help the user to choose the minimum support and to search the rules which contain specific items.

Although our method is incomplete because we don't generate all association rules, it is useful and sufficient especially for commercial domain where constraints as the number of items in the rules or finding strongly linked items can be imposed.

Algorithms for construction lattice of maximal rectangles are limited for very large database. Improving and implementing such algorithms is under study.

REFERENCES

[1] Agrawal R, Imielinski T and Swami A. Mining association rules between sets of items in large databases. In Proc of the 1993 ACM-SIGMOD Int'l conf on management of data.p207-216

[2] Agrawal R and Skirant R.
Fast algorithms for mining association rules. In Proc of the 20th Int'l conf on VLDB p.478-499. Expanded version in IBM research report RJ9839. 1994

[3] Efficiently mining long patterns from databases. In Proc of the 1998 ACM-SIGMOD Int'l conf on management of data.85-93

[4] N. Belkhiter, C. Boulhfir, M.M. Gammoudi, A. Jaoua, N. Lethanh, M. Reguig,
Décomposition rectangulaire optimale d'une relation binaire : Application aux base de données documentaire, In INFOR Journal, 1994

[5] J.P. Bordat, « Sur l'algorithmique combinatoire d'ordres finis ». Thèse de Doctorat de l'université de Montpellier II, avril 1992.

[6] Brin S, Motwani R, Ullman J and Tsui S. Dynamic Itemset Counting and Implication rules for market basket data
In Proc of the 1997 ACM-SIGMOD Int'l conf on management of data.255-264.

[7] M. M. Gammoudi, « Méthode de décomposition rectangulaire d'une relation binaire : une base formelle et uniforme pour la génération automatique des thesaurus et la recherche documentaire », Thèse de doctorat Es-Sciences, spécialité informatique, UNSA-I3S, France, 1993.

[8] M. M. Gammoudi, J. Mendes and S. Wilson,
An automatic genaration of Hierarchy using the rectangular Decomposition of a Binary Relation. In Lecture Notes proceedings, ER'97, Los Angeles, USA, November 1997

[9] Gardarin G, Pucheral P and Wu F. Bitmap based Algorithms for mining association rules
In proc of the 14th journey of advanced databases. Hammamet Tunisia p157-175

[10] Garey M.R and Johnson DS. Computers and Intractability. A guide to the theory of NP-Completeness. W.H.Freeman, 1979.

[11] Godin R, Missaoui R and Alaoui H. Learning algorithms using a galois lattice structure. In proceedings of the third international ACM SIGIR Conference on tools for Artificial Intelligence, San José , Calif: IEEE computer Society Press, 1991, p22-29.

[12] A. Kaufman, E. Pichat,
Méthode mathématiques non numériques et leurs algorithmes (Tome1), Masson, Paris, 1977.

[13] Pasquier N, Bastide Y, Touil R and Lakhal L.
Pruning closed itemset lattices for association rules. In proc of the 14th journey of advanced databases. Hammamet Tunisia p177-196

[14] Savaseri A, Omiecinski E and Navathe s. An efficient algorithm for mining association rules in large databases. In Proc of the 21th Int'l conf on VLDB p.432-444. 1995.

[15] R. Wille,
Knowledge Acquisition by Methods of Formal Concept Analysis,
In E. Diday (Eds.), Data Analysis, Learning Symbolic and Numeric Knowledge, New York: Nova Science Pub., 1989, p.365-380.