AN IMPROVEMENT OF OPTIMAL COUVERTURE EXTRACTION USING SEMANTIC APPROACH

Ourida Saidi[1]

[1]Department of Computer Sciences, Higher Institute of Management, Tunisia
Ourida.saidi@isg.rnu.tn

ABSTRACT

The amount of data speedily proliferates. Consequently, the excessive number of extracted association greatly prohibits to better assist the decision maker. In this respect, backboned on the Formal Concept Analysis, we propose to extend the notion of Formal Concept through the generalization of the notion of itemset aiming to consider the itemset as an intent, its support as the cardinality of the extent. Accordingly, we propose a new approach to extract frequent itemsets using the coverage concept. In fact, this contribution is based on a quality-criterion of a rule namely the profit which expressively improves the classical formal concept analysis through the addition of semantic value to derive highly significant association rules.

KEYWORDS

Association rules, formal concept analysis, quality measure.

1. INTRODUCTION

The examination of stored data is primary to discover the unknown information and extract the hidden patterns namely the association rules. Recently, a new trend of approaches based on Formal Concept Analysis [1,2,3] appears. In fact, the "Concept" is a couple of intent and extent used to represent nuggets of knowledge. Frequently, the number of the extracted association rules grows exponentially with the number of data rows and attributes. Thus, the understanding and the explanation of the generated patterns becomes a challenging task. Thus, various attractive proposals have been advanced related to association rules [4].

In this paper, we introduce an original approach of association rules mining based on Formal Concept Analysis. The paper is organized as follows. Section 2 presents the works related to association rules extraction issue. Section 3 details the mathematical background of FCA and its connection with the derivation of association rule bases. Section 4 introduces our method of optimal itemsets extraction. Results of the experiments carried out on benchmark datasets are reported in section 5. Illustrative examples are given throughout the paper. Section 6 concludes this paper and presents our future work.

2. RELATED WORKS

The huge number of the generated thousands and even millions of rules – among which many are redundant (Bastide et al., 2000; Stumme et al., 2001; Zaki, 2004) –[5, 6, 7, 8]

DOI : 10.5121/ijcga.2011.1201

promoted the proposal of more efficient discriminating techniques in order to decrease the number of obtained rules.

This pruning technique can be backboned on patterns defined by the user (user-defined templates), on boolean operators (Meo et al., 1996; Ng et al., 1998; Ohsaki et al., 2004; Srikant et al., 1997) [9,10,11,12].

Another trend dedicated to heavily diminish the number of rules emerged. It stresses on additional information that can be efficiently used such as the taxonomy of items (Han, & Fu, 1995) or a metric of specific interest (Brin et al., 1997) (e.g., Pearson's correlation or χ2-test) [13, 14]. More sophisticated methods that produce only lossless information limited number of the entire set of rules, called generic bases (Bastide et al., 2000). The generation of such generic bases greatly employs a battery of results provided by formal concept analysis (FCA) (Ganter & Wille, 1999) [15].

Basically, the pruning strategy of association rules is based on fundamental methods namely the frequency of the generated pattern through discarding all the itemsets having a support less than MinSup, and the strength of the dependency between premise and conclusion by pruning all the rules having a confidence less than MinConf.

To efficiently prune the reported association rules, some researchers [16] provide other measures. Indeed, Bayardo et al suggest the conviction measure. Moreover, Cherfi et al [17] porpose five different measures namely the benefit (interest) and the satisfaction. Maddouri et al provide the gain measure [18].

In this paper, we introduce a new measure: the *profit*. Indeed, the latter is based on the Formal Concept Analysis [19, 20]. Assuming that an itemset is completely represented by a formal concept as a couple of intent (the classic itemset) and extent (its support), it combines the support of the rule with the length of the itemset. So, we propose to include a new semantic aspect on association rules extraction by considering the lift measure during the selection of frequent itemsets when we generate the association rules.

3. MATHEMATICAL BACKGROUND

We recall some crucial results inspired from the Galois lattice-based paradigm in FCA and its interesting applications to association rules extraction.

3.1.Preliminary notions

In the remainder of the paper, we use the theoretical framework presented in [20]. Let O be a set of objects, P a set of properties and R a binary relation defined between O and P [19, 20].

TABLE 1. FORMAL CONTEXT

I \ *O*	*A*	*B*	*C*	*D*
o1	1	1	0	0
o2	1	1	0	0
o3	0	1	1	0
o4	0	1	1	1
o5	0	0	1	1

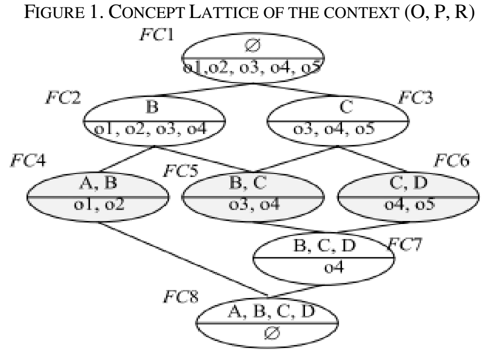
Definition 1 [19]: A formal context (O, P, R) consists of two sets O and P and a relation R between O and P. The elements of O are called the objects and the elements of P are called the properties of the context. In order to express that an object o is in a relation R with a property p, we write oRp or (o, p)∈R and read it as "the object o has the property p".

Definition 2 [19]: For a set A⊆O of objects and a set B⊆P of properties, we define :
The set of properties common to the objects in A :
A►={p∈P | oRp for all o∈A}
The set of objects which have all properties in B :
B◄={o∈O | oRp for all p∈B}
The couple of operators (►, ◄) is a Galois Connection.

Definition 3 [19]: A formal concept of the context (O, P, R) is a pair (A, B) with A⊆O, B⊆P, A►=B and B◄=A.
We call A the extent and B the intent of the concept (A, B).

Definition 4 [19]: The set of all concepts of the context (O, P, R) is denoted by Φ (O, P, R). An ordering relation (<<) is easily defined on this set of concepts by :
(A1, B1) << (A2, B2) ⇔ A1⊆A2 ⇔ B2⊆B1.

FIGURE 1. CONCEPT LATTICE OF THE CONTEXT (O, P, R)



In this subsection, we remind basic theorem for Concept Lattices [19]:

Φ (O, P, R, <<) is a complete lattice. It is called the *concept lattice* or *Galois lattice* of (O, P, R), for which infimum and supremum can be described as follow:
$\mathrm{Sup}_{i\in I}(A_i,B_i)=((\cup_{i\in I}A_i)^{\blacktriangleright\blacktriangleleft}, (\cap_{i\in I}B_i))$
$\mathrm{Inf}_{i\in I}(A_i, B_i)=(\cap_{i\in I}A_i , (\cup_{i\in I}B_i)^{\blacktriangleleft\blacktriangleright})$

Example [18]: table 1 illustrates the notion of formal context (O, P, R).The latter is composed of five objects {o1, o2, o3, o4, o5} and four properties {A, B, C, D}. The concept lattice of this context is drawn in Figure 1 containing eight formal concepts.

Definition 5 [19]: Let (o, p) be a couple in the context (O, P, R). The pseudo-concept PC containing the couple (o, p) is the union of all the formal concepts containing (o,p).

Definition 6 [20]: A coverage of a context (O, P, R) is defined as a set of formal concepts CV={RE_1, RE_2, ..., RE_n} in Φ (O, P, R), such that any couple (o, p) in the context (O, P, R) is included in at least one concept of CV.

FIGURE 2. ILLUSTRATIVE EXAMPLE OF PSEUDO-CONCEPT, OPTIMAL CONCEPT, AND NON OPTIMAL CONCEPT CONTAINING THE COUPLE (O3,B).

O / *I*	*A*	*B*	*C*
o1	1	1	0
o2	1	1	0
o3	0	1	1

a. Pseudo-concept of (o3,B)

O / *I*	*A*	*B*	*C*
o1	1	1	0
o2	1	1	0
o3	0	1	1

b. Optimal concept of (o3,B)

O / *I*	*A*	*B*	*C*
o1	1	1	0
o2	1	1	0
o3	0	1	1

c. Non optimal concept of (o3,B)

Example [18]:

Considering the formal context (O, P, R) depicted by table 1, the figure 2.a represents the pseudo-concept containing the couple (o3, B) being the union of the concepts FC_2 and FC_5.

A coverage of the context is formed by the three concepts: {FC_4, FC_5, FC_6} such as:

- FC_4 is the concept containing the items ({o1, o2}, {A, B});
- FC_5 is the concept containing the items ({o3, o4}, {B, C});
- FC_6 is the concept containing the items ({o4, o5}, {C, D}).

The lattice constitutes concept coverage.

4. EXTRACTION OF OPTIMAL ITEMSETS

It is recognized that the expensive step to derive association rules is the computation of the frequent itemsets [4]. In fact, this phase consists of applying, iteratively, an heuristic to compute the candidate itemsets. At the iteration *i*, we combine the itemsets of the iteration *i-1*. After that, the support threshold (MinSup) is used to prune non-frequent candidates. The itemsets of iteration *i-1*, are also discarded. We keep the remaining itemsets of the latest iteration n with n is the number of properties in the formal context.

Two characteristics are defined during the derivation of association rules: *(i) the support* which is the cardinality of the set of objects which verify the rule. In Formal Concept Analysis, it refers to the extent of a formal concept; *(ii) the cardinality of the Itemset* which is the number of properties of the itemset. In Formal Concept Analysis, it refers to the intent of a formal concept.

The intent is not sufficiently adequate to represent the association rule, the latter is entirely correlated to a formal concept namely both its intent and its extent. Having a support represented by the cardinality of the extent, a highly qualified selection of itemsets must be done according to the intent of the formal concept of the rule.

Besides, the association rule generated from the formal concept should efficiently consider the quality of the relationship between the head and the body of the rule.

To formalize the new criterions, we give the following definitions.

Definition 7 : Let $FC_i = (A_i , B_i)$ be a formal concept. We define:

- Length of a concept FC_i: the number of properties in the intent B_i of the concept.
- Width of a concept FC_i: the number of objects in the extent A_i of the concept.
- Lift of a concept FC_i: the maximum lift of the set of rules generated from the concept FC_i.
- Profit of a concept: is a function of the width the length and the lift of the concept, given by:

$$Profit(FC_i)=\frac{length(FCi)*width(FCi)}{length(FCi)+width(FCi)}*lift(FCi)$$

The profit measure depends on the number of properties. In fact, a less number of properties and a less number of objects, a less value of profit is noted. Having more properties and more objects induces to higher profit. Moreover, if we increase the value of a lift of the concept a higher value of profit is noted.

Definition 8 : A formal concept $FC_i = (A_i,B_i)$ containing a couple (o, p) is said to be optimal if it maximizes the *profit* function.

Definition 9 [20]: A coverage $CV=\{ FC_1 , FC_2, ..., FC_k \}$ of a context (O, P, R) is optimal if it is composed by optimal concepts.

Example [18]: An illustrative example of the pseudo-concept is sketched by figure 2. b represents the optimal concept FC_5 containing the couple (o3, B). Figure 2.c represents the non optimal concept FC_2 containing the couple (o3, B).

The optimal coverage of the context (O, P, R) is formed by three optimal concepts: $\{FC_4, FC_5 , FC_6\}$. FC_4 is the concept containing the items ({o1, o2}, {A, B}). FC_5 is the concept containing the items ({o3, o4}, {B, C}). FC_6 is the concept containing the items ({o4, o5}, {C, D}).

4.1. Heuristic Searching for Optimal Concept

The pseudo-concept, denoted by PCF, containing the couple (o, p), is the union of all the concepts containing (o, p). It is computed according to the relation R by the set of objects described by p, then {p}►, and the set of properties describing the object o, so {o}◄. Where (►,◄) is the Galois connection of the context (O, P, R).

When we determinate the pseudo-concept PCF, two cases are considered:

- *Case 1*: PCF forms a formal concept.

If no zero is found in the relation/matrix representing PCF, then, PCF is the optimal concept. So, the algorithm stops.

- *Case 2*: PCF is not a formal concept.

If some zero entries are found in the relation/matrix representing PCF, we will look for more restraint pseudo-concepts within the pseudo-concept PCF.
So, we consider the pseudo-concepts containing the couples like (X, p) or (o, Y). These concepts contain, certainly, the couple (o, p).

The considered heuristic is the optimal concept undoubtedly included in the optimal pseudo-concept.

We should generate all possible rules from the pseudo-concepts containing the couples like (X, p) or (o, Y). Then we compute the corresponding lifts and choose the maximum value between them. After that, we calculate the profit value. Finally, we retain the pseudo-concept having the greatest value of the profit function to be the new PCF.

This heuristic procedure (of case 2) is repeated until PCF becomes a formal concept. To calculate the profit of a pseudo-concept, we introduce the general form of the previous function:

Definition 10: Let $PCF_i = (A_i , B_i , R_i)$ be a pseudo-concept, where R_i is the restriction of the binary relation R, to the subsets A_i and B_i . We define the:

- Length of a pseudo-concept PCF_i : the number of properties in B_i.
- Width of a pseudo-concept PCF_i : the number of objects in A_i .
- Lift of a pseudo-concept PCF_i : is the maximum of lift found when we generate the set of rules extracted from the pseudo-concept PCF_i .
- Size of a pseudo-concept PCF_i: the number of couples (of values equal to 1) in the pseudo-concept. When PCF_i is a formal concept, we have:

$$Size(PCF)= (length(PCF)* width(PCF))$$

- Profit of a pseudo-concept is a function of the width, the length, the size and the lift given by :

$$Profit(PCF)= \left[\frac{Size(PCF)}{length(PCF)+ lift(PCF)}\right] * \left[(length(PCF)+ width(PCF))- Size(PCF)\right]$$

4.2. Method for Optimal Coverage derivation

The problem of covering a binary relation by a set of optimal concepts can eventually be resolved through covering a binary matrix by a number of its complete sub-matrix. The latter is a matrix having all its entries equal to '1'. This issue, being NP-Complete problem, has been the subject of several previous works.

It's obvious to propose an approximate polynomial algorithm called *enhanced Semantic Approach based on Formal Concept Analysis and Lift Measure* denoted *SAFCALM*.

Let R be the binary relation to cover. The proposed solution is to split R into n packages (subsets): $P_1, ..., P_n$. Each package symbolizes one or more couples.

The strategic idea of *SAFCALM* algorithm is to build incrementally the optimal coverage of R:

(i) The first phase, covering the relation $R_1 = P_1$ by CV_1.
(ii) The i^{th} phase, let $R_{i-1} = P_1 \cup \ldots \cup P_{i-1}$ and let CV_{i-1} be its optimal coverage. Building the optimal coverage CV_i of $R_i = R_{i-1} \cup P_i$ using the initial coverage CV_{i-1} and the package P_i.
(iii) The n^{th} phase, finally, finding a set of concepts covering the relation R.

Algorithm *SAFCALM*
Begin
Let R be partitioned to n packages $P_1, \ldots, P_n$.
Let $CV_0 := \emptyset$.
FOR *$i=1$* ***to*** *n* ***DO***
Sort the couples of P_i by the pertinence of their pseudo-concepts
While *$(P_i \neq \emptyset)$* ***Do***
- *Select a couple (a, b) in P_i by the sorted order of the profit function*
- *Search PC : the pseudo-concept containing (a, b) within $R_i = CV_{i-1} \cup P_i$*
- *Search FC: the optimal concept containing (a,b) within PC*

$CV_i := (CV_{i-1} - \{r \in CV_{i-1} / r \subseteq FC\}) \cup \{FC\}$:
Delete all the redundant concepts from CV_i
$P_i := P_i - \{(X,Y) \in P_i / (X,Y) \in FC\}$
End While
End FOR
End.

FIGURE 3. INCREMENTATION PHASE WHEN ADDING P4

P \ O	A	B	C	D
o1	1	1	0	0
o2	1	1	0	0
o3	0	1	1	0

a. Optimal coverage of the context ({o1,o2,o3}{A,B,C,D})

P \ O	A	B	C	D
o1	1	1	0	0
o2	1	1	0	0
o3	0	1	1	0
o4	0	1	1	1

b. Case 1 : Coverage of the context ({o1,o2,o3,o4}{A,B,C,D})

P \ O	A	B	C	D
o1	1	1	0	0
o2	1	1	0	0
o3	0	1	1	0
o4	0	1	1	1

c. Case 2 : Coverage of the context ({o1,o2,o3,o4}{A,B,C,D})

Example: Let R be the relation to cover as highlighted by table 1. R is partitioned into five packages:

- P_1={o1}x{A, B},
- P_2={o2}x{A, B},
- P_3={o3}x {B, C},
- P_4={o4}x{B, C, D} and
- P_5={o5}x{C, D}.

Firstly, R is an empty relation and in each phase we add a package.

Figure 3 presents the incrementation phase when adding P_4. In this phase, R3 encloses the four rows P_1, ..., P_3. The initial optimal coverage CV_3 encloses the formal concepts FC_3=({o1, o2}, {A, B}) and FC_4=({o3}, {B, C}).

The package P_4 encloses only three couples: (o3,B), (o3,C) and (o3,D).
The pseudo concept containing the couple (o4, B) and (o4,C) is :

Case 1: the union of formal concepts ({o3,o4}, {B, C}) and ({o4},D)

Case 2: the formal concept ({o4}, {B, C, D}).

The computation of the profit function is equal to zero in the first case and higher positive value in the second case according to the following formula:

$$Profit(PCF)= \left[\frac{Size(PCF)}{length(PCF)+ lift(PCF)}\right] * \left[(length(PCF)+ width(PCF)- Size(PCF)\right]$$

The retained formal concept is FC_6=({o4}, {B, C, D}).

The final coverage of R contains the concepts FC_3, FC_4 and FC_6. Finally, according to our example, we find three itemsets : {A, B}, {B, C} and {C, D}.

5. Experimental study

In this section, we present our carried out experimental study to stress the potential benefits of our proposed approach.

Experiments were conducted on a Pentium IV PC with a CPU clock rate of 3.06 Ghz and a main memory of 512 MB. The characteristics of benchmark datasets used during these experiments are depicted in table 2.

TABLE 2. BENCHMARK DATASET CHARACTERISTICS

Dataset	*# Transactions*	*#Items*
T10I4D	1000000	100
Mushroom	8124	128
T20I6D	1000	9
Tic Tac Toe	958	10

To highlight the performance of our approach, we compare our proposal to the three pionnering methods in the same trend in the litterature namely the Apriori algorithm and IAR approach. The latter is already based on FCA background.

To analyze the data, we choose the following values of parameters: MinSup=0.35 and MinConf=0.75.

FIGURE 4. COMPARAISON BETWEEN SAFCALM, IAR, AND APRIORI ALGORITHMS ACCORDING TO THE RUNTIME CRITERION

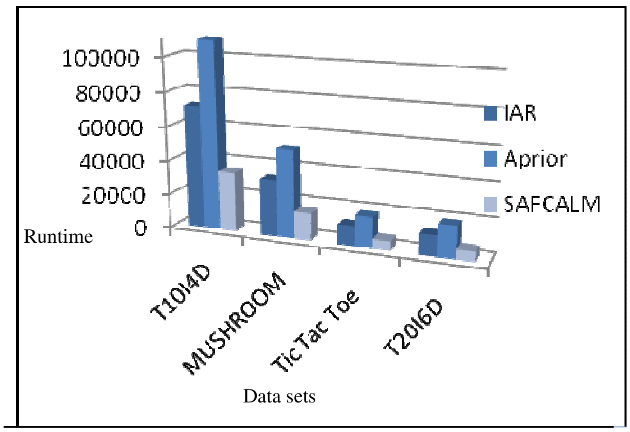


Figure 4 shows the runtime measured in seconds for the three algorithms. We observe that the three methods keep the same behaviour overall the data sets. In fact, the Apriori algorithm needs the greater time because it is based on an exhaustive approach to test all the possible combinations. Consequently, it is recognizable that our method SAFCALM efficiently outperforms the two algorithms IAR and Apriori thanks to its semantic selection of the frequent itemsets during the association rules extraction through the use of the lift measure. Accordingly, the uninteresting itemsets will be discarded thus the needed runtime will greatly decreases compared to other methods.

6. Conclusion

In this paper, we focused on the association rules extraction based on formal concept analysis. Assuming that an itemset is presented by the intent and the extent of a Formal Concept, we introduce a new approach SAFCALM backboned on lift measure to provide a semantic relationship on the formal concept used for association rule extraction. The carried out experiments of our proposal showed the performance of our method compared to the pionnering approaches in the same trend.

Avenues for future work mainly address the following issues: (1) The incorporation of uncertain items on the dataset, (2) the integration of constraints to better assist the user on his semantic extraction of frequent itemsets.

References

[1] R. Wille. Restructuring Lattice Theory : an Approach Based on Hierarchies of Concepts. In Proc. Symposium on Ordered Sets, pp.445-470, Dordrecht-Boston: reidel, 1982.

[2] Ganter B., Wille R., Formal Concept Analysis: Mathematical Foundations, Edition: Springer, Heidelberg-Berlin-New York, 1999

[3] Jaoua A., Beaulieu J. M., Belkhiter N., Deshernais J., Reguig M., Optimal rectangular decomposition of a finite binary relation, International Conference on Discrete Mathematics (sixth conference, June 1992, Vancouver-Canada).

[4] Agrawal R., Imielinski T, Swami A.: Mining association rules between sets of items in large databases. ACM SIGMOD Int. Conf. on Management of Data (1993) 207-213

[5] Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L., & Stumme, G. (2000). Mining minimal non-re- Mining minimal non-redundant association rules using frequent closed itemsets. In Proceedings of the International Conference DOOD'2000, LNCS, Springer-Verlag, (pp. 972–986), London, UK.

[6] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining frequent patterns with counting inference. SIGKDD Explorations, 2(2), 66–75.

[7] Pasquier, N., Bastide, Y., Taouil, R., Stumme, G., & Lakhal, L. (2005). Generating a condensed representation for association rules. Journal of Intelligent Information Systems, 24(1), 25–60.

[8] Zaki, M. J. (2004). Mining Non-Redundant Association Rules. Data Mining and Knowledge Discovery, (9), 223– 248.

[9] Meo, R., Psaila, G., & Ceri, S. (1996). A new SQL-like operator for mining association rules. In Proceedings of the VLDB Conference, (pp. 122–133), Bombay, India, 1996.

[10] Ng, R. T., Lakshmanan, V. S., Han, J., & Pang, A. (1998). Exploratory mining and pruning optimizations of constrained association rules. In Proceedings ACM SIGMOD International Conference on Management of Data, (pp. 13–24), Seattle, Washington, USA.

[11] Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., & Yamaguchi, T. (2004). Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In Proceedings of the International Conference PKDD'04, Springer-Verlag, 3202, 362–373, Pisa, Italy.

[12] Srikant, R., Vu, Q., & Agrawal, R. (1997). Mining association rules with item constraints. In Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, (pp. 67–73), Newport Beach, California, USA.

[13] Han, J., & Fu, Y. (1995). Discovery of multiple-level association rules from large databases. In Proceedings of the VLDB Conference, (pp. 420–431).

[14] Brin, S., Motawni, R., & Silverstein, C. (1997). Beyond market baskets: Generalizing association rules to correlations. In Proceedings of ACM SIGMOD International Conference on Management of Data, ACM Press, (pp. 265–276), Tucson, Arizona, USA.

[15] Ganter, B., Stahl, J., Wille, R. (1986). Conceptual measurement and many--valued contexts, Classification as a tool of research Amsterdam: North--Holland (1986) , p. 169-176.

[16] Bayardo R., Agrawal R.: Mining the most interesting rules. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (1999) 145-154.

[17] Cherfi H., Toussaint Y.:Text mining by combining association rules and statistics. Int. Workshop on Text Mining CIFT-02, Hammamet – Tunisia (October 21-23, 2002) 67-80.

[18] Maddouri M, A Formal Concept Analysis Approach to discover association rules from data, Radim Belohlavek, Vaclav Snasel (Eds.): CLA 2005, pp. 10–21, ISBN 80–248–0863–3.

[19] Ganter B., Wille R., Formal Concept Analysis: Mathematical Foundations, Edition: Springer, Heidelberg-Berlin-New York, 1999

[20] Jaoua A., Beaulieu J. M., Belkhiter N., Deshernais J., Reguig M., Optimal rectangular decomposition of a finite binary relation, International Conference on Discrete Mathematics (sixth conference, June 1992, Vancouver-Canada).