TOWARDS AN ENHANCED SEMANTIC APPROACH BASED ON FORMAL CONCEPT ANALYSIS AND LIFT MEASURE

Ourida Saidi

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

ABSTRACT
The volume of stored data increases rapidly. Therefore, the battery of extracted association heavily prohibits the better support of the decision maker. In this context, 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 interesting itemsets through the concept coverage. This approach uses an original quality-criterion of a rule namely the profit improving the classical formal concept analysis through the addition of semantic value in order to extract meaningful association rules.

KEYWORDS
Association rules, formal concept analysis, quality measure.

1. INTRODUCTION
The analysis of huge volumes of data is fundamental to explore the unknown information and extract the hidden patterns namely the association rules. The “Concept” is a couple of intent and extent used to represent nuggets of knowledge. A new trend recently appears focusing on building theoretical foundations for data-Mining based on Formal Concept Analysis [1,2,3]. Accordingly, various motivating proposals have been advanced related to association rules [4].

Some are constraint-based in that they generate rule satisfying a set of metrics such as the minimum support threshold or confidence threshold. Others are heuristic through deriving predictive rules but suffer some disadvantages such as the incapability to guarantee the completeness of the returned set of rules (i.e. decision trees). Another class of extraction of rules determinate the most interesting, called optimal, rules in the respect of given interestingness metrics. Such class is extremely useful specially in real domains due to its minimal time consuming and simplicity.

In this paper, we focus on the last class. We introduce an original approach of association rules mining based on Formal Concept Analysis. The remainder of the paper is organized as follows. We devote Section 2 to works related to the association rules extraction problem. Section 3 sketches the mathematical background of FCA and its connection with the derivation of association rule bases. We introduce, in Section
4, an heuristic algorithm to calculate the optimal itemsets. 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 points out future work.

2. RELATED WORKS

Usually, 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 become a challenging task.

To resolve this issue, several approaches were sharply proposed [17]. 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] 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 \( \chi^2 \)-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. The authors show that the best rule according to any of these metrics must reside along a support/confidence border. In addition, in the case of conjunctive rule mining within categorical data, the number of generated rules along this border is suitably small, and can be well mined from a diversity of real-world data-sets. Thus, returning a broader set of rules than classical algorithms allows for improved insight into the data and support more user-interaction in the optimized rule-mining process.

Moreover, Cherfi et al [17] propose five different measures namely the benefit (interest) and the satisfaction to define a set of rules that contains the most interesting rule according to any of the above metrics, even if one requires this rule to characterize a specific subset of the population of interest.

Maddouri et al provide the gain measure [18] to prune the non frequent itemsets in order to reduce the number of generated rules which become exhaustive particularly for large size data sets.

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**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>o1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>o2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>o3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>o4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>o5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**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:

**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:
- Supi∈I(Ai, Bi) = (∪i∈IAi) ⊨ (∩i∈IBi)
- Infi∈I(Ai, Bi) = (∩i∈IAi) ⊲ (∪i∈IBi)

**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 \(\Phi (O, P, R)\), such that any couple \((o, p)\) in the context \((O, P, R)\) is included in at least one concept of \(CV\).

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. DISCOVERY OF OPTIMAL ITEMSETS

As well known, 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 $\text{FC}_i = (A_i, B_i)$ be a formal concept. We define:
- Length of a concept $\text{FC}_i$: the number of properties in the intent $B_i$ of the concept.
- Width of a concept $\text{FC}_i$: the number of objects in the extent $A_i$ of the concept.
- Lift of a concept $\text{FC}_i$: the maximum lift of the set of rules generated from the concept $\text{FC}_i$.
- Profit of a concept: is a function of the width the length and the lift of the concept, given by:

$$\text{Profit}(\text{FC}_i) = \frac{\text{length}(\text{FC}_i) \cdot \text{width}(\text{FC}_i) \cdot \text{lift}(\text{FC}_i)}{\text{length}(\text{FC}_i) + \text{width}(\text{FC}_i)}$$

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 $\text{FC}_i = (A_i, B_i)$ containing a couple $(o, p)$ is said to be optimal if it maximizes the profit function.

Definition 9 [20]: A coverage $\text{CV} = \{ \text{FC}_1, \text{FC}_2, ..., \text{FC}_k \}$ 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 $\text{FC}_3$ containing the couple $(o_3, B)$. Figure 2.c represents the non optimal concept $\text{FC}_2$ containing the couple $(o_3, B)$. The optimal coverage of the context $(O, P, R)$ is formed by three optimal concepts: $\{\text{FC}_4, \text{FC}_5, \text{FC}_6\}$. $\text{FC}_4$ is the concept containing the items $\{(o_1, o_2), (A, B)\}$. $\text{FC}_5$ is the concept containing the items $\{(o_3, o_4), (B, C)\}$. $\text{FC}_6$ is the concept containing the items $\{(o_4, o_5), (C, D)\}$.

4.1. Heuristic Searching for Optimal Concept

The pseudo-concept, denoted by $\text{PCF}$, 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 $\{p\}$ $\triangleright$, and the set of properties describing the object $o$, so $\{o\}$ $\triangleleft$. Where $\triangleright$, $\triangleleft$ is the Galois connection of the context $(O, P, R)$.

When we determinate the pseudo-concept $\text{PCF}$, two cases are considered:
- Case 1: $\text{PCF}$ forms a formal concept.
  - If no zero is found in the relation/matrix representing $\text{PCF}$, then $\text{PCF}$ is the optimal concept. So, the algorithm stops.
  - Case 2: $\text{PCF}$ is not a formal concept.
  - If some zero entries are found in the relation/matrix representing $\text{PCF}$, we will look for more restraint pseudo-concepts within the pseudo-concept $\text{PCF}$.
  
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 $\text{PCF}$.

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

Definition 10: Let $\text{PCF}_i = (A_i, B_i, R_i)$ 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 $\text{PCF}_i$: the number of properties in $B_i$.
- Width of a pseudo-concept $\text{PCF}_i$: 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:

\[
\text{Size}(PCF) = (\text{length}(PCF) \times \text{width}(PCF))
\]

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

\[
\text{Profit}(PCF) = \left[ \frac{\text{Size}(PCF)}{\text{length}(PCF) + \text{width}(PCF)} \right] \times \left[ (\text{length}(PCF) + \text{width}(PCF)) - \text{Size}(PCF) \right]
\]

4.2. Algorithm for Optimal Coverage

The problem of covering a binary relation by a set of optimal concepts can be resolved by 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 problem, being NP-Complete problem, has been the subject of several previous works. However, 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 $R_{i-1} = P_1 \cup ... \cup P_{i-1}$ 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 $P_1, ..., P_n$.

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

$CV_i := CV_{i-1} \cup \{r \in CV_{i-1} / r \notin FC \} \cup FC$.

Delete all the redundant concepts from $CV_i$.

$P_i := P_i \setminus \{(X, Y) \in P_i / (X, Y) \notin FC\}$

End While

END FOR

End.

---

**Figure 3. Incrementation Phase when Adding $P_4$**

---

a. Optimal coverage of the context ($\{01, 02, 03\}$ [A,B,C,D])

b. Case 1: Coverage of the context ($\{01, 02, 03, 04\}$ [A,B,C,D])
Example: Let $R$ be the relation to cover as highlighted by Table 1. $R$ is partitioned into five packages:
- $P_1 = \{o_1\} \times \{A, B\}$,
- $P_2 = \{o_2\} \times \{A, B\}$,
- $P_3 = \{o_3\} \times \{B, C\}$,
- $P_4 = \{o_4\} \times \{B, C, D\}$ and
- $P_5 = \{o_5\} \times \{C, D\}$.

Initially, $R$ is an empty relation and in each phase we add a package. Figure 3 sketches the incrementation phase when adding $P_4$. In this phase, $R_3$ encloses the four rows $P_1$, ..., $P_3$. The initial optimal coverage $CV_3$ encloses the formal concepts $FC_3 = \{\{o_1, o_2\}, \{A, B\}\}$ and $FC_4 = \{\{o_3\}, \{B, C\}\}$.

The package $P_4$ encloses only three couples: $(o_3, B)$, $(o_3, C)$ and $(o_3, D)$. The pseudo concept containing the couple $(o_4, B)$ and $(o_4, C)$ is:

- **Case 1**: the union of formal concepts $\{\{o_3, o_4\}, \{B, C\}\}$ and $\{\{o_4\}, \{B, C, D\}\}$
- **Case 2**: the formal concept $\{\{o_4\}, \{B, C, D\}\}$

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:

$$Profit(PCF) = \frac{\text{Size}(PCF)}{\text{length}(PCF) + \text{width}(PCF) - \text{Size}(PCF)} * \right\} \text{(length}(PCF) + \text{width}(PCF) - \text{Size}(PCF))$$

Hence, the retained formal concept is $FC_6 = \{\{o_4\}, \{B, C, D\}\}$.

Thus, the final coverage of $R$ contains the concepts $FC_3$, $FC_4$ and $FC_6$. Finally, according to our example, we find three Item-sets: $\{A, B\}$, $\{B, C\}$ and $\{C, D\}$.

### 5. Experimental study

It was worth the effort to experiment 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>
<thead>
<tr>
<th>Dataset</th>
<th># Transactions</th>
<th>#Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1014D</td>
<td>1000000</td>
<td>100</td>
</tr>
<tr>
<td>Mushroom</td>
<td>8124</td>
<td>128</td>
</tr>
<tr>
<td>T2016D</td>
<td>1000</td>
<td>9</td>
</tr>
<tr>
<td>Tic Tac Toe</td>
<td>958</td>
<td>10</td>
</tr>
</tbody>
</table>

To shed light on the performance of our approach, we compare our proposal to the three pioneering methods in the same trend in the literature 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 shows the runtime measured in seconds for the three algorithms. We notice that the three methods keep the same behavior overall the data sets. The Apriori algorithm needs the greater time because it is based on an exhaustive approach to test all the possible combinations. 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. So, the non-interesting itemsets will be discarded; thus the needed runtime will greatly decreases compared to other methods.

6. Conclusion

In this paper, we discussed the extraction association rules 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 pioneering approaches in the same trend.

Avenues for future work mainly address the following issues: (1) The application of visualization within a user-driven approach, (2) the extraction of closed itemsets using our enhanced semantic approach of association rules extraction based on formal concept.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

References


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


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


