Semantic Knowledge Discovery from Heterogeneous Data Sources

Claudia d'Amato[1], Volha Bryl[2], Luciano Serafini[2]

[1] Department of Computer Science - University of Bari, Italy
claudia.damato@di.uniba.it
[2] Data & Knowledge Management Unit - Fondazione Bruno Kessler, Italy
{bryl|serafini}@fbk.eu

Abstract. Available domain ontologies are increasing over the time. However there is a huge amount of data stored and managed with RDBMS. We propose a method for learning association rules from both sources of knowledge in an integrated way. The extracted patterns can be used for performing: data analysis, knowledge completion, ontology refinement.

1 Introduction

From the introduction of the Semantic Web view [4], many domain ontologies have been developed and stored in open access repositories. However, still huge amounts of data are managed privately with RBMS by industries and organizations. Existing domain ontologies may describe domain aspects that complement data in RDMS. This complementarity could be fruitfully exploited for setting up methods aiming at (semi-)automatizing the ontology refinement and completion tasks as well as for performing data analysis. Specifically, hidden knowledge patterns could be extracted across ontologies and RDBMS. To this aim, an approach for learning *association rules* [1] from hybrid sources of information is proposed. Association rule mining methods are well know in Data Mining [12]. They are generally applied to propositional data representations with the goal of discovering patterns and rules in the data. To the best of our knowledge, there are very few works concerning the extraction of association rules from hybrid sources of information. For better explaining the intuition underlying our proposal, let us consider the following example. Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be an ontology expressed in Description Logics (DLs) [3], composed of a Tbox $\mathcal{T}$ describing general knowledge on kinships and an Abox $\mathcal{A}$ on the kinships of a group of people.

$$\mathcal{T} = \left\{ \begin{array}{lll} \mathsf{Person} \equiv \mathsf{Man} \sqcup \mathsf{Woman} & \mathsf{Man} \sqsubseteq \neg \mathsf{Woman} & \top \sqsubseteq \forall \mathsf{hasChild}.\mathsf{Person} \\ \exists \mathsf{hasChild}.\top \sqsubseteq \mathsf{Person} & \mathsf{Parent} \equiv \exists \mathsf{hasChild}.\mathsf{Person} & \mathsf{Mother} \equiv \mathsf{Woman} \sqcap \mathsf{Parent} \\ \mathsf{Father} \equiv \mathsf{Man} \sqcap \mathsf{Parent} & \mathsf{Grandparent} \equiv \exists \mathsf{HasChild}.\mathsf{Parent} & \mathsf{Child} \equiv \exists \mathsf{HasChild}^{-}.\top \end{array} \right\}$$

$$\mathcal{A} = \left\{ \begin{array}{llll} \mathsf{Woman(alice)} & \mathsf{Man(xavier)} & \mathsf{hasChild(alice, claude)} & \mathsf{hasChild(alice, daniel)} \\ \mathsf{Man(bob)} & \mathsf{Woman(yoana)} & \mathsf{hasChild(bob, claude)} & \mathsf{hasChild(bob, daniel)} \\ \mathsf{Woman(claude)} & \mathsf{Woman(zurina)} & \mathsf{hasChild(xavier, zurina)} & \mathsf{hasChild(yoana, zurina)} \\ \mathsf{Man(daniel)} & \mathsf{Woman(maria)} & \mathsf{hasChild(daniel, maria)} & \mathsf{hasChild(zurina, maria)} \end{array} \right\}$$

Given an ontology and a DL reasoner, it is possible to derive new knowledge that is not explicitly asserted in $\mathcal{K}$. For instance, in the example above it is possible to derive that alice is a Mother and xavier is a Father. Let $\mathbf{D} \subseteq$ Name × Surname × Qualification × Salary × Age × City × Address be a job information database (see Tab. 1, for simplicity a single table is used). The link between $\mathcal{K}$ and $\mathbf{D}$ is given by $\{(\mathsf{alice}, P001), (\mathsf{xavier}, p003), (\mathsf{claude}, p004), (\mathsf{daniel}, p005), (\mathsf{yoana}, p006), (\mathsf{zurina}, p007), (\mathsf{maria}, p008)\}$ where the first element is an individual of $\mathcal{K}$ and the second element is an attribute value of $\mathbf{D}$.

ID	Name	Surname	Qualification	Salary	Age	City	Address
p001	Alice	Lopez	Housewife	0	60	Bari	Apulia Avenue 10
p002	Robert	Lorusso	Bank-employee	30.000	55	Bari	Apulia Avenue 10
p003	Xavier	Garcia	Policeman	35.000	58	Barcelona	Carrer de Manso 20
p004	Claude	Lorusso	Researcher	30.000	35	Bari	Apulia Avenue 13
p005	Daniel	Lorusso	Post Doc	25.000	28	Madrid	calle de Andalucia 12
p006	Yoana	Lopez	Teacher	34.000	49	Barcelona	Carrer de Manso 20
p007	Zurina	Garcia-Lopez	Ph.D student	20.000	25	Madrid	calle de Andalucia
p008	Maria	Lorusso	Pupil	0	8	Madrid	calle de Andalucia

Table 1. The job information database

Given a method for analyzing jointly the available knowledge sources, it could be possible to induce more general information such as *Women that earn more money are not mothers*. The knowledge of being Woman and Mother comes from the ontology and the knowledge on the salary comes from $\mathbf{D}$. In the following, the approach for accomplishing such a goal based on learning association rules is illustrated.

2 The Framework

Association rules [1] provide a form of rule patterns for data mining. Let $\mathbf{D}$ be a dataset made by a set of attributes $\{A_1, \ldots, \mathcal{A}_n\}$ with domains $D_i :\ i \in \{1, \ldots, n\}$. Learning association rules from $\mathbf{D}$ consists in finding rules of the form $((A_{i_1} = a) \wedge \cdots \wedge (A_{i_k} = t)) \Rightarrow (A_{i_k+1} = w)$ where $a, \ldots, t, w$ are values in $D_{i_1}, \ldots, D_{i_k}, D_{i_k+1}$. The pattern $(A_{i_1} = a) \wedge (A_{i_2} = b) \wedge \cdots \wedge (A_{i_k} = t)$ is called *itemset*. An association rule has the general form $\theta \Rightarrow \varphi$ where θ and φ are itemset patterns. Given the itemset θ, the *frequency* of θ (*fr*(θ)) is the number of cases in $\mathbf{D}$ that match θ. The frequency of $\theta \wedge \varphi$ (*fr*$(\theta \wedge \varphi)$) is called *support*. The *confidence* of a rule $\theta \Rightarrow \varphi$ is the fraction of rows in $\mathbf{D}$ that match φ among those rows that match θ, namely *conf*$(\theta \Rightarrow \varphi) = fr(\theta \wedge \varphi)/fr(\theta)$. A frequent itemset expresses the variables and the corresponding values that occur reasonably often together in $\mathbf{D}$.

The algorithms for learning association rules typically divide the learning problem into two parts: 1) finding the frequent itemsets w.r.t. a given support threshold; 2) extracting the rules from the frequent itemsets satisfying a given confidence thresholds. The solution to the first subproblem is the most expensive one, hence most of the algorithms concentrate on finding optimized solutions to this problem. The most well known algorithm is Apriori [1]. It is grounded on the key assumption that a set X of variables can be frequent only if all the subsets of X are frequent. The frequent itemsets are discovered as follows:

```
Apriori($\mathbf{D}$:dataset, sp-tr: support threshold): $L$ frequent itemsets
$L = \emptyset$;     $L_1$ = {frequent itemsets of length 1}
for (k = 1; $L_k \neq \emptyset$; k++) do
    $C_{k+1}$ = candidates generated by joining $L_k$ with itself
    $L_{k+1}$ = candidates in $C_{k+1}$ with frequency equal or greater than sp-tr
    $L = L \cup L_{k+1}$
return $L$;
```

As a first step, all frequent sets L_{1_i} (w.r.t. to a support threshold) consisting of one variable are discovered. The candidate sets of two variables are built by joining L_1 with itself. By depurating them of those sets having frequency lower than the fixed threshold, the sets L_{2_i} of frequent itemsets of length 2 are obtained. The process is iterated, incrementing the length of the itemsets at each step, until the set of candidate itemsets is empty. Once the set L of all frequent itemsets is determined, the association rules are extracted as follows: 1) for each $I \in L$, all nonempty subsets S of I are generated; 2) for each S, the rule $S \Rightarrow (I - S)$ is returned **iff** $(fr(I)/fr(S)) \geq$ *min-confidence*, where *min-confidence* is the minimum confidence threshold.

The basic form of APRIORI focuses on propositional representation. There exist several upgrades focusing on different aspects: reduction of computational costs for finding the set of frequent items [9], definition of heuristics for pruning patterns and/or assessing their *interestingness* [9], discovery of association rules from multi-relational settings, i.e. relational and/or distributed databases [6, 8, 7], DATALOG programs [11, 5]. Algorithms focusing on this third aspect usually adopt the following approach: 1) the entity, i.e. the attribute/set of attributes, of primary interest for extracting association rules is determined; 2) a view containing the attributes of interest w.r.t. the primary entity is built. Moving from this approach, a method for building an integrated data source, containing both data of a database $\mathbf{D}$ and of an ontology $\mathcal{K}$, is proposed. Consequently association rules are learnt. The approach is grounded on the assumption that $\mathbf{D}$ and $\mathcal{K}$ share (a subset of) common individuals. This assumption is reasonable in practice. An example is given by the biological domain where research organizations have their own databases that could be complemented with existing domain ontologies. The method for building an integrated source of information involves the following steps:

1. choose the primary entity of interest in $\mathbf{D}$ or $\mathcal{K}$ and set it as the first attribute A_1 in the table $\mathbf{T}$ to be built; A_1 will be the primary key of the table
2. choose (a subset of) the attributes in $\mathbf{D}$ that are of interest for A_1 and set them as additional attributes in $\mathbf{T}$; the corresponding values can be obtained as a result of a SQL query involving the selected attributes and A_1
3. choose (a subset of) concept names $\{C_1, \ldots, C_m\}$ in $\mathcal{K}$ that are of interest for A_1 and set their names as additional attribute names in $\mathbf{T}$
4. for each $C_k \in \{C_1, \ldots, C_m\}$ and for each value a_i of A_1, if $\mathcal{K} \models C_k(a_i)$ then set to 1 the corresponding value of C_k in $\mathbf{T}$, set the value to 0 otherwise
5. choose (a subset of) role names $\{R_1, \ldots, R_t\}$ in $\mathcal{K}$ that are of interest for A_1 and set their names as additional attribute names in $\mathbf{T}$
6. for each $R_l \in \{R_1, \ldots, R_t\}$ and for each value a_i of A_1, if $\exists y \in \mathcal{K}\ s.t.\ \mathcal{K} \models R_l(a_i, y)$ then set to 1 the value of R_l in $\mathbf{T}$, set the value to 0 otherwise
7. choose (a subset of) the datatype property names $\{T_1, \ldots, T_v\}$ in $\mathcal{K}$ that are of interest for A_1 and set their names as additional attribute names in $\mathbf{T}$
8. for each $T_j \in \{T_1, \ldots, T_v\}$ and for each value a_i of A_1, if $\mathcal{K} \models T_j(a_i, \mathit{dataValue}_j)$ then set to $\mathit{dataValue}_j$ the corresponding value of T_j in $\mathbf{T}$, set 0 otherwise.

The choice of representing the integrated source of information within tables allows us to avoid the migration of large amount of data stored in RDMS in alternative representation models in order to extract association rules and also allows for directly applying state of the art algorithms for learning association associations.

In the following, the proposed method is applied to the example presented in Sect. 1. Let NAME be the primary entity and let QUALIFICATION, SALARY, AGE, CITY be the selected attributes from $\mathbf{D}$. Let Woman, Man, Mother, Father, Child be the selected concept names from $\mathcal{K}$ and let HasChild be the selected role name. The attribute values in the table are obtained as described above. Numeric attributes are pre-processed (as usual in data mining) for performing data discretization [12] namely for transforming numerical values in corresponding range of values. The final resulting table is shown

NAME	QUALIFICATION	SALARY	AGE	CITY	HasChild	Woman	Man	Mother	Father	Child
Alice	Housewife	[0,14999]	[55,65]	Bari	1	1	0	1	0	0
Robert	Bank-employee	[25000,34999]	[55,65]	Bari	0	0	0	0	0	0
Xavier	Policeman	[35000,44999]	[55,65]	Barcelona	1	0	1	0	1	0
Claude	Researcher	[25000,34999]	[35,45]	Bari	0	1	0	0	0	1
Daniel	Post Doc	[15000,24999]	[25,34]	Madrid	1	0	1	0	1	1
Yoana	Teacher	[25000,34999]	[46,54]	Barcelona	1	1	0	1	0	0
Zurina	Ph.D student	[15000,24999]	[25,34]	Madrid	1	1	0	1	0	1
Maria	Pupil	[0,14999]	[0,16]	Madrid	0	1	0	0	0	1

Table 2. The integrated data source

in Tab. 2. Once the integrated data source has been obtained, the APRIORI algorithm is applied to discover the set of frequent items, hence the association rules are lernt. By applying[1] APRIORI to Tab. 2, given a support threshold *sp-tr* $= 0.2$ (namely 20% of the tuples in the table) and a confidence threshold 0.7, some association rules learnt are:

1. SALARY=$[15000, 24999] \Rightarrow (\mathsf{HasChild} = 1) \wedge (\mathsf{Child} = 1)$ (100%)
2. $(\mathsf{Woman} = 1) \Rightarrow (\mathsf{Man} = 0)$ (100%)
3. (AGE=$[25, 34]) \wedge$ (CITY =Madrid) $\Rightarrow (\mathsf{HasChild} = 1)$ (100%)
4. $(\mathsf{HasChild} = 1) \wedge (\mathsf{Man} = 1) \Rightarrow (\mathsf{Father} = 1)$ (100%)

The first rule means that if someone earns between 15000 and 24999 euro, he/she has a 100% confidence of having a child and being a Child. The third rule means that if someone is between 25 and 34 years old and lives in Madrid, he/she has a 100% confidence of having a child. The other two rules can be interpreted similarly. Because of the very few tuples in Tab. 2 and quite high confidence threshold, only rules with the maximum confidence value are returned. By decreasing the confidence threshold, i.e. to 0.6, additional rules can be learnt such as (CITY =Madrid) $\Rightarrow (\mathsf{Parent} = 1) \wedge (\mathsf{HasChild} = 1) \wedge (\mathsf{Child} = 1)$ (66%). The method for learning association rules exploits the evidence of the data. Hence it is not suitable for small datasets.

Association rules extracted from hybrid data sources can be used for performing data analysis. For example rule (3) suggests the average age of being a parent in Madrid that could be different in other geographical areas, e.g. Bari. These rules can be also exploited for data completion both in $\mathcal{K}$ and $\mathbf{D}$. For instance, some individuals can be asserted to be an instance of the concept Child in $\mathcal{K}$. Also rules (2), (4) that could seem trivial since they encode knowledge already modeled in $\mathcal{K}$, can be useful for the ontology refinement taks. Indeed, rules come up from the assertional data. Hence, it is possible to discover intentional axioms that have not been modeled in the ontology. If in the TBox in the example there was no disjointness axiom for Man and Woman but the data in the ABox extensively contained such information (that is our case), rule (2) mainly suggests a disjointness axiom. Similarly for (4).

3 Discussion

The main goal of this work is to show the potential of the proposed approach. Several improvements can be done. In building the integrated data source, concepts and roles are treated as boolean attributes thus adopting an implicit *Closed World Assumption*. To cope with the *Open Wolrd* semantics of DLs, three valued attributes could be considered for treating explicitly unknown information. Concepts and roles are managed without considering inclusion relationships among them. The treatment of this information could save computational costs and avoid the extraction of redundant association

[1] The Weka toolkit could be easily used for the purpose http://www.cs.waikato.ac.nz/ml/weka/.

rules. Explicitly treating individuals that are fillers of the considered roles could be also of interest. It could be also useful to consider the case when an individual of interest is a filler in the role assertion. Currently these aspects are not managed. An additional improvement is applying the algorithm for learning association rules directly on a relational representation, without building an intermediate propositional representation.

To the best of our knowledge there are very few works concerning the extraction of association rules from hybrid sources of information. The one most close to ours is [10], where a hybrid source of information is considered: an ontology and a constrained DATALOG program. Association rules at different granularity levels (w.r.t. the ontology) are extracted, given a query involving both the ontology and the DATALOG program. In our framework, no query is specified. A collection of data is built and all possible patterns are learnt. Some restrictions are required in [10], i.e. the set of DATALOG predicate symbols has to be disjoint from the set of concept and role symbols in the ontology. In our case no restrictions are put. Additionally, [10] assumes that the alphabet of constants in the DATALOG program coincides with the alphabet of the individuals in the ontology. In our case a partial overlap of the constants would be sufficient.

4 Conclusions

A framework for learning association rules from hybrid sources of information has been presented. Besides discussing the potential of the proposed method, its current limits have been analyzed and the wide spectrum of lines of research have been illustrated. For the future we want to investigate on: 1) the integration of the learnt association rules in the deductive reasoning procedure; 2) alternative models for representing the integrated source of information.

References

1. R. Agrawal, T. Imielinski, A. Swami. Mining Association Rules Between Sets of Items in Large Databases. SIGMOD Conference, p. 207-216. (1993)
2. R. Agrawal, R. Srikant. Fast algorithms for mining association rules in large databases. Proc. of the Int. Conf. on Very Large Data Bases (VLDB'94). (1994).
3. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider. The Description Logic Handbook. Cambridge University Press. (2003).
4. T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web. Scient. Amer. (2001).
5. L. Dehaspe, H. Toivonen. Discovery of frequent DATALOG patterns. Journal of Data Mining and Knowledge Discovery. Vol. 3(1), p. 7–36. (1999).
6. S. Džeroski. Multi-Relational Data Mining: an Introduction. SIGKDD Explor. Newsl. Vol. 5(1), p. 1–16. ACM. (2003).
7. B. Goethals, W. Le Page, M. Mampaey Mining Interesting Sets and Rules in Relational Databases. Proc. of the ACM Symp. On Applied Computing. (2010).
8. Y. Gu, H. Liu, J. He, B. Hu, X. Du. MrCAR: A Multi-relational Classification Algorithm based on Association Rules. Proc. of WISM'09 Int. Conf. (2009).
9. D. Hand, H. Mannila, P. Smyth. Principles of data mining. Ch. 13. Adaptive Computation and Machine Learning Series. MIT Press. (2001).
10. F. A. Lisi. AL-QuIn: An Onto-Relational Learning System for Semantic Web Mining. Int. J. of Sem. Web and Inf. Systems. IGI Global. (2011).
11. S.-L. Wang, T.-P. Hong, Y.-C. Tsai, H.-Y. Kao Multi-table association rules hiding Proc. of the IEEE Int. Conf. on Intelligent Syst. Design and Applications. (2010).
12. I. H. Witten, E. Frank, M. A. Hall Data Mining: Practical Machine Learning Tools and Techniques (3rd Ed.) Morgan Kaufmann. (2011).