MINING INFLUENCE RULES OUT OF ONTOLOGIES

Barbara Furletti and Franco Turini
Department of Computer Science, University of Pisa, Pisa, Italy

Keywords: Ontology Mining, Knowledge Discovery, If-Then Rules.

Abstract: A method for extracting new implicit knowledge from ontologies by using an inductive/deductive approach is presented. By analyzing the relationships that already exist in an ontology, we are able to return the extracted knowledge as weighted If-Then Rules among concepts. The technique, that combines data mining and link analysis, is completely general and applicable to whatever domain. Since the output is a set of "standard" If-Then Rules, it can be used to integrate existing knowledge or for supporting any other data mining process. An application of the method to an ontology representing companies and their activities is included.

1 INTRODUCTION

Knowledge extraction from databases is a consolidated practice that continues to evolve in parallel with the new data management systems. It is based not only on querying systems, but above all, on complex reasoning tools. Today, with the coming of the Web 2.0 and the semantic web, new methods for representing, storing and sharing information are going to replace the traditional systems. Roughly speaking, ontologies "could substitute" in many applications the Data Bases (DBs). Consequently, the interest is moving toward the research of new methods for handling these structures and to efficiently obtain information from them besides what is obtained by using the traditional reasoning systems.

In this paper we aim at contributing to this topic by handling the problem of extracting interesting and implicit knowledge from ontologies, in a novel way with respect to the traditional reasoners methods. By getting hints from the semantic web and data mining environments, we give a Bayesian interpretation to the relationships that already exist in an ontology in order to return a set of weighted IF-Then rules, that we refer to as Influence Rules (IRs).

The idea is to split the extraction process in two separate phases by exploiting the ontology peculiarity of keeping metadata (the schema) and data (the instances) separate. The deductive process draws inference from the ontology structure, both concepts and properties, by applying link analysis techniques and producing a sort of implications (rules schemas) in which only the most important concepts are involved. Then an inductive process, implemented by a data mining algorithm, explores the ontology instances for enriching the implications and building the final rules.

For example, let us suppose we have a fragment of ontology as depicted in Figure 1 that describes companies and the business environment. *Company*, *Manager* and *Project* are concepts, continuous arrows represent properties of the ontologies while the dotted ones are used for connecting instances to the classes they belong to. Starting from this ontology and the corresponding instances we are able, at the end of the process, to produce IRs as the following one:

$$\textit{Manager.hasAge} < \textit{45} \xrightarrow{w=0.80}$$
$$\textit{Project.hasInnovationDegree} = \textit{good}$$

Both the premise (*Manager.hasAge* $<$ *45*) and the consequence (*Project.hasInnovationDegree* = *good*) are expressions binding the datatype property of a class to a specific value, while the weight (w) measures the strength of the influence. This rule must be read as:

> "In 80% of the cases, whenever a manager of a company is less then 45 years old, then the project he manages has a *good* degree of innovation".

What we want to prove, besides the correctness and feasibility[1] of the project, is that the approach allows us to extract "higher level" rules w.r.t. classical knowledge discovery techniques. In fact, ontology metadata gives a general view of the domain

[1] The term feasibility has to be intended as the "capability of being done".

Furletti B. and Turini F..
MINING INFLUENCE RULES OUT OF ONTOLOGIES.
DOI: 10.5220/0003438403230333
In *Proceedings of the 6th International Conference on Software and Database Technologies* (ICSOFT-2011), pages 323-333
ISBN: 978-989-8425-77-5
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

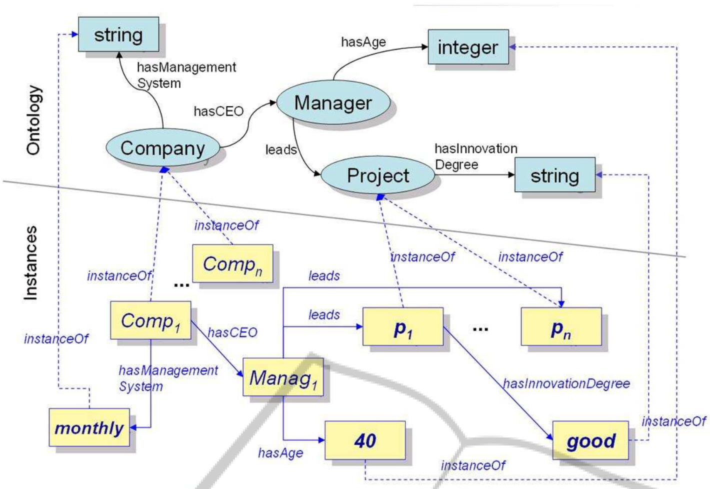


Figure 1: Fragment of an ontology schema and its instances.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

of interest and supplies information about all the elements apart from the fact that they are included as instances in the collected data. The technique is completely general and applicable to every domain. Since the output is a set of "standard" If-Then rules, it can be used to integrate existing knowledge or for supporting any other data mining process.

The paper includes the following sections:
Section 2 proposes some related works that try to combine ontologies and data mining in different ways.
Section 3 gives a short overview of the technical background about theories and algorithms used in the rest of the paper.
Section 4 is the core section in which the Ontology Miner strategy is described.
In section 5 we present a case study where our strategy is applied to an actual problem.
Before concluding this paper, in section 6 we present the new version of the system and some comments about an experiment. Section 7 contains the conclusions and discusses some future promising work.

2 ONTOLOGIES AND DATA MINING

When speaking about ontologies and data mining (DM), we enter into a domain in which DM techniques and domain ontologies are either combined for improving existing knowledge discovery tools and processes or for supporting decision systems. Ontologies and DM are related in different ways depending on the perspective from which the two field are viewed: is the ontology that improves DM or is DM that operates on ontologies? Actually, both the perspectives are interesting and three significant research lines can be identified:
1) using ontologies for driving DM;
2) using DM for building ontologies;
3) using ontologies for describing DM processes.

Great efforts are currently spent by researchers in these fields, for example a recent paper of Geller and colleagues (Geller et al., 2005) describes the use of taxonomies for improving the results of association rule mining. The goal is to produce association rules with higher support from a large set of tuples about demographic and personal interest information. Since the collection of people interests tends to be too abstract for actual applications, they use a hierarchy of concepts for raising data instances to higher levels during a pre-processing step, before running the DM algorithm.

A similar approach has been described in (Bellandi et al., 2007) where an ontology in the domain of super market products is used for extracting constraint-based multi-level association rules. In this case the use of an actual ontology (instead of a simple taxonomy) permits the definition of constraints and the use of concepts at different levels of abstraction. In this case the objective is to drive the extraction of rules that fit the user request and need and identify possible target items for seasonal promotions.

On the other hand, since the construction of an ontology is a complex and creative work for the domain experts, DM techniques are often of great help. A very simple/minimal approach is described in (Elsayed et al., 2007), where the Quinlan's C4.5 algorithm is used for building an ontology starting from the generated decision tree. The ontology is constructed by means of a mapping function from the tree elements: root node, internal nodes and decision branches are mapped into OWL classes, while the leaves (which permit the identification of the association rules) are coded as individuals.

A more structured work is presented in (Parekh et al., 2004), where the authors describe how to enrich an existing seed ontology by using text mining techniques, especially by mining the domain specific texts and glossaries/dictionaries in order to find groups of concepts/terms which are related to each other. Even if the extraction of new concepts or instances from text is automatic, the enrichment of the seed ontology is manually done by the experts. The advantage here is the discovering of many important concepts and interesting relationships directly from the data in an automatic way.

Other contributions in this field are described in (Ciaramita et al., 2008) and (Vela and Declerck, 2008).

In (Ciaramita et al., 2008), the authors describe the implementation of an unsupervised system that combines a syntactic parsing, collocation extraction and selectional restriction learning. The system, applied to a set of data (in this case to a molecular biology corpus of data), generates a list of labeled binary relations between pairs of ontology concepts. They demonstrate that the system can be easily applied in text mining and ontology building applications.

In (Ciaramita et al., 2008) a method is sketched for extending existing domain ontologies (or for semi-automatic generating ontologies) on the basis of heuristic rules applied to the result of a multi-layered processing of textual documents. The rules, extracted by using essentially statistical methods, are used for deriving ontology classes from linguistic annotation. The new classes can be added to already existing ontologies or can be used as starting point for a new ontology.

Ontologies are frequently employed also in context-aware systems. As for example in (Singh et al., 2003), they are used for describing both contexts and the DM process in a dynamic way. In particular the authors split the context aware DM into two parts: the representation of the contexts through the ontology and a framework which is able to query the ontology, invoke the mining processes and coordinate them according to the ontology design.

In the light of the above classification, our work can only partially be seen as a contribution to the line one, because what we do is to move from Knowledge Discovery in Databases to Knowledge Discovery in Ontologies by using a combination of DM and Link analysis methods. Indeed, the analysis of the T-Box of an ontology is used to prepare the process of actual mining out of the A-Box (the instances).

3 TECHNICAL BACKGROUND

In our work we combine in a novel way link analysis and DM techniques in order to extract knowledge from ontologies. In this section we introduce the link analysis method we customized and the corresponding extension to the ontology domain. For what it concerns the DM, we used PATTERNIST, a pattern discovery algorithm developed by colleagues at the CNR in Pisa. PATTERNIST is the result of a research activity that has now come to the implementation of a more sophisticated (and documented) system: ConQueSt (Bonchi et al., 2006).

3.1 Link Analysis

In this paper we exploit the peculiarities of HITS (Hypertext Induced Topic Selection) (Kleinberg, 1998), the Kleinberg's algorithm for ranking web pages, to provide a sort of "authority measure" to the ontology concepts. HITS rates web pages based on two evaluation concepts: authority and hub. The authority estimates the content value of the page, while the hub estimates the value of its links to other pages. In other words, a hub is a page with outgoing links and authority is a page with incoming links. Kleinberg observed that there exists a certain natural type of balance between hubs and authorities in the web graph defined by the hyperlinks, and that this fact could be exploited for discovering both types of pages simultaneously.

HITS works as an iterative algorithm applied to the subgraph G_σ of the web graph, derived from a sort of text matching procedure (for further details see the procedure `Subgraph` in (Kleinberg, 1998)) of the query terms σ in the search topic. For this reason it is query-dependent. The core of the algorithm starts from G_σ and computes hub ($y^{<p>}$) and authority ($x^{<p>}$) weights by using an iterative procedure qualified to mutually reinforce the values. It becomes natural to express the mutually reinforcing relationship between hubs and authorities, as: "If p points to many pages with high x-values, then it should receive a large y-value, and if p is pointed to by many pages with large y-values, then it should receive a large x-value". $\mathcal{I}$ and $\mathcal{O}$ operations have been defined for updating the weights.

$\mathcal{I}$ updates the authority x-weights as:

$$\mathcal{I} : x^{<p>} \leftarrow \sum_{q:(q,p)\in E} y^{<q>}$$

$\mathcal{O}$ updates the hub y-weights as:

$$\mathcal{O} : y^{<p>} \leftarrow \sum_{q:(p,q)\in E} x^{<q>}$$

Since the two operations are mutually recursive, a fixed point is needed for guaranteeing the termination of the computation. Even if the number k of iterations is a parameter of the algorithm, it is proven that, with arbitrarily large values of k, the sequences of vectors $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_k$ converge to the fixed points x^* and y^* (Theorem 3.1 in (Kleinberg, 1998)).
As one can guess, and as it happens for the main information retrieval methods, linear algebra supplies "tools" of support for formalizations and proofs.

First, it is possible to represent the graph G_σ in matrix form with the help of an adjacency matrix A. Then, one can easily observe that the iterative and mutual call of $\mathcal{I}$ and $\mathcal{O}$ can be (re)written as:

$$\begin{array}{l} x_i = A^T y_{i-1} \\ y_i = A x_i \end{array} \tag{1}$$

Stated that, it is easy to trace the computation of x^* and y^* back to the mathematical computation of the principal eigenvectors of a matrix A^TA and AA^T, respectively. From 1, after k iterations, we obtain

$$\begin{array}{l} x^{(k)} = (A^TA)^{(k-1)}A^T\mathbf{u} \\ y^{(k)} = (AA^T)^{(k)}\mathbf{u} \end{array} \tag{2}$$

where $\mathbf{u}$ is the initial seed vector for x and y. Equation 2 is the recursive formula for computing the authority and hub vectors at a certain iteration.

For our purposes we customized the HITS algorithm. A short description of HITSxONTO algorithm is presented in following section 3.2.

3.2 HITSxONTO Algorithm

HITSxONTO, the core algorithm, is the customized version of HITS for handling ontologies. It has been recently developed as part of a Ph.D. Thesis (Furletti, 2009). Like HITS, it is based on the concepts of authority and hubness, and its purpose is to measure the importance of the ontology concepts, basing only on the ontology structure (the TBox). In other words, it tries to deduce which concepts can be considered particularly "important" (authorities) and which ones give a particular importance to other concepts (hubs). In this context we are interested in concepts, object properties and in the *is-a* relation. This last element is used for constructing the matrix associated to the ontology that points out direct, indirect and hidden connections. The datatype properties, instead, are not relevant in the ranking procedure.
The main algorithm variant w.r.t. HITS concerns the pre-processing phase, that is the preparation of the input and the general adaptation to the ontology. In the transition from the web to the ontology environment we adopt the following association: an ontology concept is seen as a web page, and an object property is seen as a hyperlink.
HITSxONTO is iterative as HITS, and follows the same core steps.

4 ONTOLOGY MINING STRATEGY

As introduced in section 1, the objective of this method is to extract hidden information from an ontology by operating on the structure and on the instances, separately. The strategy is composed by four main steps, each one dedicated to a particular phase of the extraction. Figure 2 tries to exemplify the procedure that we describe below in detail.

[Step 1] Identification of the Concepts.
This step consists in the analysis of the ontology schema and the extraction of the most relevant concepts.
For the extraction, we exploit the possibility of representing the ontology as a graph with its associated Adjacency Matrix (AM). The AM points out the existence of a direct link between two concepts. Starting from the AM and exploiting the ontology hierarchical structure (defined by the *is-a* property) we compute a Weighted Adjacency Matrix (WAM). It is an nxn matrix where each entry w_{ij} has the following meaning:

$$w_{ij} = \begin{cases} k & \text{if } k \text{ edges from } i \text{ to } j \text{ exist} \\ 0 & \text{otherwise} \end{cases}$$

This matrix permits us to store multiple and hidden connections between concepts that is, the ones among sub-concepts, or parent concepts and sub-concepts that are not directly defined by an explicit link. In other words, we refer to the connections that exist but that are not explicitly represented by an arc in the ontology-graph. A typical case is represented in the following Example 1.

Example 1. Hidden Connections.
Suppose we have the fragment of ontology depicted in Figure 3. A and B are main concepts,

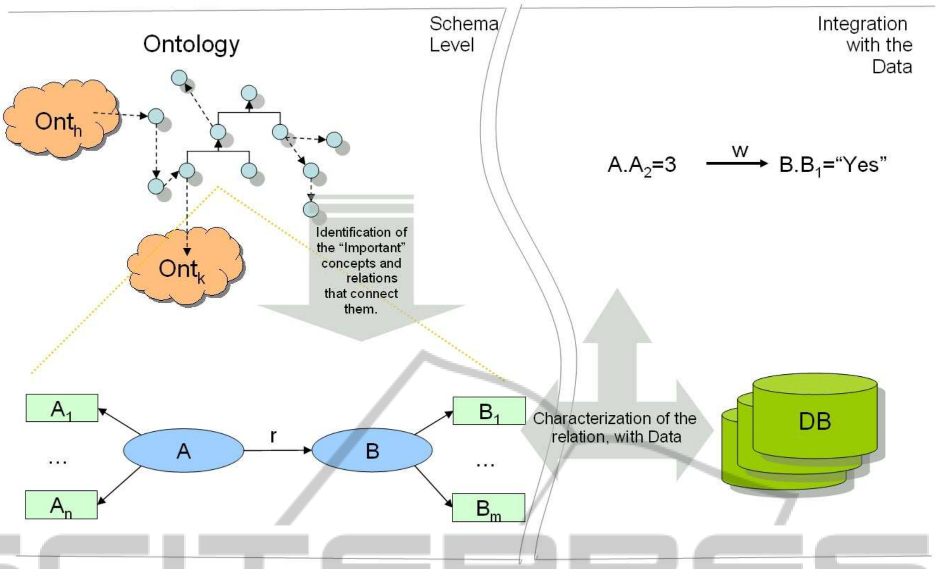
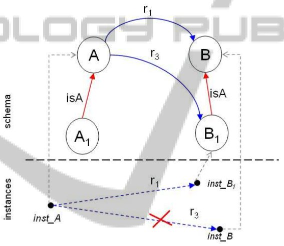

Ontology
Schema Level
Integration with the Data
Ont_h
Ont_k
Identification of the "Important" concepts and relations that connect them.
A.A2=3
w
B.B1="Yes"
Characterization of the relation, with Data
DB
A1
...
An
A
r
B
B1
...
Bm

Figure 2: Steps of analysis.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

while A_1 and B_1 are sub-concepts of A and B, respectively. r_1 and r_3 are object properties and the arrows labelled with *isA* identify the hierarchy. Thanks to these last connections, A_1 inherits from A the status of being domain of the properties r_1 and r_3, while B_1 inherits from B the status of range of the property r_1.
This said, it is easy to see that A is connected to B and B_1 thanks to direct links (r_1 and r_3), but A has actually a "double" connection with B_1: one thanks to the direct link r_3 and the other induced by r_1 and the inheritance property. A_1 has no physical connections with other concepts, nevertheless it inherits from A a simple connection to B and a double connection to B_1. Instead, B does not inherit the range status of B_1 induced by r_3. In fact, given instances $inst\_A \in A$ and $inst\_B \in B$, they cannot be connected by means of r_3. The associated WAM W highlights, for each concept, the number of direct and hidden connections. Since *isA* is a hierarchic relation and not an object property, both the [A_1,A] and [B_1, B] matrix entries are set to 0. As stated before, the contribution of this relation is used for the identification of the hidden connections. □

In order to extract the relevant concepts, we analyse only the schema of the ontology. The idea is to adopt a link analysis method as the one used in the semantic web environment. While HITS works with web pages and hyperlinks, HITSxONTO works on concepts and object properties. Running HITSxONTO with the WAM as input, we obtain two lists of concepts, ranked on authority and hub principles. The most relevant concepts are those that exceed the thresholds for acceptance fixed by the user. Since the threshold strongly depends on the ontology size and connectivity, it has to be empirically fixed.



$$W = \begin{array}{c|cccc} & A & A_1 & B & B_1 \\ \hline A & 0 & 0 & 1_{r1} & 1_{r1}+1_{r3} \\ A_1 & 0 & 0 & 1_{r1} & 1_{r1}+1_{r3} \\ B & 0 & 0 & 0 & 0 \\ B_1 & 0 & 0 & 0 & 0 \end{array}$$

Figure 3: Hidden connections.

[Step 2] Influence Rule Schema Building.
In this step we construct the schemas of the rules, that is we identify the implicant and the implicated concepts, and the direction of the implication. Each rule schema is created by using the potential implicant concepts, and connecting them with the potential implicated concepts reachable directly or indirectly via object properties.
An IR Schemas has the following format:

$$<\texttt{Implicant} \longrightarrow \texttt{Implicated}>$$

where, `Implicant` is a concept belonging to the hub-set of concepts and `Implicated` is a concept belonging to the authority-set of concepts. The following Example 2 clarifies the point.

Example 2. Building the IRs Schema.
Suppose to have an ontology that describes com-

panies and the economic environment, and suppose to obtain, from Step 1, the following two lists of candidates concepts:

Implicant Set = {`ManagementTeam`,
`Company`,
`...`}

Implicated Set =
{`CapitalizationStrategy`,
`DiversificationOfProduction`,
`LevelOfCompetition`,
`...`}

The Implicant and the Implicated sets are composed by concepts that obtained a hub value and an authority value greater than the fixed thresholds respectively. Let us also suppose that in the ontology a connection (a direct object property, an inherited object property or an indirect path of object properties) from *Company* to *LevelOfCompetition* exists. Under these hypothesis the following new schema can be built:

`Company` → `LevelOfCompetition`

This schema is the starting point for the construction of IRs where the concept *LevelOfCompetition* depends on the concept *Company*. The characterization of the schema is realized by associating the appropriate[2] attributes defined as datatype properties of the concept in the ontology. □

[Step 3] Characterization of the Influence Rules Schemas.

In this step we create the IRs starting from the schemas built in the previous step. In particular we associate the appropriate attributes to the concepts that form the schema, and a weight for the implication that identifies the strength of the rules.

To do that, we analyse the ontology instances associated to the set of concepts which the domain of interest is composed of, and we extract the frequent items by using the algorithm PATTERNIST cited at the beginning of section 3. The frequent items give us three important information:

1. The pairs of <concept.attribute> that appear together more frequently in the set of instances.
2. The values associated to the attributes.
3. The support of the frequent item sets, that corresponds to the percentage of the instances that include all items in the premise and consequence in the rule.

We then collect, from the frequent itemsets, the values and the weights for the Influence Rules schemas.

It is important to notice that we consider the support as the appropriate measure for weighting the rules. Other measures, like the confidence, could be a refinement in specific fields, although the support remains the more intuitive measure. Example 3 clarifies the point.

Example 3. Characterizing the IRs Schema. Starting from the result of the previous example 2, let us suppose that the concepts involved in the schema have the datatype properties reported in table 1. In this step 3, we run PATTERNIST on the set of instances of the ontology under analysis. The result is a set of frequent items. Let us suppose that the frequent items are the following two:

FI1: {`LevelOfCompetition.hasType = TypeA, Company.hasFoundationYear = 1989`} `(supp=0.6)`

FI2: {`LevelOfCompetition.hasLevel = High, Company.hasDimension = Big`} `(supp=0.8)`

Merging FI1 and FI2 according to the schemas extracted in step 2 we obtain the following two influence rules.

IR1: *Company.hasFoundationYear = 1989* $\xrightarrow{w=0.6}$ *LevelOfCompetition.hasType = TypeA*

IR2: *Company.hasDimension = Big* $\xrightarrow{w=0.8}$ *LevelOfCompetition.hasLevel = High*

The rules can be read respectively as:
"In 60% of the cases, if the company has been founded in 1989 than its level of competition is of TypeA", and
"In 80% of the cases, if the company is big than its level of competition is high". □

[Step 4] Validation.

The Validation is needed to guarantee that the IRs are consistent and do not conflict with each other. The best way for validating the rules is to ask a domain expert, nevertheless some *ad-hoc* procedures can be implemented with reference to the domain under analysis and the foreseeable use.

The first two steps are essentially deductive, they are a sort of "top-down" approach that starts from the theory and tries to find a model. The third one is an inductive step, a sort of "bottom-up" approach; we move from the observations (the instances) to the results (the IRs).

The methodology we propose can be employed in different DM or non-DM applications that make use of

[2] The appropriate attributes are determined by adopting a particular strategy that uses a DM method on the ontology instances, as described in step 3.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 1: Description of the datatype properties associated to the concepts of the example.

Concept	Datatype Prop.	Type	Options
Company	hasName	String	–
	hasDimention	Enumerated	{Small, Medium, Big}
	hasFoundationYear	Integer	–
LevelOfCompetition	hasLevel	Enumerated	{Low, Medium, High}
	hasDescription	String	–
	hasType	Enumerated	{TypeA, TypeB}

additional information in the form of rules, or for enriching pre-existing knowledge repository and structures (Baglioni et al., 2008).

To complete the discussion, in the next section we show an actual application that uses the IRs in another DM process.

5 CASE STUDY

In this section, we describe an actual application of the methodology described in section 4 in the context of MUSING (Mus, 2006), an European project in the field of Business Intelligence (BI). MUSING, "**MU**lti-industry, **S**emantic-based next generation business **IN**telli**G**ence" aims at developing a new generation of BI tools and modules based on semantic knowledge and content systems. It integrates Semantic Web and Human Language technologies and combines declarative rule-based methods and statistical approaches for enhancing the technological foundations of knowledge acquisition and reasoning in BI applications.

One of the services developed during the project is the Online Self Assessment. By analysing the answers to a questionnaire that describes the economic plan of a company, the tool supplies an evaluation of the quality of the company and of the credit worthiness. The system is based on a predictive model that uses both historical and external knowledge provided by an expert in the domain. The predictive model is implemented by using YaDT-DRb (Bellini, 2007), a variant of the famous Quinlan's C4.5 (Quinlan, 1993) algorithm, modified for using the external knowledge. As usual for this kind of algorithms, the historical data are used for constructing and training the classification models. The external knowledge instead, is new data-independent knowledge provided by an expert and used for integrating the training set and for driving the construction of the models. This technique is documented in our previous work (Baglioni et al., 2005; Baglioni et al., 2008). The new information is provided in form of if-then rules that we call Expert Rules (ERs).

In the project, data and metadata are described and stored by using a set of ontologies.

Starting from this scenario, the extraction of IRs out of an ontology is applied to the MUSING ontology (in particular to the subset of ontology that describes the qualitative questionnaire), and the IRs are used to enrich the set of Expert Rules (ERs) provided by an expert in economics.

Below the details and the results of the IR extraction procedure are given.

Knowledge Repository - Data and metadata reside in the MUSING ontologies. The questionnaire adopted in the Online Self Assessment service is described by the so called BPA ontology. A fragment of the integrated ontologies is depicted in Figure 4. The concepts that belong to upper or related ontologies are labelled with the corresponding prefix (i.e. *psys*, *ptop* or *company*), while for the concepts that belong to the BPA ontology the prefix is missing for saving space. The black continuous arrows represent the *isA* relationships, while the blue broken-line arrows represent the object properties. Not all the relationships nor the object properties and labels have been drawn for the picture clarity sake.

The Data - The dataset used to train and test the models has been provided by the Italian bank Monte dei Paschi di Siena (MPS). The data set, composed of 6000 records contains the following information:

- 13 Qualitative Variables representing a subset of the questions included in the Qualitative Questionnaire performed by MPS to assess the credit worthiness of a third party, and in particular utilised to calculate the Qualitative Score of a Company.
- The Qualitative Score (target item of the classification task).
- 80 Financial/Economic indicators calculated from the Balance Sheets and representing a part of the information utilised to evaluate the probability of the default of a company.

Extraction of the Relevant Concepts - The HITSxONTO algorithm has been applied to the MUS-

ING ontologies yielding a list of 552 ranked concepts. The computation ends after four iterations, returning a list of 5 concepts with hub score greater than 0 and a list of 14 concepts with authority score greater than 0. This is because the ontology is large and not strongly connected.

Construction of the IRs Schemas - Considering all the concepts in the lists as candidates, we obtain 2097 IRs Schemas with exactly one implicant and one implicated.

Characterization of the IRs - After a suitable filtering procedure we apply PATTERNIST to a set of 5757 instances of questionnaires. Having set the minimum support to 20%, PATTERNIST returs a set of 56 frequent itemsets (pairs of concepts). The result of the characterization of the IRs Schemas by using the set of frequent itemsets, is the following set of 14 IRs:

1. `ResearchAndDevelopment.isACompanyInvestment=1` $\xrightarrow{26\%}$ `PreviousAchievements.hasPrevAchievements=1.`

2. `ResearchAndDevelopment.isACompanyInvestment=1` $\xrightarrow{30\%}$ `CapitalizationStrategy.isTheIncreasingForeseen=2.`

3. `ResearchAndDevelopment.isACompanyInvestment=2` $\xrightarrow{28\%}$ `PreviousAchievements.hasPrevAchievements=2.`

4. `StrategicVisionAndQualityManagement.hasRate=2` $\xrightarrow{28\%}$ `CapitalizationStrategy.isTheIncreasingForeseen=2.`

5. `CapitalizationStrategy.isTheIncreasingForeseen=2` $\xrightarrow{36\%}$ `PreviousAchievements.hasPrevAchievements=2.`

6. `ManagementTeam.hasYearOfExperience=1` $\xrightarrow{32\%}$ `PreviousAchievements.hasPrevAchievements=2.`

7. `ResearchAndDevelopment.isACompanyInvestment=2` $\xrightarrow{31\%}$ `PreviousAchievements.hasPrevAchievements=1.`

8. `StrategicVisionAndQualityManagement.hasRate=2` $\xrightarrow{42\%}$ `PreviousAchievements.hasPrevAchievements=1.`

9. `CapitalizationStrategy.isTheIncreasingForeseen=2` $\xrightarrow{48\%}$ `PreviousAchievements.hasPrevAchievements=1.`

10. `ManagementTeam.hasYearOfExperience=1` $\xrightarrow{54\%}$ `PreviousAchievements.hasPrevAchievements=1.`

11. `ResearchAndDevelopment.isACompanyInvestment=2` $\xrightarrow{54\%}$ `CapitalizationStrategy.isTheIncreasingForeseen=2.`

12. `StrategicVisionAndQualityManagement.hasRate=2` $\xrightarrow{60\%}$ `CapitalizationStrategy.isTheIncreasingForeseen=2.`

13. `ManagementTeam.hasYearOfExperience=1` $\xrightarrow{62\%}$ `StrategicVisionAndQualityManagement.hasRate=2.`

14. `ManagementTeam.hasYearOfExperience=1` $\xrightarrow{73\%}$ `CapitalizationStrategy.isTheIncreasingForeseen=2.`

To correctly interpret these rules, please refer to the description of the qualitative questionnaire and its codification, reported in Appendix.
For example, the meaning of the last IR, `ManagementTeam.hasYearOfExperience=1` $\xrightarrow{73\%}$ `CapitalizationStrategy.isTheIncreasingForeseen=2` is:
In 73% *of the cases, if the management team has more than* 10 *years of experience in the industrial sector, then the company does not foresee to increase its capital.*
This IR, in agreement with what we just stated, belongs to the following schema:

`ManagementTeam` $\rightarrow$ `CapitalizationStrategy`

which is one of the 2097 schemas extracted in the previous phase. Here it is clear that the schema provides the structure of a set of future IRs; it defines the direction of the implication and what are the involved concepts. The frequent itemset, instead, identifies the interesting datatype properties (related to the considered concepts) and assigns the weight (i.e. the support), making one of the possible instances compatible with that schema.

6 NEW DEVELOPMENTS

The successful results obtained in the MUSING project and in the economic domain encouraged us to further work on the system and to carry on new experiments. In particular, the extension covers two aspects:

1. The generation of “complex” IRs, i.e. rules composed of more than one implicant item, such as:

   $$I_1, I_2, \ldots, I_n \xrightarrow{w} I_k$$

   where $I_k \notin [I_1, \ldots, I_n]$.

2. The use of a further rule measure: the confidence.

For implementing the first feature we grouped each simple rule with the same consequence, and we construct “super-sets” composed of all the combination of 2, 3, ..., n implicants. Then, we maintain only the sets that, together with the consequence, have a correspondent itemset in the file produced by PATTERNIST. This requirement is necessary to get the right weight to associate to the new complex rule. Then we build the IRs in the traditional way.
The confidence, as usual for association rules,

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

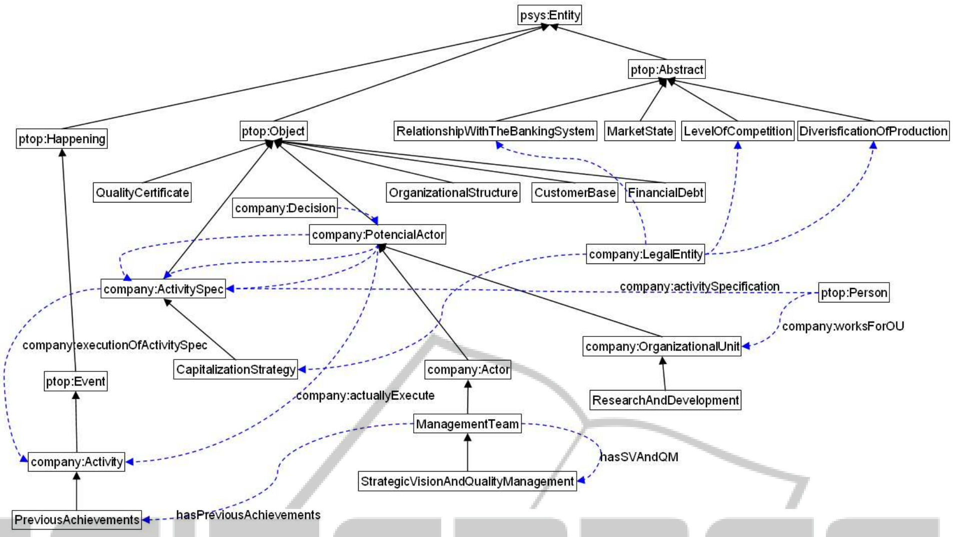


Figure 4: A fragment of the whole Musing ontology.

denotes the conditional probability of the *head* of the rule, given the *body*. This parameter allows to measure the reliability of a rule and in particular of an outlier, i.e. an IR with low probability of occurring.

As an example we report some interesting results computed by using the MUSING data and where we set the minimum support to 1%. For each IRs we associate a short interpretation.

ManagementTeam.hasYearOfExperience=1,
StrategicVisionAndQualityManagement.hasRate=1,
ResearchAndDevelopment.isACompanyInvestment=2 $\xrightarrow{3\%}$
CapitalizationStrategy.isTheIncreasingForeseen=2 (c=92%)

"*In the 3% of the cases, if the years of experience of the management team are more than* 10, *the rate of the strategic vision and quality management is excellent and the company does not invest in R&D, then the company is not foreseeing to increase its capitalization*".

PreviousAchievements.hasPrevAchievements=2,
StrategicVisionAndQualityManagement.hasRate=1,
ResearchAndDevelopment.isACompanyInvestment=2 $\xrightarrow{1\%}$
CapitalizationStrategy.isTheIncreasingForeseen=2 (c=98%)

"*In the 1% of the cases, if the company owner/CEO has no relevant past experiences, the rate of the strategic vision and quality of management is Excellent and the company does not invest in R&D, then the company is not foreseeing to increase its capitalisation.*"

These two IRs have a very low probability but an high confidence, and they can be considered important for an analyst interested in the behavior of a company towards the strategies of management, the investment in the R&D, and the way to finance them.

ManagementTeam.hasYearOfExperience=3 $\xrightarrow{1\%}$
PreviousAchievements.hasPrevAchievements=2 (c=64%)

"*In the 1% of the cases, if the years of experience of the management team are less than 5, then company owner/CEO has no relevant past experiences*".

This is a really rare case, but maybe it should be taken into consideration because of its not negligible confidence value.

7 CONCLUSIONS AND FUTURE WORKS

In this paper we have presented how we handled the problem of extracting interesting and implicit knowledge out of an ontology, presenting the results in form of influence rules. Our idea was to drive the extraction process by using the ontology structure, and to exploit the instances only in a second step. The main problem was to understand if and how to use traditional methods for DM in the context of the ontology. Obviously, the traditional systems can be used only as models, but they are not directly applicable to the ontologies. By decomposing the problem into sub-problems, we succeeded in finding a methodology taking inspiration from consolidated theories and recent developments.

Besides the theoretical results, we had the opportunity of testing our system in an concrete setting ex-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

ploiting our involvement in a European industrial research project: MUSING. In this way, we had at our disposal an integrated framework and a real set of data. Our analysis tool mainly solves, in this domain, the problem of the availability of the expert knowledge. In fact, in the economic field, obtaining a cognitive net of relationships from experts is a hard task, either for the complexity of the matter, or for the lack of specific studies (very often these rules are based on the expert believes or his/her own experience).

A final consideration deals with the application fields and the system extension. In the paper, we focused on the economic domain using the IRs for augmenting a set of "similar" (for meaning, structure and objective) rules. Nevertheless, it is important to point out that the system is fully general and can be used in several domains i.e. in all the domains that can be described by an ontology and where instances are available. Moreover, the new extension further enriches the system, making the IRs much more informative and interesting than before.

ACKNOWLEDGEMENTS

This paper was supported by MUSING Project (IP FP-027097) which provided an useful and convenient framework.

REFERENCES

(2006). Musing Project - http://www.musing.eu/.

Baglioni, M., Bellandi, A., Furletti, B., Spinsanti, L., and Turini, F. (2008). Ontology-based business plan classification. In *EDOC 08: Proceedings of the 2008 12th International IEEE Enterprise Distributed Object Computing Conference*, pages 365–371.

Baglioni, M., Furletti, B., and Turini, F. (2005). Drc4.5: Improving c4.5 by means of prior knowledge. In *SAC 05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 474–481.

Bellandi, A., Furletti, B., Grossi, V., and Romei, A. (2007). Pushing constraints in association rule mining: An ontology-based approach. In *Proceedings of the IADIS International Conference WWW/INTERNET*.

Bellini, L. (2007). Yadt-drb: Yet another decision tree domain rule builder. Masters Thesis.

Bonchi, F., Giannotti, F., Lucchese, C., Orlando, S., Perego, R., and Trasarti, R. (2006). Conquest: a constraint-based querying system for exploratory pattern discovery. In *ICDE*.

Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., and Rojas, I. (2008). Unsupervides learning of semantic relations for molecular biology ontologies. In *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*.

Elsayed, A., El-Beltagy, S. R., Rafea, M., and Hegazy, O. (2007). Applying data mining for ontology building. In *In the proceedings of The 42nd Annual Conference On Statistics, Computer Science, and Operations Research*.

Furletti, B. (2009). Ontology-driven knowledge discovery. Ph.D. Thesis: http://www.di.unipi.it/˜furletti/papers/PhDThesisFurletti2009.pdf.

Geller, J., Zhou, X., Prathipati, K., Kanigiluppai, S., and Chen, X. (2005). Raising data for improved support in rule mining: How to raise and how far to raise. In *Intelligent Data Analysis*, volume 9, pages 397–415.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677.

Parekh, V., Gwo, J., and Finin, T. (2004). Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In *Proceedings of the International Conference of Information and Knowledge Engineering*.

Quinlan, J. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

Singh, S., Vajirkar, P., and Lee, Y. (2003). Context-based data mining using ontologies. In *Conceptual Modeling - ER 2003*, volume 2813.

Vela, M. and Declerck, T. (2008). Heuristics for automated text-based shallow ontology generation. In *Proceedings of the International Semantic Web Conference (Posters & Demos)*.

APPENDIX

The qualitative questionnaire aims at collecting the qualitative information of the company/financial institution that accesses the Online Self Assessment service. Here is the list of questions and the corresponding answers.

For being processed, the questionnaire has been suitable codified. In the ontology, at the schema level, each question is a datatype property of a concept.

The codification, with the syntax `Concept.datatypeProperty`, is also provided.

- **Diversification of Production.**
  1. The company operates in more than one sector.
  2. The company operates in just one sector with flexible production processes.
  3. The company operates in just one sector with no flexible production processes.

```
DiversificationOfProduction.hasDivOfProdValue
```

- **Commercial Diversification.**
  1. Customers base well diversified, with no concentration of sales.

SCIENCE AND TECHNOLOGY PUBLICATIONS

2. Customers base well diversified, with some key clients.
3. Most of sales directed to few key clients.

`CustomerBase.hasDiversification`

- **Years of Experience of the Management Team in the Industrial Sector the Company Operates in.**

1. > 10.
2. $5 - 10$.
3. < 5.

`ManagementTeam.hasYearOfExperience`

- **Previous Achievements of the Management Team.**

1. Company owner/CEO with past successful achievements even in different fields from the one in which the company operates today.
2. Company owner/CEO with no relevant past experiences.
3. Company owner/CEO with one or more unsuccessful past experiences.

`PreviousAchievements.hasPrevAchievements`

- **Strategic Vision and Quality of Management (Referred to Previous Experiences).**

1. Excellent.
2. Good.
3. Satisfying.
4. Insufficient.

`StrategicVisionAndQualityManagement.hasRate`

- **Organisational Structure of the Company.**

1. Organised in a well-articulate and efficient way.
2. Well organised even if some gaps are present, all the relevant positions are well covered.
3. The organisation is not adequate to the company dimension and some relevant positions are not presided.

`OrganizationalStructure.hasType`

- **Market Trend.**

1. Growing.
2. Stable.
3. Going toward stabilization.
4. In recession.

`MarketState.hasTypeOfPhase`

- **Does the Company Invest in Research & Development?**

1. Yes.
2. No.

`ResearchAndDevelopment.isACompanyInvestment`

- **Level of Competition in the Market.**

1. High.
2. Medium.
3. Low.

`LevelOfCompetition.competitionRate`

- **Quality Certificate(s) Achieved.**

1. The company achieved one or more quality certificates.
2. The company has one or more quality certificates requests in progress.
3. The company does not have any quality certificates.

`QualityCertificate.numberOfQCAchieved`

- **Relationships with the Banking System.**

1. Good margin of utilisation of the credit lines and good credit worthiness.
2. Good margin of utilisation of the credit lines.
3. Presence of some tensions.
4. Overdrafts are present.

`RelationshipWithTheBankingSystem.hasTypeOfRelationship`

- **Financial Requirements Trend.**

1. In line with the company dynamics.
2. Not in line with the company dynamics.

`FinancialDebt.hasFinancialDebt`

- **Is the Company Foreseeing to Increase its Capitalisation?**

1. Yes.
2. No.

`CapitalizationStrategy.isTheIncreasingForeseen`

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS