

Matching Directories and OWL Ontologies with AROMA

Jérôme David
LINA FRE 2729 CNRS
Polytechnic School of Nantes
University
44306 Nantes, France
jerome.david@univ-
nantes.fr

Fabrice Guillet
LINA FRE 2729 CNRS
Polytechnic School of Nantes
University
44306 Nantes, France
fabrice.guillet@univ-
nantes.fr

Henri Briand
LINA FRE 2729 CNRS
Polytechnic School of Nantes
University
44306 Nantes, France
henri.briand@univ-
nantes.fr

ABSTRACT

This paper presents a simple and adaptable matching method dealing with web directories, catalogs and OWL ontologies. By using a well-known Knowledge Discovery in Databases model, such as the association rule paradigm, this method has the originality to be both extensional and asymmetric. It works at the terminological level (by selecting concept-relevant terms contained in documents) and permits to discover equivalence and also subsumption relations holding between entities (concepts and properties). This method relies on the implication intensity measure, a probabilistic model of deviation from independence. Selection of significant rules between concepts (or properties) is lead by two criteria permitting to assess respectively the implication quality and the generativity of the rule.

Categories and Subject Descriptors: H.2.5 Database Management: Heterogeneous Databases

General Terms: Measurement, Algorithms.

Keywords: Ontology Matching, Association Rules, Data-mining, Semantic Web.

1. INTRODUCTION

Schema or ontology matching approaches aims at finding semantic relations (i.e. equivalence, subsumption, etc.) between entities (i.e. concepts, properties) issued from two schemas or ontologies. In the literature, a lot of works deals with schema or ontology matching ([4], [5]). The proposed approaches are issued from different communities but they mostly use similarity measures for discovering equivalence relations between concepts or properties. However, the matching process could be enhanced by using **asymmetric measures** which also permit to discover subsumption relations between ontologies entities. In parallel, knowledge discovery in databases (KDD) [3] widely uses the asymmetric measures, called interestingness measures, for association rule discovery. Association rules [1] are expressions of the type "if *antecedent* then *consequent*" representing implicative tendencies between conjunctions of attributes in databases.

In this paper, we address ontology and directory matching by using the association rule paradigm. We propose a matching method using an asymmetric measure: the Impli-

cation Intensity [2], a probabilistic model of deviation from statistical independence. Our approach, named AROMA (Association Rule Ontology Matching Approach), has the advantage to be simple because it does not use heavy machine learning technique. AROMA has been firstly designed as an **extensional and asymmetric method** for matching conceptual taxonomies populated with textual documents (examples of such structures are web directories or catalogs). In this paper, we also show the adaptability of AROMA by using it for matching OWL ontologies.

2. THE AROMA METHODOLOGY

The original AROMA method is divided into two main parts: (1) the extraction and selection of relevant terms for each concept; (2) the discovery of significant implications between the two hierarchies. In this section, we also describe an adaptation of the first part in order to deal with OWL ontologies.

2.1 Acquisition and selection of relevant terms

The first part consists to associate a set of relevant terms for each concept of a hierarchy. These terms are extracted from documents indexed to concepts and selected by evaluating association rules $t \rightarrow c$. Such a rule means: "if a document contains the term t then this document is associated with the concept c ". After this first pre-processing part, a conceptual hierarchy is defined as a tuple $\mathcal{H}' = (C, \leq, T, \gamma)$ where C is the set of concepts, \leq is the partial order organising concepts into a taxonomy, T is the set of relevant terms selected during the previous step and γ is the relation linking concept to its relevant terms set (i.e. $\gamma(c)$ represents the relevant terms associated to the concept c). In order to transpose the partial order between concepts to their relevant terms sets, we extend γ to the relation γ' as follows: $\gamma'(c) = \bigcup_{c' \leq c} \gamma(c')$.

2.2 Association rules discovery between hierarchies

The second stage of our method consists in the discovery of implicative matching relations between concepts by evaluating association rules between their respective relevant terms sets. The algorithm takes in two pre-processed hierarchies \mathcal{H}'_1 and \mathcal{H}'_2 and considers only the terms shared by the two structures. The common terms set of the two hierarchies \mathcal{H}'_1 and \mathcal{H}'_2 is noted $T_{1 \cap 2} = T_1 \cap T_2$. Next, we define the relation $\gamma'_{1 \cap 2}$ which associates a subset of $T_{1 \cap 2}$ for

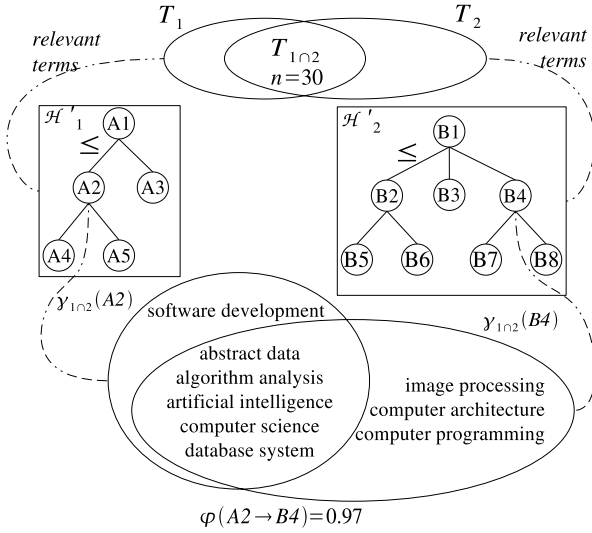


Figure 1: Evaluation of rules

each concept $c \in C_1 \cup C_2$: $\gamma'_{1 \cap 2}(c) = \begin{cases} \gamma'_1(c) \cap T_2 & \text{if } c \in C_1 \\ \gamma'_2(c) \cap T_1 & \text{if } c \in C_2 \end{cases}$

The algorithm provides a top-down search of association rules and uses two criteria for select significant rules. A rule $a \rightarrow b$ (between the concepts $a \in C_1$ and $b \in C_2$) will be significant if it respects the two following criteria:

$$\varphi(a \rightarrow b) \geq \varphi_r \quad (1)$$

$$\forall x \geq a, \forall y \leq b, \varphi(x \rightarrow y) \leq \varphi(a \rightarrow b) \quad (2)$$

The first criterion (equation 1) guarantees the **quality** of the implication tendency between the two concepts for a given threshold φ_r . The implication intensity of the rule $a \rightarrow b$ is defined as follows: $\varphi(a \rightarrow b) = 1 - Pr(N_{a \wedge \bar{b}} \leq n_{a \wedge \bar{b}})$ where $n_{a \wedge \bar{b}} = \text{card}(\gamma_{1 \cap 2}(a) - \gamma_{1 \cap 2}(b))$ is the number of relevant terms for concept a which are not relevant for concept b . $N_{a \wedge \bar{b}}$ is the expected number of relevant terms for concept a which are not relevant for concept b .

For example (figure 1), the rule $A2 \rightarrow B4$ has $n_{A2 \wedge \bar{B4}} = 1$ counter-examples. Its implication intensity value is calculated using an Poisson law (which is a possible model for the implication intensity):

$$\varphi(A2 \rightarrow B4) = \sum_{k=0}^{n_{A2 \wedge \bar{B4}}} e^{-\lambda} \cdot \frac{\lambda^k}{k!} = 0,97$$

where $\lambda = n_{A2} \cdot n_{\bar{B4}} / n = 6 \cdot (30 - 8) / 30$.

The second criterion (equation 2) verifies the **generativity** of the rule and then permits to reduce redundancy in the extracted rule set. Indeed, a valid rule (i.e. a rule satisfying the first criterion) is significant if it does not exist a more generative rule having an implication intensity value greater than or equals to it. A rule $x \rightarrow y$ is more generative than a rule $u \rightarrow v$ if $u \leq x$ and $y \leq v$ (with $x \rightarrow y \neq u \rightarrow v$). For example (figure 1), the rules $A2 \rightarrow B7$, $A2 \rightarrow B8$, $A1 \rightarrow B4$, $A1 \rightarrow B7$, and $A1 \rightarrow B8$ are more generative than the studied rule $A2 \rightarrow B4$. The rule $A2 \rightarrow B4$ will be significant and then selected if none of its generative rules have a φ value greater than or equals to its φ value.

2.3 Adaptation for OWL ontology language

We propose an adaptation of the AROMA method for discovering implicative mapping between two OWL ontologies. In this case, we also match the hierarchies of properties. Contrarily to the web directories or catalogs, we do not only consider the extension of the ontology. Indeed, on the one hand, OWL ontologies are not often populated with a lot of individuals (or instances) and on the other hand, they contain very few textual data. Thus, we suggest to use information contained both in the schema and in the individuals data.

Contrarily to the original approach, this evolution does not only represent classes (and properties) by binary terms. The schema information extracted for a class or a property data set are their RDF id attribute, RDFS label attribute and the binary terms contained in the RDFS comment attribute. In addition, we also extract the RDFS label attributes and the values of the OWL "Datatype Properties" taken by their individuals. In case of properties, we use the extension of values taken by their range.

After have represented classes and properties by datasets, the method evaluates and selects the set of rules holding between classes and properties issued from the two OWL ontologies. The following step is the same that those described in section 2.2: we keep the taxonomy information by introducing the inclusion of the classes or properties datasets into their parents datasets, the algorithm also works on the data shared by the two structures and, the same rules selection criteria are used.

3. CONCLUSION

In this paper, we briefly presented AROMA, a conceptual hierarchies matching method, and its adaptation to OWL ontology matching. This method have the originality to be both extensional and asymmetric. Then, the main advantages of this method are its simplicity (only association rules extraction, no combination of machine learning strategies), the consideration of semantic by using binary terms contained in the corpus and the stronger semantic offered by association rules regarding only similarity-based matching systems.

4. REFERENCES

- [1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *the 1993 ACM SIGMOD int. conf. on Management of data*, pages 207–216. ACM Press, 1993.
- [2] J. Blanchard, P. Kuntz, F. Guillet, and R. Gras. *Implication intensity: from the basic statistical definition to the entropic version*, chapter 28, pages 473–485. CRC Press, 2003.
- [3] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- [4] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.
- [5] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics IV*, 4(LNCS 3730):146–171, 2005.