Comparison of interestingness measures applied to textual taxonomies matching

Jérôme David[1], Fabrice Guillet[1], Régis Gras[1], and Henri Briand[1]

Polytechnic School of Nantes University, LINA FRE CNRS 2729
44306 NANTES cedex 3, France
(e-mail: jerome.david-at-univ-nantes.fr)

Abstract. This paper presents an experimental comparison of Interestingness Measures (IMs), in the context of an approach designed for matching textual taxonomies. This extensional and asymmetric approach makes use of association rule model for matching entities issued from two textual hierarchies. We select 6 IMs and we perform two experiments on a benchmark composed of two textual taxonomies and a set of reference matching relations between the concepts of the two structures. The first test concerns a comparison of matching accuracy with each of the selected measures. In the second experiment, we compare how each IM evaluates reference relations by studying their values distributions. Results show that the implication intensity delivers the best results.
Keywords: association rule, interestingness measures, textual taxonomy matching.

1 Introduction

In the context of databases, the association rule mining [Agrawal *et al.*, 1993] is a well-known approach for discovering knowledge nuggets. Association rules are propositions, noted *antecedent* $\rightarrow$ *consequent*, representing implicative tendencies between conjunctions of valued attributes or items. Association rules have the advantage of being an easy and meaningful model for representing explicit knowledge. This advantage has motivated a lot of research works and the publication of association rule extraction algorithms [Ceglar and Roddick, 2006] and Interestingness Measures (IMs) which aim at assessing the implicative quality of rules.

In the context of the WWW (and the Semantic Web), the matching of hierarchical structures (web-directories, OWL ontologies, online shop catalogs etc.) is a major concern. Thus, many matching methods have been proposed in the literature [Rahm and Bernstein, 2001], [Shvaiko and Euzenat, 2005]. These methods aim at finding semantic relations (i.e. equivalence, subsumption, etc) between entities (i.e. categories, concepts, properties) issued from two hierarchical structures. The proposed approaches mostly use similarity measures and as a consequence, a majority of them are restricted to finding equivalence relations only.

At the intersection of these two research fields, we proposed to use the association rule paradigm for matching hierarchies [David *et al.*, 2006]. Our approach, named AROMA (Association Rule Ontology Matching Approach), heavily relies on the asymmetric nature of association rules and then helps to characterise more precisely the matching relations between hierarchical structures regarding only similarity based approaches. Furthermore, contrary to most approaches designed for matching schema or ontologies, AROMA works with the extensional data provided with structures. In this paper, we propose to evaluate IMs in this context of textual taxonomy matching.

This paper is organised as follows: first, we present related works about ontology/schema matching. Then, we present our selection of 6 IMs. In the third section, we present AROMA methodology . The last section is dedicated to two experiments made on two catalogs and a set of reference matching relations. The first experiment evaluates the selected IMs thought f-measure curves. In the second experiment, we compare the values distributions of IMs obtained on a set of reference matching relations.

2 Textual taxonomies matching

For the last seven years, many ontology and schema matching approaches have been proposed in the literature. The survey [Rahm and Bernstein, 2001] propose a classification of matching methods which discriminates the *extensional or element-based approaches* from the *intensional or only-schema-based approaches*. While many efforts are concentrated around the intensional matchers ([Shvaiko and Euzenat, 2005] proposes a detailed survey), few extensional approaches have been proposed. We focus on extensional matchers since AROMA is one of them. Such a method relies on extensional content because they are designed for matching textual taxonomies which have typically poor schema information but a lot of textual data.

Textual taxonomies are often provided with different sets of documents and then it is difficult to compare them. Thus, before computing similarities, the methods pre-process the extensional data. According to the type of preprocessing step they use, we retain two main classes of techniques:

1. **document level methods** which compare shared documents between entities. These methods use machine learning for classifying documents from each structure into the other one so that they share the same set of documents. GLUE [Doan *et al.*, 2004], Hical [Ichise *et al.*, 2004], and oPLMap [Nottelmann and Straccia, 2005] are examples of such approaches.
2. **term level methods** which represent entities by a set of terms. They use linguistic processing and information retrieval techniques in order to extract terms and select and/or weight them in function of their relevance to studied entities. Examples of such methods are V-DOC [Qu *et al.*, 2006], CAIMAN [Lacher and Groh, 2001] and AROMA.

After the pre-processing step, methods can compare two concepts by using their extensions. GLUE uses Jaccard similarity (but it can use other one), oPLMap uses conditional probability and Hical makes use of a statistical test ($k - statistic$). Terms level methods tend to represent concepts by vectors where dimensions are terms and values represents weights computed according a measure such TF/IDF (in V-DOC and CAIMAN). Such methods use cosine measure for comparing the vectors representing concepts.

Concerning the type of matching relation, only AROMA considers the implication. The others are restricted to the equivalence. Concerning the measures used, 4 methods use similarities while only oPLMap and AROMA use asymmetric measures. Nevertheless, oPLMap does not seems to use it in order to find implication between concepts.

3 Interestingness Measures

In the framework of association rule discovery, and in order to select the most interesting rules, many Interestingness Measures (IMs) have been proposed and studied (see [Bayardo Jr. and Agrawal, 1999], [Tan *et al.*, 2004] for a survey).

According to our objective of hierarchy matching, the 6 selected IMs respect the principle of maximal value and have been chosen following two criteria proposed in [Blanchard, 2005]. For each nature of IM (i.e. statistical (S) or descriptive (D)), we select a "rule IM" ($\rightarrow$), a "quasi-implication IM" ($\Rightarrow$) and a "quasi-conjunction IM" ($\leftrightarrow$). The selected "rule IMs" assess the deviation from equilibrium situation (E). Those having the quasi-implication as scope assess the deviation from independence situation (I). For the quasi-conjunction IMs, one has the deviation from equilibrium as subject, the other has the deviation from independence as subject. For each select IM, Table 1 also shows its fixed value taken at the independence or equilibrium situation and its formula.

Measure	Scope	Nature	Subject	Fixed value	Formula
Implication Intensity (II) [Gras et al., 1996]	$\Rightarrow$	S	I	0.5	$P(n_{a\overline{b}} < Poisson(\frac{n_a.n_{\overline{b}}}{n}))$
Loevinger [Loevinger, 1947]	$\Rightarrow$	D	I	0	$1 - \frac{n.n_{a\overline{b}}}{n_a.n_{\overline{b}}}$
Probabilistic index of deviation from equilibrium (IPEE) [Blanchard, 2005]	$\rightarrow$	S	E	0.5	$P(n_{a\overline{b}} < Binomial(n_a, 1/2))$
Confidence [Agrawal *et al.*, 1993]	$\rightarrow$	D	E	0.5	n_{ab}/n_b
Likelyhood linkage analysis index (LLA) [Lerman, 1981]	$\leftrightarrow$	S	I	0.5	$P(n_{ab} > Poisson(\frac{n_a.n_b}{n}))$
Jaccard	$\leftrightarrow$	D	E	0, 5	$n_{ab}/(n_a + n_b - n_{ab})$

Table 1. Selected IMs and their properties

4 AROMA methodology

AROMA (Figure 1) was designed to discover a set of significant association rules holding between concepts issued from conceptual hierarchies populated of textual documents. AROMA takes, in input, two conceptual hierarchies $\mathcal{H}_1$ and $\mathcal{H}_2$ each defined as a tuple $\mathcal{H} = (C, \leq, D, \sigma)$ where C is the set of concepts, $\leq$ is the partial order organising concepts into a taxonomy, D is the set of textual documents, and σ is the relation associating a set of documents to each concept (i.e. for a concept $c \in C$, $\sigma(c)$ represents the documents associated to c). Thanks to the first stage concerning the acquisition and the selection of relevant terms for each concept, we redefine each hierarchy as a tuple $\mathcal{H}' = (C, \leq, T, \gamma)$ where T is the set of relevant terms selected. In order to consider the partial order, we assume that a term associated to a concept is also associated with the parent concepts, then we extend γ to γ' as follows: $\gamma'(c) = \bigcup_{c' \leq c} \gamma(c')$.

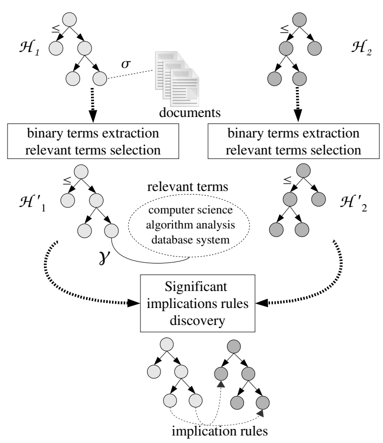

Fig. 1. The AROMA approach

The second stage of AROMA consists in the discovery of implicative matching relations between concepts by evaluating association rule 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, noted $T_{1\cap 2} = T_1 \cap T_2$. The relation $\gamma'_{1\cap 2}$ associates a subset of $T_{1\cap 2}$ for each concept $c \in C_1 \cup C_2$: $\gamma'_{1\cap 2}(c) = \begin{cases} \gamma'_1(c) \cap T_2 \; if \; c \in C_1 \\ \gamma'_2(c) \cap T_1 \; if \; c \in C_2 \end{cases}$

A valid rule $a \rightarrow b$ represents a quasi-implication from the relevant terms set of the concept a into the relevant terms set of the concept b. Such a rule means that the concept $a \in C_1$ is probably more specific than or equivalent to the concept $b \in C_2$.

We use two criteria for selecting significant rules and reduce redundancy. Then, 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: (1) $\varphi(a \rightarrow b) \leq \varphi_r$; (2) $\forall x \geq a$, $\forall y \leq b$, $\varphi(x \rightarrow y) \leq \varphi(a \rightarrow b)$

The first criterion guarantees the **quality** of the implication tendency between the two concepts for a given threshold φ_r. The second criterion verifies the **generativity** of the rule and then permits to reduce redundancy in the extracted rules set. Indeeds, 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 $x \rightarrow y \neq u \rightarrow v$).

5 Experimental results

The experiments presented in this section concern only the second part of the AROMA (i.e. the rule selection phase). In a first time, we compare measures with the help of f-measure curves. And next, we perform an analysis of distribution of measures values on a set of reference matching relations.

5.1 Analysed data

The studied benchmark, "Course catalog" [Doan *et al.*, 2004], is composed of two catalogs of courses descriptions which are proposed at the Cornell and Washington universities. The courses descriptions are hierarchically organised into schools and colleges and then into departments and centers within each college. These two hierarchies contain respectively 166 and 176 concepts to which are associated 4360 and 6957 textual course descriptions. The benchmark data are also provided with a set of 54 manual symmetric matching relations from concepts of Cornell catalog to Washington catalog.

5.2 Evaluation of IMs

Here, we address the evaluation obtained by AROMA algorithms with each selected IM. We made vary the rule selection threshold φ_r from 0 to 1. Then for each φ_r value, the rule selection algorithms are executed twice. The first time for selecting implications from the Cornell concepts to the Washington concepts and the second time for selecting implications from Washington to Cornell. From the two implicative matching sets, we retain only equivalence relations by following this rule: if $a \rightarrow b$ and $b \rightarrow a$, then $a \leftrightarrow b$.

In order to evaluate the relevance of results, we use the **f-measure** defined as follows: $f - measure = 2.precision.recall/precision + recall$, where $precision = card(F \cap R)/card(F)$ and $recall = card(F \cap R)/card(R)$; F is the set of matching pairs found using AROMA and R is the set of "reference" matching pairs.

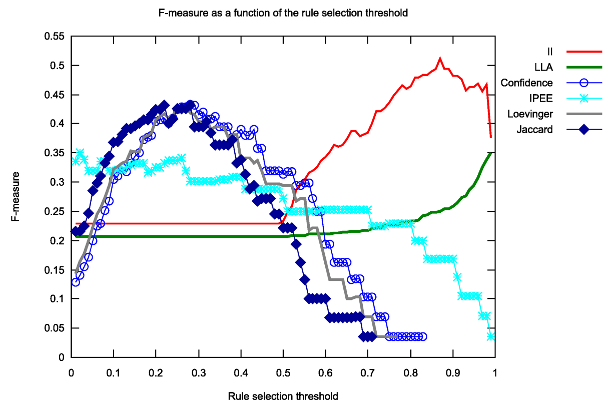

Fig. 2. Evolution of F-measure

On Figure 2, the descriptive measures (Loevinger, Jaccard and confidence) have very similar curves which increase (until a maximum f-measure value near 0.45 for $\varphi_r \simeq 0.25$) and then rapidly decrease. The curve of Ipee steadily decrease from around 0.35 to a value near 0. The curves of LLA and II are constant before $\varphi_r = 0.5$ and then increase until respectively 0.35 and 0.51. The best f-measure values are obtained with II.

Deviation from equilibrium IMs (Jaccard, confidence and Ipee) obtain their best results before the equilibrium situation be reached (when $\varphi_r = 0.5$ for these IMs). Deviation from independence IMs (II, LLA, and Loevinger) obtain their best results after the independence situation (when $\varphi_r = 0$ for Loevinger, and when $\varphi_r = 0.5$ for II and LLA). As a consequence, this last IMs family seems to be more relevant for AROMA.

5.3 Distributions of IMs

In this experiment, we would study and compare how measures evaluate matching relations independently of the AROMA rule selection algorithms and their selection criteria. We propose to draw the values distributions of IMs obtained on the set of reference matching relation R.

For each manual matching relation represented by a couple (A, B), we evaluated the rules $A \rightarrow B$ and $B \rightarrow A$. For each couple, we keep only the best value, in the cases where the studied measure is not symmetric.

On Figure 3, distributions of confidence and Loevinger look very similar: they evaluate few relations under 0.5, a majority in the interval $[0.5-0.9]$ and 11 relations with 1. The two statistical IMs of deviation from independence (II and LLA) assess the main part of relations with 1 (from 25 up to 35). Nevertheless, the values given by II are more disseminated than those of LLA.

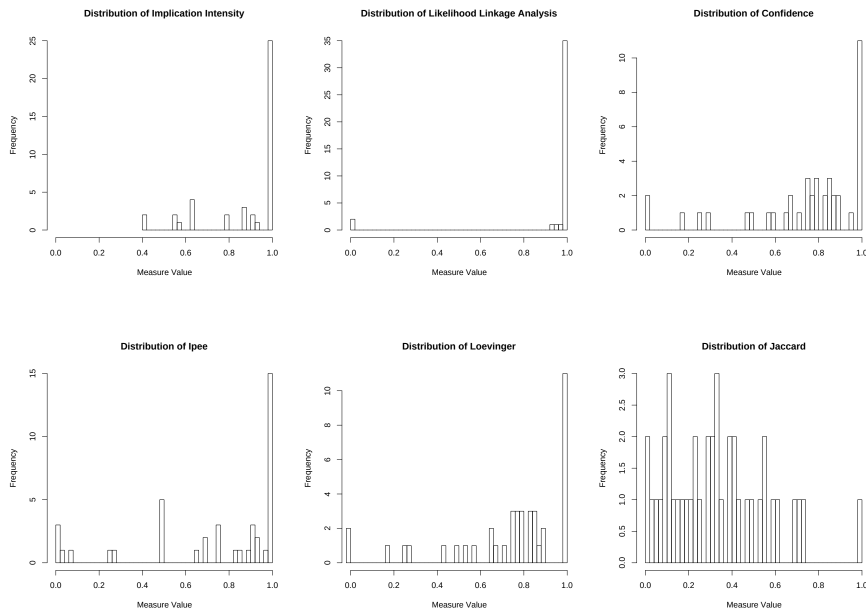

Fig. 3. Measures distributions on manual matching relations R

Even if Ipee also values a lot of relations at 1, it assess more relations than the other statistical IMs with values under 0.5. The distribution of Jaccard presents a lot of relations evaluated with bad values (< 0.5).

IMs of deviation of independence evaluate only 2 relations under their respective fixed value while IMs of deviation from equilibrium (except Jaccard) evaluate 7 relations under the value taken at the equilibrium situation.

6 Conclusion

In this paper, we proposed an original use of association rule model and IMs in a context of schema/ontology matching. More precisely, we described the approach AROMA, which is an extensional matcher designed to be used on hierarchical structures provided with text documents. AROMA has the originality to make use of the asymmetrical aspect of association rules in order to discover subsumption matches between hierarchies or ontologies.

We selected 6 IMs according to several criteria (subject and scope) and we evaluated them on a matching benchmark. Results show that f-measure curves of descriptive measures are similar. Ipee and LLA indexes are not relevant for a such application. Implication intensity obtains the best results on the benchmark. The distributions shows that relevant matching relations are better evaluated with deviation from independence measures.

References

[Agrawal *et al.*, 1993]R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *the proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data*, pages 207–216. ACM Press, 1993.

[Bayardo Jr. and Agrawal, 1999]Roberto J. Bayardo Jr. and Rakesh Agrawal. Mining the most interestingness rules. *Proc. of the 5th ACM SIGKDD Int. Conf. On Knowledge Discovery and Data Mining*, pages 145–154, 1999.

[Blanchard, 2005]J. Blanchard. *A visualization system for interactive mining, assessment, and exploration of association rules.* PhD thesis, University of Nantes, 2005.

[Ceglar and Roddick, 2006]A. Ceglar and J. F. Roddick. Association mining. *ACM Comput. Surv.*, 38(2):5, 2006.

[David *et al.*, 2006]J. David, F. Guillet, and H. Briand. Matching directories and owl ontologies with AROMA. In *the proc. of the 15th ACM int. conf. on Inf. and knowledge management, CIKM'06*, pages 830–831. ACM Press, 2006.

[Doan *et al.*, 2004]A. Doan, J. Madhavan, P. Domingos, and A. Halevy. *Ontology Matching : a machine learning approach*, pages 397–416. Springer-Velag, 2004.

[Gras et al., 1996]R. Gras et al. *L'implication statistique, une nouvelle méthode exploratoire de données.* La pensée sauvage, 1996.

[Ichise *et al.*, 2004]R. Ichise, M.Hamasaki, and H. Takeda. Discovering relationships among catalogs. In E. Suzuki and S. Arikawa, editors, *7th Int. Conf. Discovery Science 2004*, volume 3245 of *LNCS*, pages 371–379. Springer, 2004.

[Lacher and Groh, 2001]M. S. Lacher and G. Groh. Facilitating the exchange of explicit knowledge through ontology mappings. In *the proc. of 14th Int. Florida A.I. Research Society Conf.*, pages 305–309. AAAI Press, 2001.

[Lerman, 1981]I.C. Lerman. *Classification et analyse ordinale des données.* Dunod, 1981.

[Loevinger, 1947]J. Loevinger. A systemic approach to the construction and evaluation of tests of ability. *Psychol. Monographs*, 61:4, 1947.

[Nottelmann and Straccia, 2005]H. Nottelmann and U. Straccia. A probabilistic, logic-based framework for automated web directory alignment. In Zongmin Ma, editor, *Soft Computing in Ontologies and the Semantic Web*, Studies in Fuzziness and Soft Computing. Springer Verlag, 2005.

[Qu *et al.*, 2006]Y. Qu, W. Hu, and G. Cheng. Constructing virtual documents for ontology matching. In *the proc. of the 15th int. conf. on WWW*, pages 23–31, New York, NY, USA, 2006. ACM Press.

[Rahm and Bernstein, 2001]E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[Shvaiko and Euzenat, 2005]P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics*, 4:146–171, 2005.

[Tan *et al.*, 2004]Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.