Evaluation of hierarchical interestingness measures for mining pairwise generalized association rules

Fernando Benites and Elena Sapozhnikova

***Abstract*—In the literature about association analysis, many interestingness measures have been proposed to assess the quality of obtained association rules in order to select a small set of the most interesting among them. In the particular case of hierarchically organized items and generalized association rules connecting them, a measure that dealt appropriately with the hierarchy would be advantageous. Here we present the further developments of a new class of such hierarchical interestingness measures and compare them with a large set of conventional measures and with three hierarchical pruning methods from the literature. The aim is to find interesting pairwise generalized association rules connecting the concepts of multiple ontologies. Interested in the broad empirical evaluation of interestingness measures, we compared the rules obtained by 39 methods on three real world datasets against predefined ground truth sets of associations. To this end, we adopted a framework of instance-based ontology matching and extended the set of performance measures by two novel measures: relation learning recall and precision which take into account hierarchical relationships between rules.**

***Index Terms*—Data Mining, Association Rules, Interestingness Measures, Ontology Matching**

I. Introduction

This paper discusses a task of mining pairwise associations between the concepts of multiple ontologies. An ontology [1] formally specifies the concepts (usually hierarchically organized) and their relations within a domain. The concepts of ontologies are often used to classify or annotate objects. An ontology O is a structure $O = (Tm, C, Rel, G)$, where Terms Tm usually cover natural language aspects and are assigned to concepts C and relations Rel. Examples for concepts are “biological process” or “animal cell” and for relations “to transport sugar” or “positively regulates”. The relations connect concepts producing a labelled graph structure G over these concepts. More specifically, we focus on the underlying taxonomies which contain concepts connected only by the is-a relation. They are also often referred to as concept hierarchies [1], we will utilize both terms interchangeably.

The motivation behind this task is that associations connecting different ontologies can be very helpful in many applications varying from ontology mapping to multi-label classification [2]. In the latter case, the data to be classified often share a common representation in the input space, but possess multiple class taxonomies in the output space. For example, a movie can be classified simultaneously either by its genre in a genre taxonomy or by the producing company in a taxonomy of producers. A possible association between the taxonomies could involve a specific company being specialized in a certain genre of movies, e.g. science fiction. Discovering such implicit relations between class taxonomies may support experts in extracting new knowledge from data, on the one hand, and facilitate understanding of obtained classification results, on the other.

Association rule mining [3] is one of the methods which can be used to solve the considered task. It was initially applied to the market basket analysis in order to find related products, e.g. present in the same transaction. A transaction is a subset of items, for example, items purchased together. An Association Rule (AR) has the form $X \rightarrow Y$, where X and Y are disjoint sets of items (X is called the antecedent and Y the consequent). To select a small set of the most interesting ARs, the so-called Interestingness Measures (IMs) are used. Support and confidence are most commonly used IMs in AR mining. If the items are hierarchically organized, Generalized Association Rules (GARs) [4] can be found across different levels of a hierarchy. For example, a rule “Shirt”→“Shoes” would be more specific than the rule “Clothes”→“Shoes”. In our study, items correspond to the concepts of ontologies and transactions – to the objects annotated by them.

However, mining associations between ontological concepts does not completely match the standard setting of the association analysis that assumes frequent itemset mining, i.e. only associations between the most frequent itemsets can be established. This is due to the fact that in the market basket analysis it is reasonable to pay attention only to the products that are frequently bought together. In contrast, while connecting ontological concepts, we are more interested in finding rare associations that affect only a small subset of data rather than frequent ones which correspond to high-level concepts and well-known facts. Furthermore, appropriate handling rare cases [5] is especially useful when the standard association analysis would eliminate infrequent but nevertheless interesting relations, e.g. in biology where new knowledge could be gathered from seldom combinations of gene functions emerging together in only a few proteins [6]. Therefore, mining rare ARs has recently attracted more attention [7], [8].

As the standard AR mining cannot be directly applied to our task, we propose an alternative approach. It uses a new class of hierarchical IMs for assessing the quality of rules with respect to (w.r.t.) the hierarchy [2], [9]. Since the redundancy of GARs is caused to a large extent by the hierarchical structure itself, it can be used by such IMs for penalizing redundant rules. The idea is to take into account a given standard measure and its respective hierarchical expectation. In this paper, we further develop the class of hierarchical measures and compare them with a large number of conventional IMs proposed in the literature as well as with three hierarchical pruning methods.

This wide experimental comparison is the most important contribution of our study. To the best of our knowledge, no comparative analysis of IMs has been performed for GARs yet. In this study, we will close this gap by comparing the developed hierarchical IMs with a large set of 29 conventional IMs. Among this set, the most interesting for comparison may be those which were shown to be well suited for mining rare ARs by [8]: *Cosine*, *AllConfidence* and *Jaccard*. As the leaf nodes of a concept hierarchy can usually be seen as rare items, the hypothesis that these three measures are also best suited for finding non-redundant associations between multiple ontologies should be verified experimentally.

It is often difficult to compare IMs because usually only few experimental results are available in the literature. An additional difficulty is that it is typically not known in advance which associations should be discovered. One possibility is to compare the rules obtained by different measures with those manually extracted by a human expert. However, this approach is restricted by the great effort that is involved. As a result, manually extracted rules are rarely available. Alternatively, the best measure can be found by comparing how well each measure agrees with the expectation of an expert when analyzing a small subset of ARs [10]. Motivated by our task of connecting multiple ontologies, we decided to adopt the evaluation framework of instance-based ontology matching for the experimental comparison of IMs. It assumes that multiple ontologies should be aligned by connecting similar concepts using classified data and the quality of this alignment is then assessed by the comparison with a manually created set of correct connections (ground truth associations, true rule set). Thus, the number of discovered true associations can serve as an indicator of quality for IMs. Obviously, it is restricted to the case of balanced associations (with similar distributions of the antecedent and the consequent in an association) because we assume that the ontologies are similar. In general, it will not necessarily hold. For the case of unbalanced associations please refer to [11].

Another contribution of this work is the development of two new performance measures: Relation Learning Recall and Relation Learning Precision by improving the method of [12]. The reason is that the standard performance measures precision and recall are insufficient to deal with hierarchically structured rules because they cannot take into account partial matches between discovered and the true rules, e.g. if a more general concept was found instead of a special one.

The rest of the paper is organized as follows: Section II gives a brief overview of related work. Section III introduces our approach along with the new class of hierarchical IMs and shows the list of IMs used for comparison. Section IV describes the datasets, experimental settings and results. In Section V, the discussion of the results is presented and Section VI concludes the paper.

II. Related Work

An overview of related work stemming from different fields is summarized in Table I. Considering a general research on IMs, there are many studies of their properties [10], [11], [13], [14]. However, they do not treat the special case of GARs and hierarchical measures.

Table I: List of related works organized by field

Field	Studies
Research on IMs	[10], [11], [13] and [14]
Semantic Data Mining	[15], [16] and [17]
FCA	[18] and [19]
Non-taxonomic relations	[12] and [20]
Ontology Matching	AROMA [1], HICAL [21], GLUE [22], oPLMap [23] and [24]
Hierarchical pruning methods	AROMA [1], GRP [4], GCC [25]
Hierarchical Measures	[2], [9] and [26]

In the field of semantic data mining, we find our approach related to the use of ontologies as background knowledge. For example, ontologies are used in [15] for discovering semantically richer ARs and in [16] for learning more general rules. However, the taxonomies in these cases are merely tools for improving the quality of obtained rules and not the subject of mapping. It should also be noted that a recent work [17] presents an AR-based approach to the generation of ontologies from Resource Description Framework (RDF) repositories. It can also be seen as a related research area.

In [18] and [19], Martin et al. discuss the problem of finding fuzzy associations between different taxonomies by means of fuzzy Formal Concept Analysis (FCA). This task comes close to ours, but our approach differs from it by using crisp taxonomies as well as by exploiting GARs to connect them. Additionally, fuzzy FCA does not describe the hierarchical relations of the trees themselves nor the relation between two hierarchies.

A more relevant field involves discovering non-taxonomic relations in ontology learning by AR mining [12], [20]. In [12], GARs are also used in order to mine co-occurrences among pairs of words in text and thus to connect different parts of an ontology. Although the algorithm has been claimed to be built on the original one of [4], it is in fact quite different. The rules are pruned in such a way that more general ancestral rules replace more specific ones. In contrast, we suggest discovering the most interesting rules by analyzing deviations of child rules from their parent rules as proposed by [4]. This is also the reason why we discard the approach of multi-level ARs [26], [27], where the separate search for confidence and support thresholds at each hierarchy level is required. We rather aim at finding the most interesting rules along the path from a leaf (most specific) concept to the root (most general) of the hierarchy.

As discussed above, instance-based ontology matching is a research field closely related to our task. The key idea is that the more sets of instances corresponding to two concepts overlap, the more related they are. Developed methods (e.g. HICAL [21], GLUE [22] and a more recent oPLMap [23]) use the instances directly associated with a concept for finding correspondences between the concepts, however, without using ARs. On the contrary, in [24] a case study of ontology matching by means of pairwise ARs was reported recently. In this study, Paulheim et al. also used the DBpedia-Yago dataset in a similar setting, but applied the standard support-confidence framework and did not extract GARs. Another application of association analysis to ontology matching is AROMA, short for Association Rule Ontology Matching Approach [1], which also utilizes pairwise ARs for mapping between ontologies. However, it does not use any hierarchical IMs but rather a rule selection criterion with the implication intensity measure[1]. We will consider this method later on as AROMA pruning, because we would like to discuss it together with other pruning methods. It consists of checking the IM value of any more generative rule than a given rule r. Rules defined as more generative are those that have either a more general antecedent or a more specific consequent or both. If a more generative rule with a greater IM value than that of r does not exist, r is said to be significant and is selected. The main difference between ontology matching methods and our work is that they map or combine similar taxonomies, while we are generally interested in analyzing connections between different taxonomies.

Additionally, we can consider two hierarchical pruning methods more as related approaches because they use hierarchy for removing redundant rules. The first is hierarchical pruning of [4], which we later refer to as Generalized Rule Pruning (GRP). Srikant et al. proposed pruning more specialized rules deeper in the hierarchy unless they differ significantly from their ancestor rules as measured by calculating expected values, or expectations, for support and confidence ($SupExp$, $CnfExp$, see Table II). This enables significant reduction of the found rule set as compared with standard AR mining in the presence of a hierarchy. The second hierarchical pruning method was proposed in [25] as the Generalized rule Confidence Constraint (GCC). It compares each rule r with all rules where the consequent has any descendant of the consequent of r and the same antecedent. If none of those rules has a confidence greater than the minimum confidence threshold, then r should be retained, otherwise it should be discarded. We use the hierarchical pruning methods for comparison in the experimental part of this work.

Hierarchical IMs were introduced in [2], [9] and [26]. In the last case, only minor changes of calculating support and confidence were proposed, for example, taking into account only a subset of transactions containing at least one of the concepts of the antecedent and the consequent and not the total number of transactions. We will discuss the hierarchical IMs in the next section.

III. Approach

A. Problem Definition

A taxonomy is a set of concepts C connected by a tree or a Directed Acyclic Graph (DAG) $T = (C, G)$, and every concept $a \in C$ is connected to a parent $\widehat{a} \in C$ through an edge in the graph G, i.e. $a \leq_{re} \widehat{a}$ where $\leq_{re}$ is the *is_a* partial order relation (subsumption relation) [1]. Further, the ancestors of a concept a are concepts lying on the path between this concept and the root: $\{anc_i \mid a \leq_{re} anc_i;\ a, anc_i \in C\}$. If an object is assigned to a certain concept it should also possess all its ancestors. As depicted in the example of Figure 1, the concepts are organized hierarchically and if some proteins (objects) belong to the 'cell bone marrow' concept and are connected to 'cell proliferation in bone marrow', this also implies that these proteins are in 'animal cell' and involved in 'single organism BP' (i.e. belong to all those concepts). The task is then to derive associations between a concept pair where the concepts are from different taxonomies T_1 and T_2, i.e. to find out how $a \in C_1$, $b \in C_2$, $C_1 \in T_1$, $C_2 \in T_2$ and $T_1 \neq T_2$ are related by the set of given objects annotated by both concept sets. We define also the parent rules of $a \to b$ as $\widehat{a} \to b$ and $a \to \widehat{b}$. One-to-one GARs are then the associations $a \to b$ allowed to have items from different levels of the taxonomies, such as $a \cap b = \emptyset$ and b is not an ancestor of a.

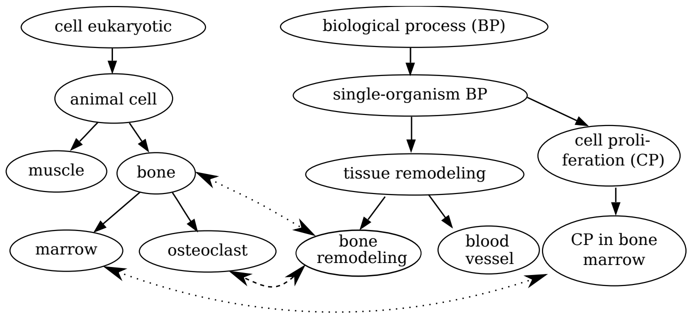

Figure 1: Example hierarchies (adapted excerpt from CL-GO) and relation mismatches

In our setting, GARs that describe the most interesting one-to-one relationships between the concepts of two ontologies are discovered by an Apriori-like [3] algorithm based on the analysis of concept co-occurrences. Two possible scenarios can be distinguished: The first one considers the concepts as labels of classified data. In the AR terminology, each transaction can be seen as an object labeled by concepts, as in the above example. Another example would be to search for connections between entities classified by concepts as in Wikipedia pages. Thus, the concepts of different ontologies can be connected if they are assigned together to the same Wikipedia page (object). The first three datasets for the experiments are created in this way. The second scenario consists of analyzing unclassified data, for example, a text corpus. The co-occurrence of terms corresponding to the concepts of different ontologies in the same sentence enables associations to be built between them. The approach can be useful because co-occurring concepts may indicate an important relation, e.g. if in a news group a certain company/product/model is often discussed with a certain malfunction. In this case, a transaction is represented by a single sentence of the text. Mining GARs rather than ARs in this scenario is even more advantageous because allows one to overcome the data sparseness problem typical for short texts [28] by extending the set of searched term combinations.

The main stages of the algorithm in the case of classified data are as follows: First, the objects with the concepts of a dataset are taken as input. Then, ARs are created by connecting a concept from C_1 to a concept from C_2. The obtained rules are ranked according to an IM and a fixed number of the top rules are taken from the rule set. Different IMs can be applied

[1] Although one should distinguish between selection and pruning rules (by selection we understand sorting with retaining rules not meeting the pruning criteria).

at this stage. And finally, the quality of the reduced rule set is evaluated by performance measures. In the second scenario, sentences are scanned for concepts. After the transactions are filled with the concepts, it can be proceed as in the first scenario.

B. Interestingness Measures for Association Rules

Given a set of transactions of the size n, n_a (n_b) and n_{ab} correspond to the number of transactions containing the item a (b) and both items, respectively. Further, $n_{\bar{a}}$ is the number of transactions without the item a. The corresponding frequencies of the items are defined as *Support*, denoted by p (i.e. $p_{ab} = \frac{n_{ab}}{n}$). The formulas for different IMs applied to a rule $a \rightarrow b$ can be found in Table II. The most popular of them is *Confidence* that is an estimate of the conditional probability P($b|a$) of b given a. In the standard setting, itemsets are first pruned by a minimum *Support* threshold to find frequent itemsets and then the rules connecting them are pruned by a minimum *Confidence* threshold to select the most interesting ones. As we intend to find all possible rules from C_1 to C_2 and not only the most frequent ones, we investigate all rules where *Support* is greater than zero ($p_{ab}>0$). The motivation is that the leaf concepts in a hierarchy may be relatively infrequent in comparison to the concepts of higher levels, but the associations between leaf concepts may be more significant because the concepts of high levels tend to produce trivial rules. It was mentioned in the literature [29], that the minimum *Support* constraint prevents discovering most interesting rules if *Support* is not directly related to the interestingness of a rule as in the case of a hierarchy.

It should be noted that the *Support* of a rule grows by climbing up in the hierarchy and the following relations hold [36]:

(a) $Sup(a,\widehat{b}) \geq Sup(a,b)$; (c) $Sup(\widehat{a},\widehat{b}) \geq Sup(a,b)$;
(b) $Sup(\widehat{a},b) \geq Sup(a,b)$; (d) $Cnf(a,\widehat{b}) \geq Cnf(a,b)$.

where $\widehat{a}$ is the parent of a in hierarchy H_1, and $\widehat{b}$ is the parent of b in H_2.

An additional disadvantage of the well established support-confidence framework is that "*Confidence* is unable to extract truly interesting rules", as stated in [37]. For example, a rule with a high *Confidence* value can still be of low relevance, as the *Support* of the consequent is not taken into account: If it is higher than the *Confidence* of the rule, then the items are even negatively correlated [37], [38]. Another aspect is that the *Confidence* needs the minimum *Support* to filter spurious rules, e.g. rules with a low antecedent *Support* and a much higher consequent *Support*. Such rules have also a "specific to general" character when connecting ontologies. Often rules with a high *Support* and a high *Confidence* tend to be trivial. Therefore, numerous alternative IMs have been proposed in the literature [13]. We will compare some of them in the presented setting.

Despite the variety of existing IMs, only a few of them possess a valuable property of null-transaction invariance, which was recently shown to be critically important for mining large datasets [11]. This property means that a measure is not

Table II: List of used interestingness measures and abbreviations. Every measure has two parameters a, b.

Measure name	Abbrev.	Formula	Ref.
Null-invariant			
Confidence	*Cnf*	$\frac{p_{ab}}{p_a}$	[3]
Sebag-Schoenauer	*Seb*	$\frac{p_{ab}}{p_a - p_{ab}}$	[13]
Jaccard	*Jac*	$\frac{p_{ab}}{p_a + p_b - p_{ab}}$	[10]
Cosine	*Cos*	$\frac{p_{ab}}{\sqrt{p_a * p_b}}$	[8]
AllConfidence	*ACnf*	$min(Cnf(a,b), Cnf(b,a))$	[8]
Kulczynski	*Kulc*	$\frac{p_{ab}}{2} * (\frac{1}{p_a} + \frac{1}{p_b})$	[11]
Not Null-invariant			
Support	*Sup*	p_{ab}	[3]
Lift	*Lif*	$\frac{Cnf}{p_b}$	[10]
Conviction	*Cnv*	$\frac{p_a * p_{\bar{b}}}{p_{a\bar{b}}}$	[30]
Certainty factor	*Crf*	$\begin{cases} \frac{Cnf - p_b}{1 - p_b}, & \text{if } Cnf > p_b \\ \frac{Cnf - p_b}{p_b}, & \text{otherwise} \end{cases}$	[31]
Loevinger	*Loe*	$\frac{p_{ab} - p_a * p_b}{p_a - p_a * p_b}$	[13]
Piatetsky-Shapiro	*PS*	$n * (p_{ab} - p_a * p_b)$	[32]
Bayes Factor	*BF*	$\frac{Seb * (1 - p_b)}{p_b}$	[13]
Centered Confidence	*CCnf*	$Cnf - p_b$	[13]
Klosgen	*Klos*	$\sqrt{p_{ab}} * (Cnf - p_b)$	[33]
Odds Ratio	*OD*	$\frac{p_{ab} * p_{\bar{a}\bar{b}}}{p_{a\bar{b}} * p_{\bar{a}b}}$	[8]
Kappa	κ	$2 * \frac{p_{ab} - p_a * p_b}{p_{\bar{a}} * p_b + p_a * p_{\bar{b}}}$	[34]
J-measure	*JM*	$p_{ab} * log(\frac{p_{ab}}{p_a * p_b}) + p_{a\bar{b}} * log(\frac{p_{a\bar{b}}}{p_a * p_{\bar{b}}})$	[8]
Least Contradiction	*LC*	$\frac{p_{ab} - p_{a\bar{b}}}{p_b}$	[13]
Gini index	*GINI*	$p_a * (Cnf^2 + (\frac{p_{a\bar{b}}}{p_a})^2) + p_{\bar{a}} * ((\frac{p_{\bar{a}b}}{p_{\bar{a}}})^2 + (\frac{p_{\bar{a}\bar{b}}}{p_{\bar{a}}})^2) - p_b^2 - p_{\bar{b}}^2$	[10]
Yule's Q	*YQ*	$\frac{p_{ab} * (p_{\bar{a}\bar{b}}) - (p_{a\bar{b}}) * (p_{\bar{a}b})}{p_{ab} * (p_{\bar{a}\bar{b}}) + (p_{a\bar{b}}) * (p_{\bar{a}b})}$	[10]
Yule's Y	*YY*	$\frac{\sqrt{p_{ab} * (p_{\bar{a}\bar{b}})} - \sqrt{(p_{a\bar{b}}) * (p_{\bar{a}b})}}{\sqrt{p_{ab} * (p_{\bar{a}\bar{b}})} + \sqrt{(p_{a\bar{b}}) * (p_{\bar{a}b})}}$	[10]
Collective Strength	*CS*	$\frac{p_{ab} + p_{\bar{a}\bar{b}}}{p_a * p_b + p_{\bar{a}} * p_{\bar{b}}} * \frac{p_{\bar{a}} * p_b + p_a * p_{\bar{b}}}{p_{\bar{a}b} + p_{a\bar{b}}}$	[34]
Laplace	*L*	$\frac{(p_{ab} * n) + 1}{(p_a * n) + 2}$	[34]
Zhang	*Zhang*	$\frac{p_{ab} - p_a * p_b}{max(p_{ab} * p_{\bar{b}}, p_{a\bar{b}} * p_b)}$	[34]
ϕ-coefficient	ϕ	$\frac{p_{ab} - p_a * p_b}{\sqrt{p_a * p_b * (p_{\bar{a}}) * (p_{\bar{b}})}}$	[10]
ϕ-Confidence	ϕCnf	$\phi * Cnf$	[35]
ϕ-Jaccard	ϕJac	$\phi * Jac$	our
ϕ-AllConfidence	$\phi ACnf$	$\phi * ACnf$	our
ϕ-Kappa	$\phi\kappa$	$\phi * \kappa$	our
Expectations and Dif-Measures			
Support Expectation	*SupExp*	$p_{\widehat{a}b} * \frac{p_a}{p_{\widehat{a}}}$	[4]
Confidence Expectation	*CnfExp*	$\frac{p_{\widehat{a}b}}{p_{\widehat{a}}}$	[4]
Jaccard Expectation	*JacExp*	$\frac{p_{\widehat{a}b} * \frac{p_a}{p_{\widehat{a}}}}{p_a + p_b - p_{\widehat{a}b} * \frac{p_a}{p_{\widehat{a}}}}$	our
Support Interestingness	*Int*	$\frac{Sup}{SupExp + Sup}$	our
Confidence Interestingness	*CnfInt*	$\frac{Cnf}{CnfExp + Cnf} = Int$	our
Confidence Difference	*CnfDif*	$Cnf * (Cnf - CnfExp)$	our
Jaccard Difference	*JacDif*	$Jac * (Jac - JacExp)$	our
AllConfidence-Dif	*ACnfDif*	$ACnf * (ACnf - ACnfExp)$	our
ϕ-JacDif	$\phi JacDif$	$\phi * JacDif$	our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2320722, IEEE Transactions on Knowledge and Data Engineering

affected by the number of null-transactions, i.e. transactions that contain none of the items of interest. Typically, the number of occurrences of an item is small when compared to the total number of transactions. If a measure is affected by the number of null-transactions it will produce unstable results depending on the current size of the dataset. For this reason, we include six null-invariant measures in the set of measures to be analyzed in this paper: *Confidence*, *Cosine*, *Jaccard*, *AllConfidence*, *Kulczynski*, and *Sebag-Schoenauer*. Additionally, some popular but not null-invariant measures such as, for example, *Lift* and *Conviction* are also considered. ϕ*-coefficient* and *Kappa* are in principle not null-invariant measures but both converge to a certain value if the number of transactions increases (for more details see online supplemental material [39]).

C. Hierarchical Interestingness Measures

Unfortunately, none of the conventional IMs is well-suited for mining GARs. In such a case, rules in higher hierarchy levels subsume rules in deeper levels. Thus, the hierarchy can be used successfully for pruning redundant rules as. For example, in the hierarchical pruning method GRP only the rules whose *Support* (or *Confidence*) is more than c times the expected value are said to be interesting, where $c > 1$ is a user-defined threshold. These expectations are calculated based on the hierarchy (see Table II).

Both GRP and GCC are based on the standard support-confidence framework and therefore share its shortcomings. Furthermore, GRP additionally depends upon the right choice of the parameter c. Although AROMA pruning step exploits an alternative IM, it does not use expectations, which is also true for GCC.

To overcome these limitations we developed a general class of hierarchical IMs which are calculated using a conventional IM and its expectation. First, we proposed a new IM by replacing *Support* and *Confidence* constraints of GRP through a direct calculation of the measure value on the basis of expectations [2]. Since the objective behind this measure was to improve GRP, it also uses the ratio of real and expected values. However, in contrast to GRP, it can be applied not only to *Support* and *Confidence*, but also to any other IM with a range of [0, 1], and for which an expectation can be defined and interpreted meaningfully. Comparing the real value MV of a metric M for a rule $a{\rightarrow}b$ with its expectation value ME, the $Interestingness$ is:

$$MInt(a,b) = \begin{cases} \frac{MV(a,b)}{MV(a,b)+ME(a,b)} = \frac{1}{1+\frac{ME(a,b)}{MV(a,b)}}, & \text{if } MV(a,b)>0 \\ 0, & \text{if } MV(a,b)=0. \end{cases}$$

Examples of expectations for *Support*, *Confidence*, and *Jaccard* are shown in Table II. They correspond to an extension of the formulas for support and confidence expectations from [4], substituting $Sup(a,b)$ for $SupExp(a,b)$ in the calculation of the IMs. For the root nodes they are set to $p_a * p_b$, p_b, and $\frac{p_a * p_b}{p_a + p_b - p_a * p_b}$, respectively. This is based on the independence assumption for the distributions of items a and b. Although this first hierarchical IM was shown to successfully detect interesting rules, it has a significant limitation of low noise resistance [39].

Table III: Example transactions

Nr.	Rule	*Support*	Item	*Support*
1	cell eukaryotic → bone remodeling	2	cell eukaryotic	150
2	bone → bone remodeling	2	bone	100
3	marrow → bone remodeling	1	marrow	90
4	osteoclast → bone remodeling	1	osteoclast	10

To eliminate this drawback, we developed another measure *Interestingness by Difference* [9] which is based on the difference between the real and the expected values as follows:

$$MDif(a,b) = MV(a,b)(MV(a,b) - ME(a,b)).$$

$MDif$ depends directly on the magnitude of the real value and is therefore less sensitive to very small expectations. If the expected values become greater than the real ones, $MDif$ converts to negative. This happens, for example, when a sibling or even the parent of a node has a stronger relation to the consequent of the rule. Similarly to the *Interestingness*, this measure can also be used in different variants (based on *Support*, *Confidence*, the *Jaccard* coefficient or on any other conventional measure with the range of [0, 1]).

To illustrate the behavior of this measure, a simple example is depicted in Table III. $CnfDif$ for three rules Rule 2-4 can be calculated as follows:

Rule 2: $0.02 * (0.02 - \frac{2}{150}) \approx 0.0001$, Rule 3: $\frac{1}{90} * (\frac{1}{90} - \frac{2}{100}) \approx -0.0001$, Rule 4: $0.1 * (0.1 - \frac{2}{100}) \approx 0.008$.

In this case, Rule 4 has the highest value among three rules but it is much smaller than the possible maximum because the real *Confidence* value was low. Since the sibling of Rule 3 has a stronger relation to the consequent, it has a negative value.

We chose $JacDif$, $CnfDif$, and $ACnfDif$ for the experimental comparison. The advantage of the latter one is that it uses not only the *Confidence* of a rule $a{\rightarrow}b$, but also the *Confidence* of its reverse ($b{\rightarrow}a$). Calculating expectations, the rules can be generalized on the antecedent (consequent) side or on both sides. For simplicity, we assume in this study that the generalization side is always the left-hand side (antecedent) that corresponds to the set C_1. In the case of $ACnfDif$, however, both hierarchies should be taken into account, depending on the particular *Confidence* values of a rule in each direction. Indeed, the expectation $ACnfExp$ will be either $\frac{p_{\widehat{a}b}}{p_{\widehat{a}}}$ if $Cnf(a,b){\leq}Cnf(b,a)$ or $\frac{p_{a\widehat{b}}}{p_{\widehat{b}}}$ otherwise.

In order to further develop the *Interestingness by Difference*, we now propose ϕ-$JacDif$ which is the product of ϕ and $JacDif$. It utilizes the Pearson coefficient ϕ similarly to the approach of [35] where it was used to evaluate the degree of dependence between the antecedent and the consequent of a rule. Taking the zero value at independence, it is able to separate positively and negatively correlated items and facilitate rule pruning. Only those rules which connect highly positively correlated items will obtain high scores; negatively correlated items, in turn, will obtain a negative ϕ value not allowing them to be ordered high in the ranked rule list. It was shown by [35] that the high discriminating power of ϕ can be combined with another IM improving its performance. Casas-Garriga multiplied ϕ by *Confidence* for a two-fold purpose: First, to assign high scores to rules that have a positive correlation between items and, second, to distinguish the strength of implication of the rule against its reverse. This measure is denoted by ϕ-*Confidence* in Table II.

In some cases, e.g. finding corresponding concepts in two different ontologies, it is imperative that the direction is indistinguishable because an AR reflects a symmetrical relationship. Thus, multiplying ϕ by a symmetrical IM would be more appropriate in our setting. For comparison with ϕ-$JacDif$, we combined ϕ with some other symmetrical measures: *Jaccard*, *AllConfidence*, and κ. We normalized measures which can become negative to the [0,1] interval before multiplying with ϕ.

D. Performance measures

For the evaluation of experimental results, two types of performance measures were used. The standard recall and precision measures as well as measures derived from the relation learning accuracy [40]. Recall (R) is known as the ratio of the number of discovered true elements to the total number of true elements. Precision (P) is the ratio of the number of discovered true elements to the number of all elements found. F-1 is the harmonic mean of R and P. Recall and precision are inappropriate to assess the quality of rules embedded into a hierarchy. Both measures do not differentiate between specific and general relations as implied by a hierarchy and cannot recognize neighborhoods. Thus, other performance measures can become more useful for the discrimination of the obtained results in our task.

To examine how a found rule set is related to the true rule set given by the hand-coded relations, we improved the method of Generic Relation Learning Accuracy ($\overline{RLA}$) proposed in [12]. The idea of $\overline{RLA}$ is to measure an average distance between discovered and true relations. While $\overline{RLA}$ considers the quality of a found rule set only in terms of accuracy, we were also interested in most compact rule sets, and therefore developed Relation Learning Recall (RLR), Precision (RLP), and F-1 measure (RLF-1). As stated above, the definitions of precision and recall are based on an exact match between relations. In contrast, the use of $\overline{RLA}$ enables the matching degrees between each pair of relations to be assessed, leading to more sensitive performance measures. To obtain RLR and RLP, we renamed the original RLA into $\widehat{RLA}$ and changed its meaning by focusing on the true rule set rather than on the discovered one. Consequently, the sum over found rules was replaced by the sum over the true rules – RLA_s. The following formulas describe the calculation of RLR and RLP in detail.

$$CLA(a_1,a_2) = \begin{cases} 0, \text{if } root \text{ is not an ancestor of } lcs(a_1,a_2) \\ \frac{\delta(lcs(a_1,a_2),root)+1}{\delta(lcs(a_1,a_2),root)+1+\delta(a_1,a_2)}, \text{otherwise} \end{cases} \tag{1}$$

$$MA((a_1,b_1),(a_2,b_2)) = \sqrt{CLA(a_1,a_2) * CLA(b_1,b_2)} \tag{2}$$

$$RLA(v,U) = \max_{u \in U} MA(u,v) \tag{3}$$

$$\overline{RLA}(U,V) = \frac{1}{|V|} \sum_{v \in V} RLA(v,U) \tag{4}$$

$$\widehat{RLA}(u,V) = \max_{v \in V} MA(u,v) \tag{5}$$

$$RLA_s(U,V) = \sum_{u \in U} \widehat{RLA}(u,V) \tag{6}$$

$$RLR(U,V) = \frac{1}{|U|} RLA_s; \;\; RLP(U,V) = \frac{1}{|V|} RLA_s \tag{7}$$

where U is the true rule set and V the discovered rule set. For each true rule $u \in U$ the so-called Matching Accuracy $MA(u,v) = MA((a_1,b_1),(a_2,b_2))$ is calculated (Eq. 2) w.r.t. each discovered rule $v \in V$ on the basis of Concept Learning Accuracy CLA (Eq. 1).

First, for each side of a rule, $CLA(u,v)$ measures the distance between two nodes in the hierarchy ($\delta(u,v)$) in relation to the distance between their least common superconcept (lcs) and the root node $root$. The distance is measured by the number of edges traversed on the shortest path between two nodes. In the case of a DAG, multiple inheritance leads to more than one lcs, but only one with the best CLA value is chosen. We slightly changed the formula of [12] by adding the unity to the root node distances in order to ensure the applicability of CLA to the rules containing the root concepts (the situation did not appear in [12]). In this way, mistakes made at higher levels of the hierarchy or which have large distances are penalized more severely. In contrary to [12], we allow multiple root concepts for an ontology and do not bridge the top concepts that are too far away ontologically, creating new roots. For example, the top concepts “abstraction” and “physical entity” from the Yago taxonomy are not connected, therefore the CLA between them is zero.

Then, MA is computed as the geometric mean of CLA of antecedents $CLA(a_1,a_2)$ and consequents $CLA(b_1,b_2)$ of both relations. For each r, the MA value of the best matching rule among all discovered relations is then taken as $\widehat{RLA}$ (Eq. 5). This allows a single found relation to cover multiple true relations, although not perfectly. The last step (Eq. 7) concerns the averaging over the true rules for RLR, and over the found rules for RLP. The main advantage of this approach as compared to that of [12] is that RLR and RLP play complementary roles similarly to R and P and combining them by RLF-1 leads to better performance assessment.

A simple example is depicted in Figure 1: The true rule (for this example) *osteoclast→bone remodeling*, named r_{OsBr} (Os stands for osteoclast and Br for bone remodeling), is compared to the found rule *bone→bone remodelling*, named r_{BoBr} (Bo stands for bone). The calculation of $MA(r_{OsBr}, r_{BoBr})$ is then as follows: $MA((Os,Br),(Bo,Br)) = \sqrt{\frac{\delta(Bo,Ce)+1}{\delta(Bo,Ce)+1+\delta(Os,Bo)} * \frac{\delta(Br,Bp)+1}{\delta(Br,Bp)+1+\delta(Br,Br)}} = \sqrt{\frac{3}{4}} \approx 0.87$, Ce stands for cell eukaryotic and Bp for biological process. For comparison the distance to the rule *marrow→cell proliferation in bone marrow*: r_{MaCm} (Ma stands for marrow and Cm for cell proliferation in bone marrow) is: $MA((Os,Br),(Ma,Cm)) = \sqrt{\frac{\delta(Bo,Ce)+1}{\delta(Bo,Ce)+1+\delta(Os,Ma)} * \frac{\delta(So,Bp)+1}{\delta(So,Bp)+1+\delta(Br,Cm)}} = \sqrt{\frac{1}{5}} \approx 0.45$, So stands for single-organism. The rule

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2320722, IEEE Transactions on Knowledge and Data Engineering

$MA((Os, Br), (Bo, Br))$ gets a much higher MA value because it is nearer to the true rule.

The CLA calculation's dependence on the distance between $root$ and lcs causes that both $\overline{RLA}$ and RLF-1 are sensitive to the hierarchy depth. For example, if the cell eukaryotic node had a parent (i.e. all nodes would have a distance to the root node increased by 1), the last result would change to $MA((Os, Br), (Ma, Cm)) = \sqrt{\frac{2}{9}} \approx 0.47$. This can complicate to some extent result comparison for datasets with different depths. On the other side, the advantage of this approach is that deeper rules are seen to be closer to each other than higher rules even with the same relative difference. It is similar to the approaches proposed in the field of hierarchical multi-label classification [41].

The range of the $\overline{RLA}$ and RLR is from 0 for a total miss for every rule and 1 for a perfect match of the two rule sets. It should be noted that RLP can theoretically become greater than 1 when $|U| > |V|$, i.e. it is no longer normalized. Nevertheless, in practice, the number of found rules is often greater than the number of true ones. Ideally, all found rules would correspond to the true rules and it would produce RLP equal to 1, but usually it is much smaller.

IV. Experiments

A. Settings

In order to provide a fair comparison of IMs and do not need to choose a proper threshold value for each measure, we applied two simple techniques to selecting small rule sets. The first one is the well-known selection of only "top k rules" where k in our case is chosen to be equal to the number of the true rules in each dataset, so that precision and recall become equal. This setting attracted recently more attention as a useful alternative to the threshold setting [42]. In the second setting "best possible" introduced in [9], the best possible size of a found rule set w.r.t. the obtained F-1 value is iteratively searched for. To this end, all rules are first sorted by their measure values in descending order and F-1 is calculated, then the last rule (with the lowest measure value) is removed from the set. The next iteration starts with recalculating the F-1 value. Finally, the rule set with the highest F-1 value is selected, thus the rule set size can be different for each method.

The pruning methods from the literature, with which we compared our results, have different parameters. For GCC [25], we used the minimum *Confidence* of 0.3 as it achieved the best results in the original paper. For GRP, the threshold c was set to 1.1 as in [4]. In the AROMA study, the authors used a threshold to choose the rule set size. Here, the rule set size is given by the specific setting of the experiment.

Further, we grouped together results of IMs with the same rule ordering: Cnf and Seb as well as Cnv, Loe and Crf. The equivalence of the first four was already shown in [13]. Crf is the same as Loe if $Cnf{>}p_b$ and this is true for the highest ranked rules. The ordering is also the same for YQ and YY, since for a rule $(a_i{\rightarrow}b_i)$ to have a value greater than another $(a_j{\rightarrow}b_j,\ i{\neq}j)$ both measures should obey the same condition $(\frac{p_{ab_i} * p_{\bar{a}\bar{b}_i}}{p_{ab_j} * p_{\bar{a}\bar{b}_j}} > \frac{p_{a\bar{b}_i} * p_{\bar{a}b_i}}{p_{a\bar{b}_j} * p_{\bar{a}b_j}})$.

For the experiments, four datasets were prepared – called Movies, Universities, DBpedia-Yago and CL-GO – with two similar ontologies in each case and with increasing complexity. We evaluated the sets of ARs obtained by every IM against a ground truth set, which had been built manually. Our experimental comparison on the first three datasets showed that several IMs were able to extract small sets of rules with a high fraction of correct associations for all datasets. In the next step, we applied them to CL-GO, a larger and more complex dataset from the bioinformatics domain.

B. Data

The Movies dataset previously used in [19] comprises movies from the Internet Movie Database (IMDb) and the Rotten Tomatoes (RT) databases. From the roughly 90,000 movies in each database, we selected only those which had at least one genre assigned (like thriller or comedy) and which had an almost exact match for title and director in both databases. Additionally, only the genres with more than 250 entries in IMDb and more than 30 entries in RT were chosen. As there were no predefined genre hierarchies in this dataset, they were extracted automatically by the Apriori algorithm as described in [43]. To generate a ground truth set of associations, 48 connections between the IMDb and RT genre hierarchies were created manually: e.g. "IM: Comedy→RT: Comedy", "IM: Sci-Fi→RT: Science Fiction and fantasy". Due to the small number of the true connections, they were rather rare associations.

The Universities dataset was published in the Ontology Alignment Evaluation Initiative (OAEI 2009) and used in earlier studies [1], [22]. In this case we trained a neuro-fuzzy classifier [43] with descriptions of the courses of a university (i.e. Cornell) so that the courses of the other university (i.e. Washington) could be classified into the same taxonomy (i.e. the Cornell Courses Taxonomy). The 10-fold cross-validation classification on the Cornell data yielded a micro-F-1 score of 76.2±1.8% and on the Washington data of 80.1±0.6%. Then we applied the classifier trained with the Cornell University data to the prediction of the categories for the courses of Washington University. Co-occurrences were used for establishing associations between the ontologies similarly to [22]. For further details see [39].

A large dataset DBpedia-Yago from [24] is based on Wikipedia articles with structured information, as the main goal of DBpedia is to make them available in a unified format. Yago is a semantic knowledge base created on top of Wordnet, Wikipedia and Geonames. The used dataset was downloaded from the Internet page of DBpedia [44]. It had 264 (the root node "Thing" was removed) and 96,472 concepts for DBpedia and Yago, respectively. As the ground truth set a partial gold standard mapping between the DBpedia and the Yago ontologies [45] containing 170 connections was taken, but only 162 associations actually appeared in the data. More details are given in [39].

The fourth dataset was introduced by [46] for connecting Cell Ontology (CL) with Gene Ontology (GO) by the text analysis of PubMed articles. In this case, there were no preclassified data, and the text corpus was searched for the co-occurrences of terms directly corresponding to the concepts of CL and GO in the same sentence. We used the dataset as it was processed in [9]. The so-called cross-products were used as a partial ground truth rule set. The cross-products were generated automatically by an information extraction method based on term decomposition [47] and were then verified manually. Due to pattern matching of substrings in concept names, only concepts possessing similar names can be connected by the method. This is a serious drawback because most of such connections are trivial like "CL: T cell→GO: T cell receptor complex" and therefore not interesting. Since our method is explicitly designed to avoid discovering obvious rules w.r.t. the hierarchy, the comparison with cross-products cannot serve as the only criterion of its quality. As we could not find any CL-GO: Molecular Function cross-products, only those between CL and GO: Biological Process as well as GO: Cell Component were taken from the Open Biological and Biomedical Ontologies (OBO) foundry [48]. In total 196 relationships (from the original 677) actually co-occurred in the dataset.

To cope with DAGs, multiple expectations for DBpedia-Yago and CL-GO (from all possible parents) were calculated and the smallest one was used.

C. Results of the Movies Dataset

In the first experiment, the impact of using different IMs on discovering the hand-coded associations of the Movies dataset was studied. It was first investigated how the measures rank the true rules among an increasing amount of found rules (Figure 2). The whole rule set consisted of 4,895 rules. At the Y-axis, the graph shows the number of found true rules among the top X rules obtained by the measures Jac and Cnf as well as by their respective Dif counterparts and Int. Obviously, the steeper the increase of a curve, the better the IM. One can see that Cnf was the worst among the presented measures. Jac, Cos, and $ACnf$ were among the best measures in this task, the latter two had curves very similar to Jac (not shown in the figure). Note that all three measures are considered to be well-suited for rare ARs. Indeed, the manually created associations of this dataset were rather rare. In only eight of them, the *Support* was greater than 5%. This was the reason why they were difficult to find. So, all hand-coded relations could be found by Jac only after reaching about 3,000 rules in total. The growth of other curves was even slower.

Especially interesting is the comparison of curves between Cnf and $CnfDif$ likewise between Jac and $JacDif$. Both Dif measures first significantly outperformed their conventional counterparts, but lost their advantage towards the end. It can be explained by the hierarchically redundant nature of a large part of the true rules. Both curves had a steep increase approaching the end because redundant true rules were discovered by both measures very late, i.e. the rules which are expected from the hierarchy. It is because such rules typically have a small difference between their actual values and the expectations and therefore are ranked very low by the Dif measures. Int behaved similarly: It had a good start but it could not find general rules and therefore it performed worse than Cnf towards the end.

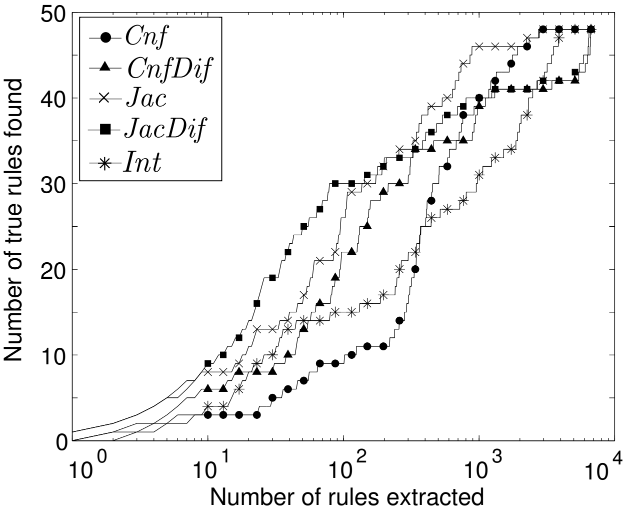

Figure 2: Number of true rules found in the top X rules extracted by Cnf and Jac, their respective $Difs$ and Int.

The results of the experiment setting "top 48 rules" are depicted in Table IV. The results of the performance measures are multiplied by 100 for readability. Here, Cnf again showed low performance and could find only 7 true rules. The hierarchical pruning method GCC, upon which it is based, performed similarly to Cnf, but the entire rule set it found was different. $CnfDif$ was again able to improve the performance of Cnf. Another measure that outperformed Cnf was $ACnf$: Although it is basically a confidence measure, it is better suited to connect two hierarchies because if there is a high Cnf value of both rule directions, the corresponding concepts can likely be important to each other.

The measures well-suited for rare ARs Jac, Cos, and $ACnf$ were good in this setting and found at least 16 true rules. Much better though were the measures κ, CS, ϕ, and those based on the latter as well as $JacDif$ and $ACnfDif$ which all found between 24 and 26 true rules. Multiplying by ϕ usually removes a large amount of unwanted rules for bad measures, as can be seen on the result of ϕCnf: The number of the true rules found by Cnf was tripled by combining it with ϕ. The best result in terms of F-1 had ϕ and $\phi\kappa$ followed by κ, CS, $\phi ACnf$, $ACnfDif$, and $\phi JacDif$ (all with the same result).

The more detailed comparison of the true rules extracted by $JacDif$ with those extracted by Jac (Table V) helps explain the differences in their rule sets. Nine rules not found by Jac but found by $JacDif$ were all relatively unexpected w.r.t. the hierarchy. They had a relatively low Jac value but at the same time their $JacExp$ values were much lower so that it led to relatively high $JacDif$ values. Jac, in turn, ranked higher rules like e.g. "IM: Drama→RT: Comedy" even with higher $JacExp$ than Jac (0.29 vs. 0.23), that is with a negative $JacDif$ value. There was only one true rule which Jac found but $JacDif$ did not. This was due to its high expectation of about 0.22.

The influence of multiplying by ϕ can be best explained by the comparison of the true rule sets of $\phi JacDif$ and $JacDif$. It shows that they differed by five rules. $\phi JacDif$ found three rules which had high ϕ values and relatively low $JacDif$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2320722, IEEE Transactions on Knowledge and Data Engineering

Table IV: Movies results, three best results are marked in bold; fnd=number of extracted rules, tf=number of true rules found.

Measure		"top 48 rules"				"best possible"								
		tf	F-1	RLF-1	$\overline{RLA}$	fnd	tf	R	P	F-1	RLR	RLP	RLF-1	$\overline{RLA}$
1.	Cnf/Seb	7	14.58	40.04	57.49	65	9	18.75	13.85	15.93	55.53	41.00	47.17	58.15
2.	Jac	16	33.33	68.27	61.65	61	21	43.75	34.43	38.53	73.88	58.13	65.07	61.95
3.	Cos	17	35.42	67.30	67.07	36	16	33.33	44.44	38.10	64.34	85.79	73.53	69.99
4.	$ACnf$	16	33.33	67.57	58.54	56	20	41.67	35.71	38.46	72.09	61.79	66.54	59.62
5.	$Kulc$	15	31.25	60.29	70.93	15	11	22.92	73.33	34.92	45.73	146.34	69.69	**90.81**
6.	$Lift$	18	37.50	58.44	61.64	39	17	35.42	43.59	39.08	53.83	66.26	59.40	69.97
7.	$Cnv/Crf/Loe$	10	20.83	52.20	54.79	108	22	45.83	20.37	28.21	77.68	34.53	47.81	54.73
8.	PS	11	22.92	47.71	57.00	124	27	56.25	21.77	31.40	80.21	31.05	44.77	60.12
9.	BF	21	43.75	62.60	67.59	47	21	43.75	44.68	44.21	62.60	63.93	63.26	69.03
10.	$CCnf$	16	33.33	62.98	62.85	59	21	43.75	35.59	39.25	69.30	56.38	62.17	64.04
11.	$Klos$	14	29.17	58.70	66.29	17	12	25	70.59	36.92	42.88	121.07	63.33	**81.62**
12.	OD	19	39.58	60.74	71.95	62	26	54.17	41.94	47.27	75.26	58.27	65.68	72.22
13.	κ	25	**52.08**	**75.22**	78.51	40	24	50	60	54.55	69.95	83.94	**76.31**	**81.96**
14.	JM	22	45.83	65.25	69.17	62	27	56.25	43.55	49.09	77.11	59.70	67.30	70.79
15.	LC	9	18.75	57.58	60.35	13	8	16.67	61.54	26.23	38.19	141.01	60.10	80.70
16.	$GINI$	15	31.25	57.19	56.32	41	15	31.25	36.59	33.71	54.51	63.81	58.79	56.15
17.	YQ/YY	18	37.50	58.66	69.37	64	26	54.17	40.62	46.43	75.26	56.44	64.51	70.86
18.	CS	25	**52.08**	**75.36**	79.38	58	28	58.33	48.28	52.83	78.62	65.07	71.21	76.96
19.	L	7	14.58	40.04	56.29	65	9	18.75	13.85	15.93	55.53	41.00	47.17	58.15
20.	$Zhang$	21	43.75	62.60	67.59	47	21	43.75	44.68	44.21	62.60	63.93	63.26	69.03
21.	ϕ	26	**54.17**	75.11	**81.49**	54	29	60.42	53.70	**56.86**	79.75	70.89	75.06	81.36
22.	ϕCnf	21	43.75	66.76	74.09	38	20	41.67	52.63	46.51	60.46	76.38	67.49	78.31
23.	ϕJac	23	47.92	73.73	77.89	58	27	56.25	46.55	50.94	77.99	64.54	70.63	78.45
24.	$\phi ACnf$	25	**52.08**	**75.82**	79.38	48	25	52.08	52.08	52.08	75.82	75.82	75.82	79.38
25.	$\phi\kappa$	26	**54.17**	75.11	**81.49**	51	28	58.33	54.90	**56.57**	79.14	74.49	**76.75**	81.42
26.	Int	14	29.17	55.76	47.52	36	13	27.08	36.11	30.95	52.46	69.95	59.95	55.04
27.	$JacDif$	24	**50**	73.64	71.49	43	24	50	55.81	52.75	73.64	82.20	**77.68**	77.38
28.	$CnfDif$	13	27.08	61.57	53.14	97	22	45.83	22.68	30.34	78.68	38.93	52.09	50.23
29.	$ACnfDif$	25	**52.08**	72.87	72.93	45	25	52.08	55.56	53.76	72.87	77.73	75.23	76.22
30.	$\phi JacDif$	25	**52.08**	73.70	**80.25**	54	29	60.42	53.70	**56.86**	79.75	70.89	75.06	81.36
31.	GCC	7	14.58	40.04	59.62	38	7	14.58	18.42	16.28	37.94	47.93	42.35	60.33
32.	AROMA	17	35.42	68.77	62.60	56	21	43.75	37.50	40.38	72.31	61.98	66.75	64.81
33.	GRP	8	16.67	57.89	50.12	165	26	54.17	15.76	24.41	82.79	24.08	37.31	40.83

Table V: Comparison of rules undetected by Jac and $JacDif$ respectively among the top 48 rules, JD $= Jac - JacExp$.

#	Rule	level	n_{ab}	Jac	$JacExp$	JD	$JacDif$
	True rules undetected by Jac but found by $JacDif$						
1	IM: gangster→RT: Organized Crime	1	14	0.127	0.017	0.110	0.014
2	IM: Fantasy→RT: Science-Fiction and fantasy	0	48	0.171	0.022	0.149	0.026
3	IM: War→RT: War	1	24	0.133	0.015	0.117	0.016
4	IM: detective→RT: Detectives	2	27	0.125	0.040	0.085	0.011
5	IM: robbery→RT: Thieves	1	13	0.126	0.024	0.102	0.013
6	IM: lesbian→RT: Gay/Lesbian	2	17	0.112	0.031	0.081	0.009
7	IM: family-relationships→ RT: Family interaction	1	57	0.176	0.039	0.137	0.024
8	IM: Sci-Fi→RT: Futuristic	0	34	0.158	0.011	0.147	0.023
9	IM: kidnapping→RT: Kidnapping & missing persons	1	27	0.165	0.015	0.149	0.025
	True rule undetected by $JacDif$ but found by Jac						
1	IM: Animation→ RT: Animation	2	28	0.241	0.216	0.025	0.006

values (ranked between 70 and 79) because they were either expected from the hierarchy or in the case of the root rule "IM: Horror→RT: Slasher" its $JacExp$, calculated under the independence assumption, was relatively high. On the other hand, two rules not found by $\phi JacDif$ were ranked as 49th and 51st, that is they were likely to be found within a slightly larger set.

Among the hierarchical pruning methods, AROMA achieved the best results comparable with those of Jac and Cos. The other two methods performed worse than the hierarchical IMs.

As expected, the majority of the results of *Dif* measures were superior to that of *Interestingness*. Its problem was that the rules with an extremely low *Support* belonging to a parent with a much higher *Support* were mostly selected as interesting. For a detailed example see [39].

It should be noted that the hierarchy-based performance measures RLF-1 and $\overline{RLA}$ can often disagree with F-1 because they both count even poor matches between the true and found rules. A noticeable difference between the same F-1 and different RLF-1 values of the measures $\phi JacDif$ and $\phi ACnf$ can be explained by the fact that the true rules with the antecedent concept item IM: Crime were represented by $\phi JacDif$ with a specific sibling rule whereas $\phi ACnf$ had a more general corresponding concept assigned, producing a higher match accuracy value. Moreover, despite one additional true rule, ϕ and $\phi\kappa$ had lower RLF-1 values than $\phi ACnf$ because their discovered rule sets were generally not as representative as that of $\phi ACnf$ and did not match the true rule set as well. In turn, the disagreement of both hierarchy-based measures is caused by the difference in their assessment of the compliance between the rule sets.

CnfDif outperformed *Cnf* but its results were worse than those of the other *Dif* measures because of the weakness of the *Confidence* in this task. Besides that, *Dif* measures showed good results in terms of F-1. Their additional advantage lies in the compression of the found rule set, as can be seen in the next experiment setting "best possible" (Table IV). This setting is useful because redundant rules can be removed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2320722, IEEE Transactions on Knowledge and Data Engineering

iteratively, and the performance can be increased either by a higher precision or by a higher recall.

The optimal size of the found rule set in this case equals the number of the true rules. A larger set increases the probability of a better recall, but the precision decreases simultaneously. Therefore, the highest F-1 values are to be expected from measures with the rule set size around 48. Still this relaxation of a fixed rule set size is generally beneficial for the IMs.

In the "best possible" experiment, the F-1 scores were similar to those of the previous experiment for the majority of measures. The exceptions were $Cnv/Crf/Loe$, PS, $Klos$, OD, and LC with an absolute improvement of 7-8%. While $Cnv/Crf/Loe$, PS, and OD could indeed increase the number of found true rules, $Klos$ and LC even decreased the number of found true rules and their improvement was only due to the small sizes of their rule sets. Among the pruning methods, only GRP could achieve a significant improvement of about 8%. This is probably because of its use of the expectations. However, it is still inferior to the Dif measures which had much more compact rule sets.

ϕ, $\phi JacDif$ and $\phi\kappa$ were again the best measures. This time the first two showed even the highest F-1 value. κ and $ACnfDif$ had in this setting fourth and fifth F-1 values and very compact rule sets of 40 and 45 rules, respectively. It is important to note that $JacDif$ and $ACnfDif$ extracted less rules than their respective counterparts but had nevertheless more true rules among them. In contrast, $CnfDif$ increased the found rule set as compared with Cnf. Though all of them could improve the F-1 and RLF-1 values of their counterparts, $JacDif$ achieved the overall best RLF-1 value.

The difference between RLF-1 and $\overline{RLA}$ can be demonstrated in this setting. An important drawback of $\overline{RLA}$ is its high correlation with precision. It can be seen in the example of $Kulc$ which had the highest precision and the highest $\overline{RLA}$ value due to its small rule set with a large part of true rules. This was also reflected in its best RLP. However, both harmonic mean measures F-1 and RLF-1 as opposite to $\overline{RLA}$ balanced their values by taking the low recall into account. The high correlation coefficient between P and $\overline{RLA}$ of 0.9 calculated for all IMs confirmed their strong dependence. In contrast, it was only 0.81 for the pair P and RLF-1.

Further, we evaluated how well a random set of found rules could cover the true rule set. It was simulated by randomly sorting all rules ten times and taking the first 48 rules. The average values of RLF-1 and $\overline{RLA}$ were then 25.6±5.2 and 23.9±1.3, respectively. All measures in this experiment, even the worst ones, achieved higher values in both cases, indicating that they are much better than a random choice.

D. Results of the Universities Dataset

For the Universities dataset, the results of the setting "top 55 rules" are depicted in Table VI. In this experiment, Cnf showed the worst F-1 performance and Jac the best. $ACnf$, κ, CS, $JacDif$, and $ACnfDif$ had the second best F-1 value. They all found 32 true rules, but they did not have the same rule set, as is reflected by their different RLF-1 values. (Even though the values of κ and CS were identical, their rule sets differentiated by two rules.) Most of the ϕ based measures produced the third best F-1 value with 31 true rules. Only ϕCnf with 20 true rules was inferior to the others, but its result is nevertheless 20 times better than that of Cnf with only one true rule.

Analyzing the results, one can see again that the F-1 values of several measures diverged from their RLF-1 and $\overline{RLA}$ values. For example, $Lift$ discovered nine true rules, which had lower RLF-1 and $\overline{RLA}$ values than YQ/YY – which found only one true rule. This was due to the fact that the rule set of YQ/YY was more representative. This can be measured by the calculation of the mean coverage values for all found rules excluding the true ones. The comparison of these values shows that they were much lower for 46 rules of $Lift$ (average RLA_s=0.37 and $\overline{RLA}$=0.51) than for 54 rules of YQ/YY (average RLA_s=0.49 and $\overline{RLA}$=0.6).

In this experiment, GCC could also improve the result of Cnf, but it was still much worse than those of the Dif measures. AROMA and GRP behaved similarly to the previous experiment, whereas AROMA was much better than GCC and GRP.

The results of the "best possible" setting for this dataset are shown in Table VI. They are similar to those of the "top 55 rules" setting. We can see that $ACnfDif$ and $JacDif$ performed well and obtained the best and the third best F-1 values, respectively. Jac scored second best. The RLF-1 results of the Jac and $JacDif$ changed places. Two out of six ϕ based measures, ϕJac and $\phi ACnf$, showed the fourth best performance, the others were slightly worse.

$Lift$, Cnv, Crf and Loe discovered the largest number of true rules. However, as compared with the rule set of $ACnfDif$, the increase of 10% in the number of true rules of these IMs required almost a 600% increase in the total number of found rules. $ACnfDif$ could discover a very compact and though highly representative rule set as can be seen from its best RLF-1 value. This points to the fact that the rules extracted by $ACnfDif$ covered the true rule set very well.

Here again, one can see the problem of $\overline{RLA}$ and why it is a poor performance measure. It assigned the highest value to LC with the smallest rule set which contained only 22% of the true rules and therefore had the lowest RLR value among all IMs. Although RLP was also very high in this case, RLF-1 as a harmonic mean was strongly influenced by low RLR and therefore LC was not competitive with the best measures in terms of RLF-1.

The analysis that was focused on the rule sets extracted by high performance measures showed they were not very different. The sets extracted by Jac and $JacDif$ differed by three true rules. The rule "WASH: Linguistics_LING→CORN: Linguistics" produced an expectation value that was higher than the actual Jac value, so that its $JacDif$ value became negative and it was ranked by $JacDif$ as 30,221st. The other two rules did not reach the top rules but they were in the first 100 rules of both IMs. $\phi JacDif$ and ϕ had the same rule set but with different orderings. The true rules of $JacDif$ and $\phi JacDif$ differed by five rules.

The pruning methods GCC and GRP could again improve the results of Cnf but their results were inferior to the worst

Table VI: Universities results, best three results are marked in bold, fnd=number of extracted rules, tf=number of true rules found.

Measure		"top 55 rules"				"best possible"								
		tf	F-1	RLF-1	$\overline{RLA}$	fnd	tf	R	P	F-1	RLR	RLP	RLF-1	$\overline{RLA}$
1.	Cnf/Seb	1	1.82	44.34	63.84	278	27	49.09	9.71	16.22	82.18	16.26	27.15	61.27
2.	Jac	33	**60**	**78.26**	**86.09**	53	33	60	62.26	**61.11**	78.18	81.13	**79.63**	86.92
3.	Cos	31	**56.36**	74.92	85.30	54	31	56.36	57.41	56.88	74.92	76.31	75.61	85.57
4.	$ACnf$	32	**58.18**	77.35	83.95	56	33	60	58.93	59.46	78.26	76.86	77.55	84.24
5.	$Kulc$	16	29.09	63.53	73.97	27	16	29.09	59.26	39.02	56.22	114.52	75.42	84.83
6.	$Lift$	9	16.36	47.09	58.58	312	34	61.82	10.90	18.53	83.88	14.79	25.14	54.24
7.	$Cnv/Crf/Loe$	3	5.45	47.65	65.07	307	34	61.82	11.07	18.78	88.63	15.88	26.93	59.91
8.	PS	18	32.73	64.06	76.14	85	25	45.45	29.41	35.71	71.07	45.99	55.84	74.25
9.	BF	8	14.55	50.60	64.72	116	20	36.36	17.24	23.39	66.96	31.75	43.07	63.62
10.	$CCnf$	10	18.18	48.87	68.86	140	25	45.45	17.86	25.64	72.88	28.63	41.11	64.77
11.	$Klos$	14	25.45	60.14	70.81	114	29	52.73	25.44	34.32	78.67	37.95	51.20	70.43
12.	OD	15	27.27	59.44	68.30	68	17	30.91	25	27.64	60.52	48.95	54.13	67.54
13.	κ	32	**58.18**	75.83	**85.72**	46	30	54.55	65.22	59.41	71.12	85.03	77.45	**87.77**
14.	JM	26	47.27	71.79	77.13	56	27	49.09	48.21	48.65	72.70	71.40	72.04	77.54
15.	LC	16	29.09	59.17	73.54	14	12	21.82	85.71	34.78	48.48	190.48	77.29	**94.34**
16.	$GINI$	18	32.73	61.87	66.00	85	24	43.64	28.24	34.29	72.56	46.95	57.01	63.21
17.	YQ/YY	1	1.82	49.62	60.98	199	26	47.27	13.07	20.47	75.57	20.89	32.73	62.82
18.	CS	32	**58.18**	75.83	**85.56**	52	32	58.18	61.54	59.81	73.24	77.46	75.29	86.66
19.	L	1	1.82	45.74	64.76	285	27	49.09	9.47	15.88	82.18	15.86	26.59	59.91
20.	$Zhang$	8	14.55	50.60	64.72	116	20	36.36	17.24	23.39	66.96	31.75	43.07	63.62
21.	ϕ	31	**56.36**	74.62	85.13	54	31	56.36	57.41	56.88	74.62	76.00	75.30	85.88
22.	ϕCnf	20	36.36	60.45	76.24	77	27	49.09	35.06	40.91	69.17	49.41	57.65	75.72
23.	ϕJac	31	**56.36**	74.92	85.22	46	30	54.55	65.22	59.41	71.12	85.03	77.45	**87.77**
24.	$\phi ACnf$	31	**56.36**	76.44	84.74	46	30	54.55	65.22	59.41	71.12	85.03	77.45	**87.77**
25.	$\phi\kappa$	31	**56.36**	74.92	85.22	54	31	56.36	57.41	56.88	74.62	76.00	75.30	85.88
26.	Int	9	16.36	47.37	58.82	111	17	30.91	15.32	20.48	62.53	30.99	41.44	54.76
27.	$JacDif$	32	**58.18**	**78.37**	84.04	50	32	58.18	64	**60.95**	78.37	86.21	**82.10**	87.58
28.	$CnfDif$	14	25.45	59.64	61.59	91	22	40	24.18	30.14	67.08	40.54	50.54	61.87
29.	$ACnfDif$	32	**58.18**	**77.65**	85.04	45	31	56.36	68.89	**62.00**	76.57	93.59	**84.23**	**88.41**
30.	$\phi JacDif$	31	**56.36**	74.92	85.22	54	31	56.36	57.41	56.88	74.62	76.00	75.30	85.88
31.	GCC	9	16.36	50.22	67.74	139	24	43.64	17.27	24.74	75.76	29.98	42.96	62.20
32.	AROMA	25	45.45	74.78	74.87	65	29	52.73	44.62	48.33	78.11	66.09	71.60	74.76
33.	GRP	6	10.91	53.91	67.89	163	26	47.27	15.95	23.85	78.57	26.51	39.65	60.90

Dif measure ($CnfDif$). AROMA was better than $CnfDif$ but it was still not as good as the other Dif measures and it was worse than several standard ones.

Despite the good classification results of the classifier, some of the measures discovered rules based on misclassifications or ontology mismatches like "WASH: Vietnamese_VIET→CORN: Bengali" or "WASH: Tibetan_TIB→CORN: Hindi" (language courses).

E. Results of the DBpedia-Yago Dataset

The results of the DBpedia-Yago dataset for the setting "top 162 rules" are shown in Table VII. The highest performance in terms of F-1 was achieved by two Dif measures with 76 found true rules: $JacDif$ and $\phi JacDif$. Several measures including ϕ, κ, and $\phi\kappa$ were able to find 73 true rules in this setting, which was the second best result. Surprisingly, the same result was obtained by $Kulc$ and LC which performed poorly on the previous two datasets. This can be explained by the large amount of true, very balanced rules: $p_{ab}{\approx}p_a{\approx}p_b$. Among 73 true rules of $Kulc$, LC and ϕ there were 71 common rules with this property. For this reason, the IMs sensitive to such rules were superior on the dataset.

In general, the Dif measures outperformed their counterparts in terms of RLF-1. Except for $ACnfDif$, they were also superior in terms of F-1.

In this experiment, one can again observe that the lack of exact matches of true rules can be successfully compensated by the higher quality of the whole rule set. It can be seen on the RLF-1 values of $JacDif$ and $ACnfDif$. The latter discovered 12 true rules less but obtained though a better RLF-1 value, actually the best one among all measures. This was due to the high mean coverage of the rules without an exact match in the true rule set. So, the average RLA_s value of 98 such rules of $ACnfDif$ was 0.59 in contrast to 0.47 as obtained for 86 rules of $JacDif$. Thus, a higher average value was the key reason for the better RLF-1 performance of $ACnfDif$.

Another illustration of the difference between F-1 and RLF-1 is the fact that although $Lift$ and L produced comparable RLF-1 values and both found at least one true rule, their RLF-1 values were still lower than the ones from $Zhang$, which did not discover any true rule. It was mostly because $Zhang$ had very few true rules not covered at all by its found rule set. The other two measures had more than two times more such rules (25 of $Zhang$ against 62 and 68 of $Lift$ and L, respectively). The important problem with $Zhang$ was that 733 rules achieved unity. This happens when the maximum of $p_{ab}{-}p_b{*}p_{ab}$ or $p_a{*}p_b{-}p_{ab}{*}p_b$ is equal to $p_{ab}{-}p_a{*}p_b$ which is true for $p_{ab}{=}p_a$. This is a strong evidence for the poor discriminating power of the measure. Nevertheless, the selected first 162 rules (their ordering is arbitrary and depends on how the items were ordered and how the algorithm extracted the indices) were coincidentally able to cover the true rules better than the corresponding rule sets of $Lift$ and L. For example, the rule "DB: Vein→YG:

Table VII: DBpedia-Yago, three best results are marked in bold; fnd=number of extracted rules, tf=number of true rules found.

Measure		"top 162 rules"				"best possible"								
		tf	F-1	RLF-1	$\overline{RLA}$	fnd	tf	R	P	F-1	RLR	RLP	RLF-1	$\overline{RLA}$
1.	Cnf/Seb	4	2.47	43.76	56.54	2133	128	79.01	6.00	11.15	93.95	7.14	13.26	62.07
2.	Jac	71	43.83	70.89	84.23	195	89	54.94	45.64	49.86	77.24	64.17	70.10	83.79
3.	Cos	72	**44.44**	70.94	84.37	193	88	54.32	45.60	49.58	77.05	64.68	70.32	83.83
4.	$ACnf$	71	43.83	70.68	84.02	213	93	57.41	43.66	49.60	79.95	60.81	69.08	83.83
5.	$Kulc$	73	**45.06**	71.01	84.42	195	88	54.32	45.13	49.30	77.34	64.25	70.19	83.87
6.	$Lift$	6	3.70	32.15	56.58	262	10	6.17	3.82	4.72	41.92	25.92	32.03	55.71
7.	$Cnv/Crf/Loe$	4	2.47	43.76	56.54	2030	127	78.40	6.26	11.59	93.46	7.46	13.81	62.51
8.	PS	11	6.79	52.65	70.86	844	71	43.83	8.41	14.12	73.76	14.16	23.75	73.54
9.	BF	4	2.47	43.76	56.54	1656	59	36.42	3.56	6.49	71.83	7.03	12.80	58.24
10.	$CCnf$	14	8.64	43.52	60.31	889	108	66.67	12.15	20.55	86.46	15.76	26.65	69.29
11.	$Klos$	22	13.58	55.37	74.25	449	64	39.51	14.25	20.95	71.80	25.90	38.07	76.61
12.	OD	52	32.10	69.01	80.20	121	49	30.25	40.50	34.63	63.73	85.33	**72.97**	83.29
13.	κ	73	**45.06**	71.19	**84.43**	193	89	54.94	46.11	**50.14**	77.24	64.83	70.49	84.15
14.	JM	26	16.05	57.61	73.58	387	65	40.12	16.80	23.68	69.55	29.12	41.05	75.65
15.	LC	73	**45.06**	**71.80**	84.18	212	93	57.41	43.87	49.73	78.41	59.92	67.93	84.11
16.	$GINI$	18	11.11	54.69	67.38	511	64	39.51	12.52	19.02	69.65	22.08	33.53	70.37
17.	YQ/YY	0	0	35.21	74.80	10000	0	0	0	0	56.45	0.91	1.80	65.33
18.	CS	72	**44.44**	70.94	84.37	197	89	54.94	45.18	49.58	77.27	63.54	69.73	83.83
19.	L	1	0.62	29.48	57.41	1753	117	72.22	6.67	12.22	89.76	8.30	15.19	63.43
20.	$Zhang$	0	0	41.93	51.65	1656	59	36.42	3.56	6.49	71.83	7.03	12.80	58.24
21.	ϕ	73	**45.06**	71.13	**84.43**	198	89	54.94	44.95	49.44	77.96	63.78	70.16	**84.32**
22.	ϕCnf	71	43.83	70.90	84.09	184	84	51.85	45.65	48.55	75.67	66.62	**70.86**	**84.57**
23.	ϕJac	71	43.83	70.89	84.23	196	89	54.94	45.41	49.72	77.70	64.22	70.32	84.24
24.	$\phi ACnf$	72	**44.44**	71.10	84.30	200	91	56.17	45.50	**50.28**	77.98	63.16	69.79	83.98
25.	$\phi\kappa$	73	**45.06**	71.13	**84.43**	197	89	54.94	45.18	49.58	77.67	63.87	70.10	84.28
26.	Int	0	0	22.22	49.06	268	3	1.85	1.12	1.40	27.75	16.77	20.91	50.94
27.	$JacDif$	76	**46.91**	**71.84**	83.56	265	105	64.81	39.62	49.18	86.36	52.79	65.53	84.19
28.	$CnfDif$	20	12.35	46.96	57.56	508	86	53.09	16.93	25.67	79.33	25.30	38.36	64.45
29.	$ACnfDif$	64	39.51	**75.47**	81.05	219	81	50	36.99	42.52	83.25	61.58	70.79	82.37
30.	$\phi JacDif$	76	**46.91**	**72.54**	83.78	182	86	53.09	47.25	**50**	77.73	69.19	**73.21**	**84.61**
31.	GCC	43	26.54	65.47	68.96	223	53	32.72	23.77	27.53	73.89	53.68	62.19	70.47
32.	AROMA	4	2.47	63.77	72.20	408	50	30.86	12.25	17.54	84.26	33.45	47.89	70.80
33.	GRP	6	3.70	45.01	55.55	973	111	68.52	11.41	19.56	88.11	14.67	25.15	66.26

vein105418717", was ranked by $Zhang$ as 554th and by L as 365th. However, $Zhang$ had a relative rule ("DB: Vein→YG: bloodvessel105417975") in the first 162 rules. Discriminating better, L ranked this rule similarly to the former rule, at a position around 360, thus producing a worse RLF-1 value.

Another measure which did not find any true rules was YY. It has even poorer discriminating power since it is equal to unity if $p_{ab}=p_a$ or $p_{ab}=p_b$ and there were 287,066 such rules. Yet, its RLF-1 value was still higher than that of $Lift$ with six found true rules. This demonstrates the need for the simultaneous use of F-1 and RLF-1 performance measures because of their complementary nature.

In this experiment, GCC surprisingly outperformed AROMA and GRP in terms of F-1. It strongly improved the F-1 score of Cnf. AROMA and GRP achieved results comparable to that of Cnf. However, all hierarchical pruning methods were inferior to the majority of the standard IMs.

The results of DBpedia-Yago for the "best possible" setting are depicted in Table VII. To speed up the process of finding the best value, here we used only the top first 10,000 rules and from them we calculated the best F-1 value.

One can see that the average number of found rules for all measures in this experiment is larger than in the other experiments. This was also the reason why none of the measures had an RLP greater than 100.

This time only one measure YQ/YY was not able to find the true rules at all. The ϕ based measures performed slightly better than in comparison with the same setting in the Universities dataset. $\phi ACnf$ had the highest F-1 value followed by κ and $\phi JacDif$. Multiplying by ϕ again highly improved the F-1 score of Cnf, ϕCnf had even the third best RLF-1 value.

ϕ extracted a rule set similar to that of $JacDif$. While the discovered true rules of $JacDif$ were also high ranked by ϕ, the reverse was not true. From three true rules found by ϕ in its top 198 rules and not found by $JacDif$ in its 265 rules, one produced a $JacDif$ value of zero and the other two were ranked 533rd and 826th by $JacDif$. This was because the expectations of these rules were high.

Although $JacDif$ can often optimize the size of a discovered rule set, it was not the case this time. A significant amount of the true rules were concentrated around the places between 240th and 265th positions (10 true rule in 25 discovered ones). Therefore, the measure had a large rule set and its precision was penalized. $\phi JacDif$ was able to improve the results of its base measures because it compressed its rule set and boosted its precision.

Here, the pruning methods were much better than Cnf but still worse than many standard measures.

F. Results of CL-GO Dataset

In this experiment, the partial cross-products were used as a ground truth rule set. As said before, the serious disadvantage of this evaluation is that the cross-products represent obvious associations because the method connects concepts with similar names. These associations can therefore be ranked by our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2014.2320722, IEEE Transactions on Knowledge and Data Engineering

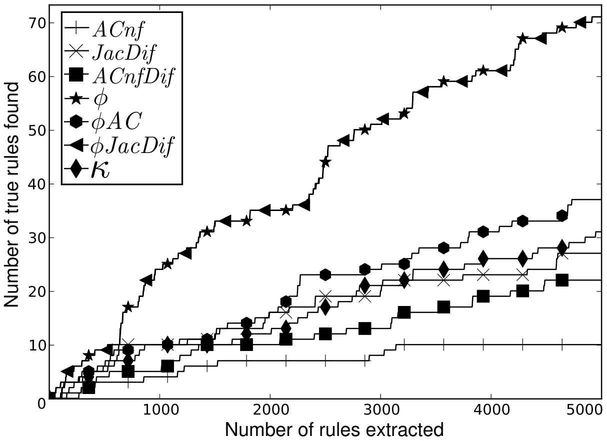

Figure 3: Number of found true rules in the top X rules extracted by the measures on the CL-GO dataset.

approach lower than less obvious rules. In Fig. 3, one can see the increase in the number of found true rules among the top X rules with growing X, for a representative group of selected measures (mostly the measures which performed well in the previous experiments). The graph shows that ϕ and $\phi JacDif$ had the same performance (stars over triangles) and were the best. The second best curve was produced by $\phi ACnf$. The Dif measures outperformed their counterparts. The low numbers of true rules found in this experiment as compared to the results of Movies, Universities and DBpedia-Yago datasets point to the fact that other more interesting and unexpected rules were ranked higher.

Among the top ranked rules, several interesting connections like "CL: heterocyst→GO: nitrogen fixation" were discovered. Heterocyst is a differentiated cyanobacterial cell that carries out nitrogen fixation [49]. It is important to note that such rules could not be found by the name matching approach of [47] and therefore were absent in the cross-products. Our approach also found some associations between concepts with similar names like "CL: nitrogen fixing cell→GO: nitrogen fixation" which were nevertheless not contained in the cross-products from the OBO foundry. Other interesting examples found were: "CL: glandular cell of stomach→GO: acid secretion", "CL: spermatocyte→GO: meiosis I" and "CL: osteoclast→GO: bone remodeling". They all are reliable and confirm the usability of our approach for knowledge bridging applications.

V. Discussion

In the experiments on four real world datasets, the hierarchical IMs based on Jac and $ACnf$ achieved very good results. Cnf, on the contrary, produced some of the worst results, partly because we did not use the minimum *Support* threshold. However, its performance could be improved by multiplying by ϕ as proposed in [35]. Moreover, the combinations of ϕ with other measures produced also good results for the problem studied. Some of the best results, for example, on Movies and DBpedia-Yago datasets, were achieved by such measures. Especially successful among them was $\phi JacDif$.

Similarly to ϕ, κ and $Lift$ also measure the deviation from the independence condition. κ was able to retrieve many true rules achieving good results in terms of F-1 performance measure. $Lift$, in turn, is not normalized and favors low values of p_a or p_b. This explains in part its poor performance.

The experiments also showed the advantage of RLF-1 over $\overline{RLA}$ as a more balanced performance measure which is useful as an extension of F-1 assessing not only perfect but also partial matches between the true and discovered rules.

The Dif measures were often able to extract high level rules for several branches covering most of the true rules and thus achieved high RLF-1 values. As stated before, they were not always able to find all of the true rules, as the true rules are sometimes expected from the hierarchy. However, the RLF-1 performance measure showed clearly that the hierarchical measures produced generally better rule sets than their flat counterparts. $ACnfDif$ performed especially well: In six experiments it showed twice the best rule set in terms of RLF-1 and once the third best rule set.

We compared our results with those previously reported for the datasets Universities and DBpedia-Yago. A similar study on the Universities dataset was performed by David et. al. in [1] where the AROMA and GLUE approaches were compared. We can calculate the F-1 value of AROMA for their setting as 41% which is similar to our results.

The DBpedia-Yago dataset was already used in a similar experiment in [24]. Paulheim et al. applied the standard support-confidence approach and set several minimum *Support* thresholds to boost the *Confidence* performance. They reported the maximum F-1 value of 25% for their approach. As we showed, *Confidence* is not a good choice for this task at all. Some other measures like κ and $\phi JacDif$ are more appropriate, even when different minimum *Support* thresholds can be used with *Confidence*. If the extracted rule set should be compact and representative, the new class of hierarchical measures presented here is more promising.

VI. Conclusion

In this paper, we studied hierarchical Interestingness Measures (IMs) as a means of association rule mining for connecting multiple ontologies. Finding interesting connections between concepts of different ontologies enables their deep analysis with the aim of knowledge extraction or better understanding the relations between the ontologies describing the same domain. For this purpose, four datasets with ground truth sets of connections between two ontologies were used to extract association rules, 36 flat and hierarchical IMs were compared and examined which ones are most successful in this setup. Additionally, a new class of hierarchical IMs was further developed.

The hierarchical pruning methods proposed in the literature: GCC, AROMA and GRP showed a lower performance as compared with the hierarchical measures. The latter are superior in multiple ways: they do not need parameters as well as an additional costly post-processing step, they naturally rank redundant rules lower and can obtain more compressed rule sets in the setting "best possible". In all experiments, the *Interestingness by Difference* measures: $JacDif$, $ACnfDif$ and $\phi JacDif$ achieved very good results in terms of the standard performance measure F-1 and the novel measure RLF-1 which reflects the partial coverage between the true and discovered rules. The latter showed that they could better compress the found rule set and find more descriptive rules. They also obtained better results as compared with other studies reported in the literature for the datasets Universities and DBpedia-Yago.

Roughly, there could be a division of the measures in three groups, the measures based on the independence assumption like ϕ as well as the Dif measures had the best results, the measures with the performance similar to *Confidence* or based on it, they had the worse results and the rest had scattered results.

Future work comprises the use of association rules connecting different ontologies in order to improve the classification performance and to extract knowledge which can be easily interpreted by a human to better understand the problem at hand. A related subject would involve discovering many-to-many relations between ontologies.

References

[1] J. David, F. Guillet, and H. Briand, "Association rule ontology matching approach," *Int'l J. Semantic Web Inf. Syst.*, vol. 3, no. 2, pp. 27–49, 2007.

[2] F. Benites and E. Sapozhnikova, "Learning different concept hierarchies and the relations between them from classified data," in *Intel. Data Analysis for Real-Life Appl.: Theory and Practice*, 2012, pp. 18–34.

[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. of 20th Int'l Conf. on VLDB*, 1994, pp. 487–499.

[4] R. Srikant and R. Agrawal, "Mining generalized association rules," in *Proc. of the 21th Int'l Conf. on VLDB*, 1995, pp. 407–419.

[5] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.

[6] A. Rodríguez, J. M. Carazo, and O. Trelles, "Mining association rules from biological databases: Research articles," *J. Am. Soc. Inf. Sci. Technol.*, vol. 56, no. 5, pp. 493–504, 2005.

[7] L. Szathmary, P. Valtchev, and A. Napoli, "Generating rare association rules using the minimal rare itemsets family," *Int'l J. Software and Informatics*, vol. 4, no. 3, pp. 219–238, 2010.

[8] A. Surana, U. Kiran, and P. K. Reddy, "Selecting a right interestingness measure for rare association rules," in *16th Int'l Conf. on Management of Data (COMAD)*, 2010.

[9] F. Benites and E. Sapozhnikova, "Generalized association rules for connecting biological ontologies," in *BIOINFORM. 2013*, pp. 229–236.

[10] P. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Inf. Syst.*, vol. 29, pp. 293–313, 2004.

[11] T. Wu, Y. Chen, and J. Han, "Re-examination of interestingness measures in pattern mining: a unified framework," *Data Min. Knowl. Discov.*, vol. 21, pp. 371–397, 2010.

[12] A. Maedche and S. Staab, "Discovering conceptual relations from text," in *Proc. of the 14th ECAI*, 2000, pp. 321–325.

[13] S. Lallich, O. Teytaud, and E. Prudhomme, "Association rule interestingness: Measure and statistical validation," in *Quality Measures in Data Mining*, ser. Studies in Comp. Intel., 2007, vol. 43, pp. 251–275.

[14] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, "On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid," *European J. of Operational Research*, vol. 184, no. 2, pp. 610–626, 2008.

[15] R. Miani, C. Yaguinuma, M. Santos, and M. Biajiz, "Narfo algorithm: Mining non-redundant and generalized association rules based on fuzzy ontologies," in *Enterprise Inf. Syst.*, 2009, vol. 24, pp. 415–426.

[16] N. Lavrac, A. Vavpetic, L. N. Soldatova, I. Trajkovski, and P. K. Novak, "Using ontologies in semantic data mining with segs and g-segs," in *Discovery Science*, vol. 6926, 2011, pp. 165–178.

[17] J. Völker and M. Niepert, "Statistical schema induction," in *The Semantic Web: Research and Applications, LNCS*, vol. 6643, 2011, pp. 124–138.

[18] T. Martin and Y. Shen, "Fuzzy association rules in soft conceptual hierarchies," in *Fuzzy Inf. Processing Society NAFIPS*, 2009, pp. 1–6.

[19] T. Martin, Y. Shen, and B. Azvine, "Granular association rules for multiple taxonomies: A mass assignment approach," in *URSW (LNCS)*, vol. 5327, 2008, pp. 224–243.

[20] J. Villaverde, A. Persson, D. Godoy, and A. Amandi, "Supporting the discovery and labeling of non-taxonomic relationships in ontology learning," *Expert Syst. Appl.*, vol. 36, pp. 10 288–10 294, 2009.

[21] R. Ichise, H. Takeda, and S. Honiden, "Rule induction for concept hierarchy alignment," in *Proc. of the 2nd WKSP O.L. at the 17th IJCAI*, 2001.

[22] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning to map between ontologies on the semantic web," in *Proc. of the 11th Int'l Conf. on World Wide Web*. NY, USA: ACM, 2002, pp. 662–673.

[23] H. Nottelmann and U. Straccia, "A probabilistic, logic-based framework for automated web directory alignment," in *Soft computing in ontologies and the semantic web.*, 2006, pp. 47–77.

[24] H. Paulheim and J. Fümkranz, "Unsupervised generation of data mining features from linked open data," in *Proc. of the 2nd Int'l Conf. on Web Intel., Mining and Semantics*, 2012, pp. 31:1–31:12.

[25] E. Baralis, L. Cagliero, T. Cerquitelli, and P. Garza, "Generalized association rule mining with constraints," *Inf. Sci.*, vol. 194, Jul. 2012.

[26] P. Manda, F. M. McCarthy, and S. M. Bridges, "Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new go relationships," *J. of Biomedical Informatics*, vol. 46, no. 5, pp. 849–856, 2013.

[27] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," in *Proc. Int'l Conf. VLDB*, 1995.

[28] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proc. of the 17th Int'l Conf. on WWW (WWW '08)*. New York, NY, USA: ACM, 2008, pp. 91–100.

[29] G. I. Webb and S. S. Zhang, "k-optimal-rule-discovery," *DATA MIN KNOWL DISC*, vol. 10, no. 1, pp. 39–79, 2005.

[30] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," *SIGMOD Rec.*, vol. 26, no. 2, pp. 255–264, Jun. 1997.

[31] F. B. Galiano, I. J. Blanco, D. Sánchez, and M. A. V. Miranda, "Measuring the accuracy and interest of association rules: A new framework," *Intel. Data Analysis*, vol. 6, no. 3, pp. 221–235, 2002.

[32] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in *Knowledge Discovery in Databases*, 1991, pp. 229–248.

[33] S. Kannan and R. Bhaskaran, "Association rule pruning based on interestingness measures with clustering," *CoRR*, 2009.

[34] Y. Le Bras, P. Lenca, and S. Lallich, "Optimonotone measures for optimal rule discovery," *Comp. Intell.*, vol. 28, no. 4, pp. 475–504, 2012.

[35] G. Casas-Garriga, "Statistical strategies for pruning all the uninteresting association rules," in *ECAI*, 2004, pp. 430–434.

[36] V. O. de Carvalho, S. O. Rezende, and M. de Castro, "Evaluating generalized association rules through objective measures," in *Proc. 25th IASTED: A.I. and Applications*, 2007, pp. 301–306.

[37] T. Brijs, K. Vanhoof, and G. Wets, "Defining interestingness measures for association rules," *Int'l J. of Information Theories and Applications*, vol. 10, no. 4, pp. 370–376, 2003.

[38] J. L. Balcázar, "Closure-based confidence boost in association rules," in *WAPA*, 2010, pp. 74–80.

[39] *Online supplemental material*. http://doi.ieeecomputersociety.org/

[40] U. Hahn and K. Schnattinger, "Towards text knowledge engineering," in *Proc. of the 15th Nat'l/10th Conf. on A.I. (AAAI)*, 1998, pp. 524–531.

[41] F. Brucker, F. Benites, and E. P. Sapozhnikova, "An empirical comparison of flat and hierarchical performance measures for multi-label classification with hierarchy extraction," in *KES (1)*, ser. Lecture Notes in Computer Science, vol. 6881, 2011, pp. 579–589.

[42] G. I. Webb, "Filtered-top-*k* association discovery," *Wiley Interdisc. Rew.: Data Min. and Knowl. Disc.*, vol. 1, no. 3, pp. 183–192, 2011.

[43] F. Brucker, F. Benites, and E. P. Sapozhnikova, "Multi-label classification and extracting predicted class hierarchies," *Pattern Recognition*, vol. 44, no. 3, pp. 724–738, 2011.

[44] *DBPedia*, 2012. http://downloads.dbpedia.org/3.8/

[45] *DBpedia Yago*, 2012. http://www.netestate.de/De/Loesungen/DBpedia-YAGO-Ontology-Matching

[46] R. Hoehndorf, A. Ngomo, M. Dannemann, and J. Kelso, "From terms to categories: Testing the significance of co-occurrences between ontological categories," in *Proc. of the 3rd Int'l Symp. on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland*, 2008, pp. 53–60.

[47] M. Bada and L. Hunter, "Enrichment of obo ontologies," *J. of Biomedical Informatics*, vol. 40, no. 3, pp. 300 – 315, 2007.

[48] (2012) CL GO. http://www.obofoundry.org/index.cgi?show=mappings

[49] (2012) Uniprot. http://www.uniprot.org/keywords/364