Using Ontological Contexts to Assess the Relevance of Statements in Ontology Evolution

Fouad Zablith, Mathieu d'Aquin, Marta Sabou, Enrico Motta*

Knowledge Media Institute (KMi), The Open University
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
{f.zablith, m.daquin, r.m.sabou, e.motta}@open.ac.uk

Abstract. Ontology evolution tools often propose new ontological changes in the form of statements. While different methods exist to check the quality of such statements to be added to the ontology (e.g., in terms of consistency and impact), their relevance is usually left to the user to assess. Relevance in this context is a notion of how well the statement fits in the target ontology. We present an approach to automatically assess such relevance. It is acknowledged in cognitive science and other research areas that a piece of information flowing between two entities is relevant if there is an agreement on the context used between the entities. In our approach, we derive the context of a statement from online ontologies in which it is used, and study how this context matches with the target ontology. We identify relevance patterns that give an indication of relevance when the statement context and the target ontology fulfill specific conditions. We validate our approach through an experiment in three different domains, and show how our pattern-based technique outperforms a naive overlap-based approach.

1 Introduction

Ontologies are conceptual representations of defined domains. Consequently, they are subject to constant updates and evolution to keep-up with domain changes. Ontology evolution is a painstaking and time-consuming task. Thus we observe an increase in the availability of tools that automatically suggest new additions to be applied to ontologies in the form of statements [2, 8, 11, 19]. Nevertheless, although such tools help by automatically identifying ontology changes, they have introduced a new burden on users: inspecting the quality of a large number of proposed statements in terms of consistency and relevance with respect to the ontology.

There exist many tools that can be used to manage and preserve the consistency of an ontology after adding new statements [7, 15]. However, assessing the relevance of a statement with respect to an ontology is not a trivial task, and is usually left to the user. For example, introducing *Concert* as a type of *Event* in an academic related ontology might not result in any logical conflict

* Work funded by the NeOn project under EC grant number IST-FF6-027595

to the ontology, however, such an addition is not particularly relevant to the ontology, where events are mainly about conferences, seminars, workshops, etc. We understand statement relevance with respect to an ontology as an indication of how well it fits in the ontology.

Relevance is a core subject of interest in various domains including Artificial Intelligence, Cognitive Science [16, 17] and Information Retrieval [1, 9]. However, this problem is not very well explored in the domain of ontology evolution. As Wilson and Sperber noted in their work on relevance theory [16], two entities communicating and in exchange of knowledge, require a kind of agreement on the choice of context in which the conversation occurs. Moreover, they argue that "an input is relevant to an individual when it connects with background information he has available to yield conclusions that matter to him." [16]

Based on these key ideas, we present an approach towards automatically assessing the relevance of statements with respect to an ontology. Our process starts by identifying the context of a statement, by finding an online ontology in which it appears (Section 3). This context is matched to the ontology to derive the shared concepts. We initially investigate a naive overlap approach that takes into account the number of shared concepts. It is based on the idea that the more shared concepts exist between the target ontology and the external ontology defining the context of a statement, the more relevant a statement is. With the various limitations of this technique, we point out the need for a more sophisticated approach that takes into account not only the shared entities, but also the structure surrounding them. We accomplish this by identifying a set of patterns (Section 4), where each pattern has specific application conditions and a confidence value. When a pattern occurs at the intersection area of the statement context and the target ontology, a certain degree of confidence can be calculated. We back our work by an experiment in three domains (Section 5), showing that the pattern-based technique outperforms the naive overlap approach in terms of precision and recall, and can be used to support users in the selection of relevant statements during the process of ontology evolution (Section 6).

2 Related Work & Motivation

Our previous experiment conducted in the context of the Evolva ontology evolution tool, showed that a significant amount of statements proposed automatically to be added to a target ontology are irrelevant [19]. Within Evolva, currently users have to manually identify such statements (e.g., in the academic domain *Concert* is a type of *Event*) and select relevant ones (e.g., *Tutorial* is a type of *Event*). However with many statements to check, this can be a time-consuming task by itself. Thus having a mechanism that automatically gives an indication of the statements' degree of relevance would be of added value to the ontology evolution process.

Besides Evolva, there exist tools, e.g. SPRAT [8] and Text2Onto [2], which extract information from text documents, and convert them into ontological entities. Such tools mostly rely on the TF.IDF statistical measure to check the relevance of terms with respect to the corpus used. This of course assumes that the corpus has been selected to represent precisely the intended domain. In particular, if the extracted entities are intended to be used in or in conjunction with an existing ontology, this ontology is not currently taken into account in calculating the confidence degree of the extracted elements.

Automatically finding ontology changes is important for ontology evolution, however maintaining the consistency and quality of the ontology is equally significant. Thus several approaches have emerged recently that focus on evaluating the impact of statements on the ontology they are added to. For example in [13], a solution is suggested to highlight what is gained or lost as a result of adding an axiom (i.e., a statement) to an ontology. The aim here is to present the effect of a statement to the user, in order to make a more informed judgment in implementing the change and preserving conceptual consistency. Another approach proposes the evaluation of changes in ontology evolution using an impact function, which computes the cost involved in performing the change [12]. Tools such as RaDON [7] that check the consistency of the ontology after adding statements, are commonly used to evaluate the impact of the statements, in particular in evolution tasks. While these techniques provide valuable support to the users in assessing the impact of statements on their ontologies, to our knowledge there is still no solution to support them in assessing the relevance of such statements.

3 Overview of the Relevance Assessment Process

We understand the *relevance* of a statement s with respect to a target ontology O_t as an indication of how well s fits in O_t. An ontology is a set of statements, which we manipulate as a graph. A statement s is of the form $< subject, relation, object >$. We focus in this paper on the scenario in which the target ontology is extended by introducing statements that have one part (i.e. *object*) that already exists in O_t. The *relation* of s can be either of taxonomic type (*sub-class*, *super-class*), or other named relations. For now, we focus on the taxonomic relations, as they are less ambiguous than the named ones, of which relevance is much harder to assess even by users.

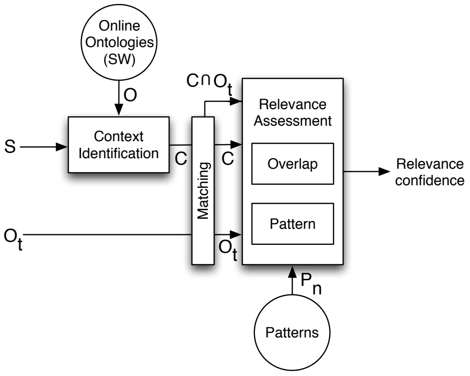


Fig. 1: Checking the relevance of statement s with respect to ontology O_t.

The relevance assessment process (Figure 1) starts with identifying a context C for s from online ontologies. Subsequently, the context C is matched to the target ontology O_t to identify shared concepts, which result from the intersection of the graphs C and O_t, and used for the relevance assessment.

Identifying the Context of a Statement. Similarly to other tools that exploit the open Semantic Web for performing a variety of tasks [5], our approach uses online ontologies as background knowledge to provide contextual information for a statement. To find online ontologies in which the statement appears, we use Scarlet, a relation discovery engine on the Semantic Web [14]. Scarlet uses the Semantic Web gateway Watson [4], and automatically selects and explores online ontologies *to discover relations between two given concepts*. For example, when relating two concepts labeled *Tutorial* and *Event*, Scarlet 1) identifies online ontologies that can provide information about how these two concepts inter-relate and then 2) combines this information to infer their relation. To find online ontologies in which s appears, we use the *subject* and *object* of s as input to Scarlet, which returns a list of relations that exist between the two entities, along with information about the source ontologies from where the relations have been identified.

Matching the Statement Context with the Target Ontology. The statement context is matched to the target ontology to detect their shared concepts. Our approach is independent from the ontology matching technique to use. In our implementation, we perform the matching between the concepts' names using the Jaro-Winkler string similarity metric [3]. We define the function $e(G)$ to extract the set of nodes n_i that exist in the graph G. We use the matching to generate the intersection of the statement context and the target ontology: $e(C) \cap e(O_t) = \{n_i \mid n_i \in e(C) \wedge n_i \in e(O_t)\}$.

We developed a tool for visualizing how the context matches with the target ontology. This tool proved to be very useful during our experiments, as it makes understanding the matching process easier. It has customizable parameters that enable for example only to display the shared nodes with their connected entities up to a certain depth, and hide or show the target or online ontology. Matching nodes between the graphs are represented by star shaped nodes as shown in Figure 2, a visualization of the context of $< proposal, subClass, document >$ extracted from the online OntoSem ontology[1], and the SWRC target ontology[2].

Assessing Relevance Based on Overlap Analysis. We investigate a first naive approach based on the idea that the more overlapping the statement context and ontology are, the more relevant the statement is. The relevance confidence in this case is based on the ratio of the number of shared concepts, to the number of concepts in O_t, as calculated using the following formula:

$$conf_{overlap}(s, C, O_t) = \frac{|e(C) \cap e(O_t)|}{|e(O_t)|}$$

For example in Figure 2, with the string similarity threshold value of 0.96, there are 18 shared concepts between the context of $< proposal, subClass, document >$, and the SWRC ontology that includes 71 concepts. Thus the confidence of the overlap in this case is 0.2535 (i.e., $\frac{18}{71}$).

[1] http://morpheus.cs.umbc.edu/aks1/ontosem.owl

[2] http://kmi-web05.open.ac.uk:81/cache/6/98b/5ca1/94b45/7e29980b0f/dfc4e24088dffe851

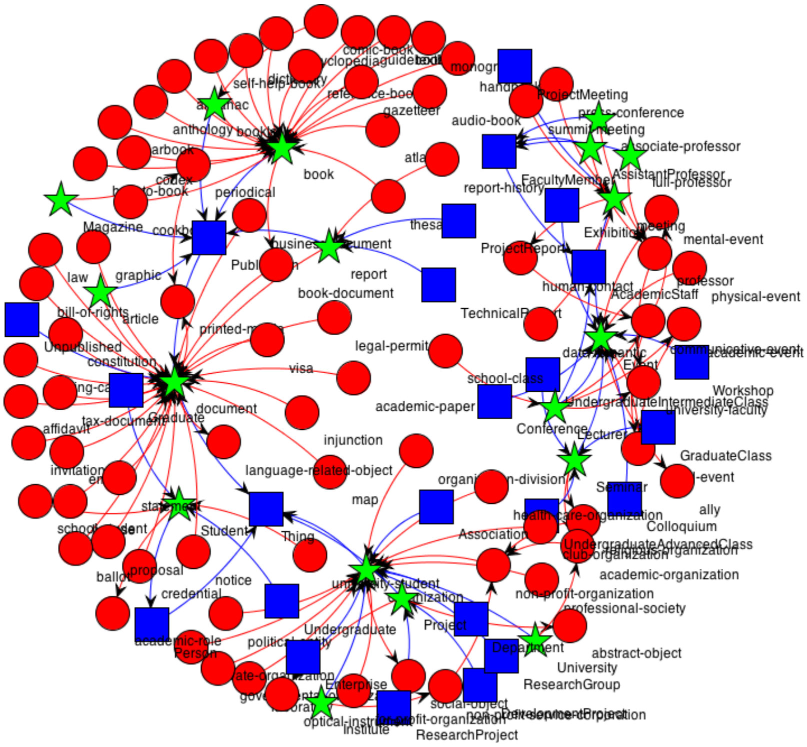


Fig. 2: Visualization of the overlap section between the OntoSem context of $< \ proposal, subClass, document \ >$ and the target SWRC ontology, where the star shaped nodes are shared, round nodes belong to the statement context, and square nodes belong to the target ontology.

The drawback of this approach is that it does not take into consideration how the ontological entities connect with each other, as it focuses on the number of shared nodes only, without any additional analysis. As a side effect, all the statements used in the context will be treated with the same relevance confidence. With big ontologies that are not domain focussed such as OntoSem or Cyc (www.cyc.com), it will cause the overlap technique to misjudge relevance. For example the statement $< capture, subClass, event >$ is extracted from OntoSem as well, but not relevant to add to the SWRC ontology. However, it has the same confidence value as the relevant statement $< proposal, subClass, document >$.

4 Pattern-Based Relevance Assessment

Given the limitations of the naive overlap technique, a more sophisticated approach is needed, which takes into account not only the overlap at the level of entity names, but also the way these entities are structured, giving a better indication of how the context fits in the ontology. Our preliminary work based on the analysis of some graph examples, highlighted the presence of patterns that reflect the relevance of statements [18]. Such *relevance patterns* identify specific structural conditions, supported by a confidence value.

In this section, we discuss next how we collect our experimental data (Section 4.1). Then we show how we refine the generation of the statements' context (Section 4.2), and finally present the relevance patterns (Section 4.3).

4.1 Gathering Experimental Data

To refine our initial approach and discover further relevance patterns, we needed a gold standard of statements assessed in terms of relevance that would serve as the basis of our analysis and tests. As such a gold standard does not exist yet, we created a set of statements evaluated by experts for relevance in three different domains: academic, music and fishery. The assessed statements played a major role in defining and discovering our relevance patterns.

Data collection of experts' evaluation was accomplished through a web interface. It supplied experts with a visualization of the target ontology, along with the options to select whether a statement is relevant, irrelevant or if relevance can not be judged from the given information ("Don't Know"). Experts were also given guidelines[3] describing the evaluation process, with some clarifications on what is meant by relevance supported by examples.

We use Evolva, our ontology evolution tool, to generate the set of statements to add to the ontologies of each domain. We parametrize Evolva to use online ontologies as a source of background knowledge, which link new concepts extracted from text to existing ones in the ontology in the form of statements.

In the academic domain, we randomly pick 30 news articles published on the Knowledge Media Institute's website (KMi). For the fishery domain, we extract 108 online web documents that include information about fishes and fishery stock. For the music domain, we extract 20 music blog pages that have on average seven blog post headers each. Table 1 lists the domains, the target ontology to evolve, the corpus used and the total number of statements suggested.

Domain	**Target Ontology**	**Corpus**	**Total** s
Academic	SWRC: http://kmi-web05.open.ac.uk:81/cache/6/98b/5ca1/94b45/7e29980b0f/dfc4e24088dffe851	KMi_News: http://news.kmi.open.ac.uk/	251
Fishery	Biosphere: http://kmi-web06.open.ac.uk:8081/cupboard/ontology/Experiment1/biosphere?rdf	Fishery_Website: http://fishonline.org/	124
Music	Music: http://pingthesemanticweb.com/ontology/mo/musicontology.rdfs	Music_Blog: http://blog.allmusic.com/	341

Table 1: Statements generation setup.

We apply a filter on the generated statements to 1) select only the taxonomic relations (cf. Section 3), and 2) remove generic relations, as our previous investigations show that statements linked to generic terms (e.g. *thing*, *object*, etc.) are

[3] http://evolva.kmi.open.ac.uk/experiments/statementrelevance/guidelines.php

mostly irrelevant [19]. We generate random selections of 100 statements in each domain to form the data-sets for experts to evaluate. We assign three different experts for each data-set, with two academic data-sets, given the expertise and availability of our evaluators in this area, and one data-set for each of the music and fishery domains.

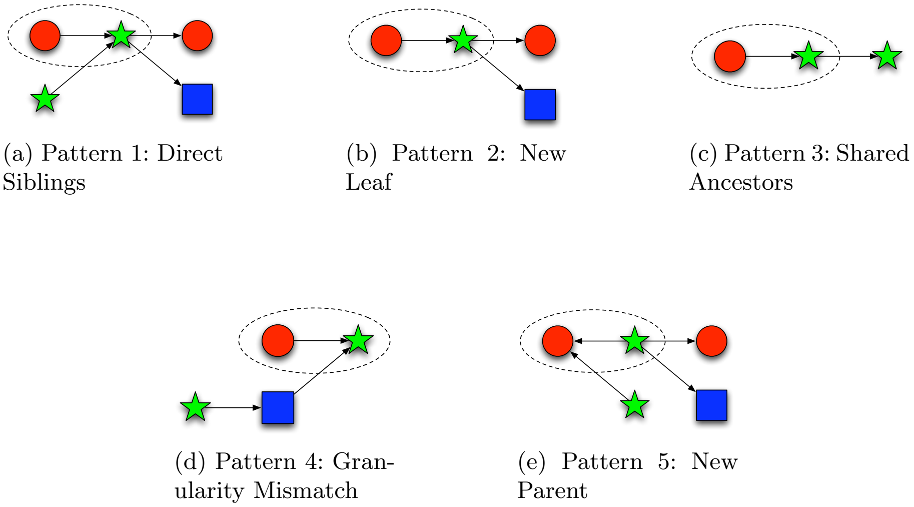
4.2 Statement Context Generation Revisited

A first improvement we introduce, following the analysis of the naive overlap approach, concerns the context generation. Instead of dealing with the ontology as a whole to define the context, we generate the context of the statement based on the surrounding entities of the statement up to a certain depth. This will help in focussing the usage of the statement by analyzing the close entities only. For that, we use $context(s, O, d) = C$, a recursive function that generates a sub-graph, formed of nodes related through taxonomic and other types of relations to the *subject* and *object* of s in O, up to a depth d (set to 1 in our implementation). This function is similar to the Prompt ontology view extraction [10] or some ontology modularization techniques [6]. The sub-graph generated forms the context C of the statement in the specified ontology.

4.3 Relevance Patterns

Another improvement comes at the level of introducing patterns for relevance detection. Relevance patterns are structural situations of interlinked nodes. When the surrounding entities in the matching graph around s trigger such patterns, a degree of relevance can be identified. This lifts the problem of the overlap that only matches the concepts' names, by providing further elements to analyze and hence a better relevance judgement. For example, a shared concept that is a sibling of an entity in s has a better influence on the relevance of s, than a shared concept which is not related to the elements of s. We create relevance patterns to detect such conditions and help deducing relevance. A clear visualization of the context and its intersection with the ontology (as shown in Figure 2), helped in identifying the relevance patterns that we discuss in this part. The statement relevance evaluation based on expert users in concrete domains contributed to spotting further undetected relevant statements, which improved our selection and definition of patterns.

Each pattern has specific *application conditions*, supported by a *confidence value*. Application conditions are defined in a way that makes the patterns mutually exclusive, thus facilitating their performance analysis. Based on our analyzed data, we identified five different patterns visualized in Figure 3, where the statement to assess is in the dashed oval, round and square nodes belong to the context and target ontology respectively, and star nodes are the ones shared by both. At a glance, Pattern 1 identifies direct shared siblings of the *subject* in s; Pattern 2 detects whether s introduces a new leaf to the ontology; Pattern 3 identifies shared ancestors of the *object* of s; and Pattern 4 detects shared siblings that occur at different levels of depth in the context and the target ontology. As per our analysis, shared ancestors (Pattern 3) gave better relevance indications then the other patterns, thus our application conditions are defined in a way to favour Pattern 3 over Patterns 1, 2 and 4. The last pattern, Pattern 5, is applied when s introduces a new parent in the target ontology.



(a) Pattern 1: Direct Siblings

(b) Pattern 2: New Leaf

(c) Pattern 3: Shared Ancestors

(d) Pattern 4: Granularity Mismatch

(e) Pattern 5: New Parent

Fig. 3: Relevance patterns for detecting the relevance of statement s represented in the dashed oval, star nodes denote shared concepts, round and square nodes belong to C and O_t respectively. The arrows depict sub-class relations.

Pattern 1: Direct Siblings. One core indication of relevance is when a new concept to add to the target ontology is surrounded by shared siblings between the statement context and target ontology. Shared siblings show that the concept in focus is missing in the target ontology, giving the statement adding it a high relevance. Pattern 1, shown in Figure 3a, detects shared siblings of the introduced concept. This is illustrated in Figure 4, where the statement in focus is $< tutorial, subClass, event >$ (in the dashed oval), in the context of the ISWC ontology[4]. This context shares with the SWRC target ontology the concepts *workshop* and $conference$. Those concepts show that the new concept *tutorial* is important to add to the SWRC ontology. *Application conditions*:

1. $\exists\, n_a \mid n_a \in e(C) \cap e(O_t) \wedge\ < n_a, subClass, object > \in C \cup O_t$
2. $\neg\exists\, n_b \mid n_b \in e(C) \cap e(O_t)\ \wedge C \models <\ object, subClass, n_b >$

Condition 1 ensures that the *subject* of s has direct siblings, while Condition 2 checks that there are no shared ancestors, thus prioritising Pattern 3. The *pattern confidence* formula is:

[4] http://annotation.semanticweb.org/ontologies/iswc.owl

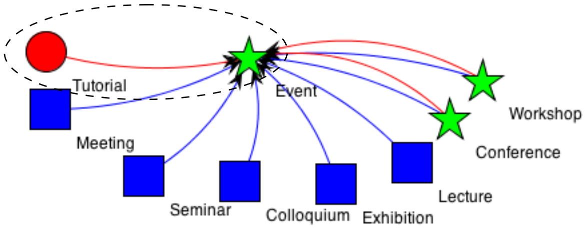


Fig. 4: Pattern 1 detected on $s = < tutorial, subClass, event >$, $C = ISWC.owl$ and $O_t = SWRC.owl$.

$$conf_{p1}(s, C, O_t) = \frac{|dSubC(object, C) \cap dSubC(object, O_t)|}{|dSubC(object, C)| - 1}$$

where $dSubC(n, G) = \{x_i \mid \ < x_i, subClass, n > \in G\}$, is a function to extract the direct sub-classes of a node in a graph. The confidence in this case is the ratio of the number of shared siblings (the numerator in the $conf_{p1}$ formula), to the total number of siblings in the context of s. If we apply the formula on $s_1 = < tutorial, subClass, event >$ in Figure 4, the confidence is:

$$conf_{p1}(s_1, ISWC, SWRC) = \frac{|\{Workshop, Conference\}|}{|\{Workshop, Conference, Tutorial\}| - 1} = 1$$

Even though Pattern 1 is one of the most intuitive patterns, it occurred on average only 11.25% of the statement cases (including relevant and irrelevant), in our four testing datasets.

Pattern 2: New Leaf. As Pattern 1 relies on the shared siblings of the *subject* of s, it will fail when the *object* of s is a leaf in the target ontology, because there will be no shared siblings in this case. This is where Pattern 2 (Figure 3b) called *New Leaf* comes in place, to detect the *subject* added as a new leaf to the target ontology. This pattern happened to be common in detecting relevant statements in the music domain, where many statements introduce new ontology levels, for example statements $< duet, subClass, performer >$ and $< quartet, subClass, performer >$[5] link *duet* and *quartet* as sub-classes to *performer*, an existing leaf in the target ontology. On average, this pattern occurred 10.25% of the cases in our tested statements. *Application conditions*:

1. $\neg\exists\, n_a \mid < n_a, subClass, object > \in O_t$
2. $\neg\exists\, n_b \mid n_b \in e(C) \cap e(O_t) \ \wedge C \models < object, subClass, n_b >$

Condition 1 ensures that the *object* of s does not have children (i.e it's a leaf in the target ontology), and Condition 2 confirms that the *object* doesn't have common ancestors. With the absence of close relatives (i.e., shared parents, ancestors and siblings), the confidence of the new leaf pattern is based on the overlap ratio of the target ontology cut to a specified depth around the *object* of s, and the context of s. Thus the *pattern confidence* formula is:

[5] Statement contexts extracted from: http://maciej.janik/test

$$conf_{p2}(s, C, O_t) = \frac{|e(C) \cap e(context(s, O_t, d))|}{|e(context(s, O_t, d))|}$$

Pattern 3: Shared Ancestors. The *Shared Ancestors* pattern (Figure 3c) relies on the condition that the relevance of a statement with respect to a target ontology increases if the shared *object* in s has shared ancestors between the target ontology and the context in which it is used. This situation was very common in the fishery domain, where Pattern 3 applied to 50% of the statements identified. For example, for the statement $< cod, subClass, fish >$, $fish$ has the ancestor *animal* in C, which is shared with O_t. This reflects a degree of common representations of animal species in online ontologies, where top levels in many ontologies tend to be more aligned than in the other domains. On average this pattern occurred in 20% of our analyzed dataset cases. *Application condition*:

1. $\exists\, n_a \mid n_a \in e(C) \cap e(O_t)\ \wedge C \models < object, subClass, n_a >$

The *pattern confidence* formula is:

$$conf_{p3}(s, C, O_t) = \frac{|aSupC(object, C) \cap e(O_t)|}{|aSupC(object, C)|}$$

based on the ratio of shared ancestors of the *object* of s, to the total number of ancestors of *object* in C. $aSupC(n, G) = \{x_i \mid G \models < x_i, superClass, n >\}$ extracts all the (direct and inferred) super-classes of a node n in a graph G.

Pattern 4. Granularity Mismatch. As ontologies are used in different application contexts, design decisions such as the level of granularity often vary from an ontology to another. This affects the performance of Pattern 1, which checks only the *direct* shared siblings of the *subject* in s. Pattern 4 (Figure 3d), called *Granularity Mismatch*, identifies such situations. With the highest occurrence of 41.75% of the cases, this pattern shows that granularity differences in concept representation when designing ontologies is a very common case. With our tests performed on the datasets, we have set this pattern to be applied as a last resort if Patterns 1, 2, and 3 are not detected. *Application conditions*:

1. $\neg\exists\, n_a \mid n_a \in e(C) \cap e(O_t) \wedge < n_a, subClass, object > \in C \cup O_t$
2. $\exists\, n_a, n_b \mid n_a \in e(C) \cap e(O_t)\ \wedge n_b \in e(C) \ominus e(O_t) \wedge n_b \in aSupC(n_a, C) \wedge n_b \in aSupC(n_a, O_t) \wedge object \in aSupC(n_a, C) \wedge object \in aSupC(n_a, O_t)$
3. $\neg\exists\, n_a \mid n_a \in e(C) \cap e(O_t) \wedge < object, subClass, n_a > \in C$

where Condition 1 is for ruling out the presence of Pattern 1, and Condition 2 checks for the presence of shared siblings (including the inferred ones) that fall at different levels in depth with respect to the *object* of s through a non-shared concept (i.e., a concept in the symmetric difference of C and O_t denoted by the symbol $\ominus$). Condition 3 rules out the presence of shared ancestors, for which Pattern 3 should be applied. This *pattern confidence* is:

$$conf_{p4}(s, C, O_t) = \frac{|aSubC(object, C) \cap aSubC(object, O_t)|}{|aSubC(object, C)|}$$

which takes the ratio of all the shared sub-classes of *object* in C and O_t, to the total number of all sub-classes of *object* in C. The function $aSubC(n, G)$ extracts all (direct and inferred) sub-classes of a concept n in G.

Pattern 5. New Parent. In cases where s links *subject* to *object* through a super-class relation, i.e. s is introducing *object* as a new parent to the ontology, Pattern 5 is applied (Figure 3e). There is indication of relevance in this case if *object* is a parent of other shared concepts between the statement context and the target ontology. The *application condition* of this pattern is solely limited to checking whether the type of relationship linking *subject* to *object* is super-class. The *pattern confidence* is based on the following formula:

$$conf_{p5}(s, C, O_t) = \frac{|aSubC(subject, C) \cap e(O_t)|}{|aSubC(subject, C)|}$$

The numerator in the fraction detects the number of shared concepts between C and O_t that are children of *subject* in C.

As per our tests, the number of statements with super-class relations is much lower than the sub-class relations. On average, only 16.75% of the total number of statements are super-classes. Furthermore, the percentage of relevance judgment correctness of this pattern is high in the four data-sets. Thus one pattern dealing with super-class relations proved to be enough for our domains.

5 Evaluation

In order to evaluate the discussed approaches, we analyze and compare the performance of the naive overlap approach, versus the pattern-based approach. We use the experts' statements evaluation data-sets in the three domains as the basis of our evaluation, which we present in this section.

5.1 Experiment Measures

Statement relevance being in many cases subjective, we made sure that each statement is evaluated by three experts per domain, having in total 12 experts for the four datasets. Based on the intuition that "relevance is not just an all-or-none matter but a matter of degree" [16], we use a measure to assess the overall relevance of each statement. To achieve this, we assign a score for each answer type from the experts: 1 for *relevant*, 0.5 for *don't know* and 0 for *irrelevant* (cf. Section 4.1). We use the sum of these values as an overall relevance score:

$$overall_{rel}(s, d) = \sum_{i=1}^{3} score(e_i, s, d)$$

where $overall_{rel}(s, d)$ is a function that returns the overall relevance score of a statement s in a data-set d, and $score(e_i, s, d)$ is the score given by expert e_i to s in d. For example, if the evaluation of a statement s is *relevant*, *relevant* and *don't know* by experts A_d, B_d and C_d respectively, the overall relevance value of

s is 2.5. We set two thresholds to handle the overall relevance measure outcome: a *relevance threshold* sets the limit above which s is considered relevant and an *irrelevant threshold* below which s is irrelevant. If the overall relevance value falls between the two thresholds, the relevance can not be determined in this case, as the experts are undecided.

Concerning the naive overlap and pattern-based algorithms output, a threshold is set to determine relevance based on the confidence value for each algorithm, i.e., when the overlap or a pattern is applied with a confidence degree higher than the specified threshold, the corresponding statement is classified as relevant, otherwise it is irrelevant. Given the different ways that each pattern calculates confidence, we use a separate threshold for each pattern (displayed in Table 2). As the goal of this experiment is to check the feasibility of the pattern-based approach, we empirically set the combination of thresholds that obtained the highest performance.

Threshold	**Academic-1**	**Academic-2**	**Fishery**	**Music**
User Relevance	2	2	2	2
User Irrelevance	1	1	1	1
Overlap	0.2	0.29	0.4	0.05
Pattern 1	0.2	0.1	1	0.08
Pattern 2	0.8	1	0.5	1
Pattern 3	1	1	0.01	0.05
Pattern 4	0	0.4	1	0
Pattern 5	1	0.4	0.05	0.02
Pattern 6	1	0.01	0.03	1

Table 2: Employed thresholds selected empirically to provide the highest average relevance and irrelevance F-measure in each data-set.

We use *Precision*, *Recall* and *F-measure* to evaluate the performance of the relevance algorithms. We define 4 sets $REL_{ed}, IRR_{ed}, REL_{ad}$ and IRR_{ad}: REL_{ed} is the set of all statements evaluated as relevant by the experts in dataset d; IRR_{ed} the set of irrelevant statements as judged by experts in d; REL_{ad} and IRR_{ad} the sets of relevant and irrelevant statements as classified by the algorithm a (i.e. pattern or overlap), in dataset d. We use the following formulas:

$$P_{rel}(d,a) = \frac{|REL_{ed} \cap REL_{ad}|}{|REL_{ad}|} \qquad R_{rel}(d,a) = \frac{|REL_{ed} \cap REL_{ad}|}{|REL_{ed}|}$$

where $P_{rel}(d,a)$ and $R_{rel}(d,a)$ compute the precision and recall of relevance respectively, in data-set d as judged by algorithm a. We use the usual *F-measure* computation based on precision and recall. In the case of irrelevance, the formulas are similar to the ones of relevance, but replaced with sets related to irrelevance (i.e. IRR_{ed} and IRR_{ad}).

5.2 Results

The main conclusion of our experiment, as shown in Table 3, is that the pattern-based approach performs better than the naive overlap approach. By simply comparing the precision and recall in each data-set, patterns are able to identify more correct relevant statements as classified by experts, with a better precision than then overlap approach. Overall, the overlap relevance F-measure is in the range of [7.41%, 58.06%], while the range is higher for the pattern-based relevance F-measure [43.75%, 69.05%]. In terms of irrelevance, the range is [60.87%, 85.71%] for the overlap approach, compared to the [74.74%, 92.48%] F-measure range using the pattern-based irrelevance detection. This is mainly due to the presence of large ontologies online that tend to highly overlap with target ontologies in general, and the fact that the overlap technique treats all statements coming from such ontologies equally, leading to lower precision and recall.

		Overlap		**Patterns**	
		Relevance	Irrelevance	Relevance	Irrelevance
Academic-1	Statements	18	82	13	87
	Precision	05.56%	83.72%	46.15%	91.95%
	Recall	11.11%	87.80%	66.67%	93.02%
	F-measure	07.41%	85.71%	**54.52%**	**92.48%**
Academic-2	Statements	15	85	16	84
	Precision	26.67%	81.18%	43.75%	90.00%
	Recall	25.00%	86.25%	43.75%	85.71%
	F-measure	25.81%	83.64%	**43.75%**	**87.80%**
Fishery	Statements	57	43	59	41
	Precision	47.37%	74.42%	55.39%	90.24%
	Recall	75.00%	55.17%	91.67%	63.79%
	F-measure	58.06%	63.36%	**69.05%**	**74.74%**
Music	Statements	57	43	35	65
	Precision	29.82%	81.40%	42.86%	83.08%
	Recall	73.91%	48.61%	65.22%	75.00%
	F-measure	42.49%	60.87%	**51.73%**	**78.83%**

Table 3: Evaluation results for relevance assessment.

Note that identifying irrelevant statements is equally important as identifying relevant ones. Moreover, our experiment shows that in most data-sets, the proportion of irrelevant statements is higher than the one of relevant statements. Thus having a high precision and recall on the bigger portion of the datasets (formed of irrelevant statements) reflects that the pattern-based approach would successfully act as a filter of irrelevant statements, reducing the workload on the user in the process of statement selection during ontology evolution.

To put the results in perspective, we rank the outcomes based on the confidence values of the overlap and pattern-based approaches, and compare them to the randomly ordered statements by Evolva (Figure 5). Due to the pattern specific threshold and confidence calculations, a direct ranking based on the confidence is not possible. Thus we normalize the pattern-based confidence values to a target unified threshold of 0.5, based on which we perform the ranking. As Figure 5 shows, the ranking based on the pattern technique groups relevant statements more towards the top of the list, meaning that ontology engineers could more confidently select most of the top statements, while safely discard most of the lower ranked ones. It is interesting to test in the future how these results would combine with other statement evaluation techniques (i.e., in terms of consistency, impact, etc).

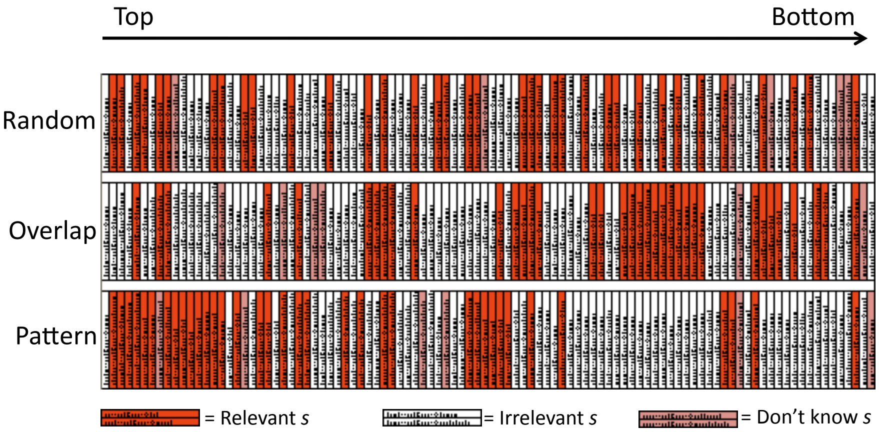


Fig. 5: Visualized ranking from left to right of 100 statements in the fishery domain, comparing the results of the random order on the top, overlap approach in the middle and pattern-based approach at the bottom.

6 Conclusion and Future Work

In this paper, we presented an approach towards the automatic assessment of the relevance of statements with respect to ontologies. This approach is based on the analysis of the context in which the statement occurs, and how it compares to the considered ontology. A set of relevance patterns in the graph merging the context with the ontology are identified, which provide indications of the level of relevance of the statement, by showing how the context fits in the ontology. The evaluation experiment demonstrates the feasibility of our pattern-based approach and how it outperforms a naive technique of measuring the overall overlap between the context and the ontology.

Even though the evaluation of our approach shows promising results, we identify potential improvements that will be part of our future work. Firstly, we plan to extend and identify further relevance patterns, in addition to test the combination of patterns rather than having them mutually exclusive. Secondly, instead of using the first online ontology returned by Scarlet as the statement context, we plan to devise a method to select the context with the highest relevance confidence. Thirdly, our future plans include a technique to automatically identify the relevance thresholds, which is crucial when our work is integrated in ontology evolution tools. One potential way to do so is to take the set of statements that have been lately added to the ontology under evolution as a base case of the threshold values calculation. The idea is that such statements are already assessed relevant by the user once added. Fourthly, a particular point to investigate is at the level of the user interaction with the tool. We foresee that our visualization tool that shows how the context of the statement matches with the target ontology, would be of added value to the user as a validation support of the assessed relevance.

References

1. P. D. Bruza and T. W. C. Huibers. A study of aboutness in information retrieval. *Artificial Intelligence Review*, 10(5):381–407, October 1996.
2. P. Cimiano and J. Volker. Text2Onto - a framework for ontology learning and data-driven change discovery. In *Proc. of (NLDB)*, 2005.
3. W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proc. of (IIWeb)*, 2003.
4. M. d'Aquin, C. Baldassarre, L. Gridinoc, M. Sabou, S. Angeletou, and E. Motta. Watson: Supporting next generation semantic web applications. In *Proc. of WWW/Internet*, 2007.
5. M. d'Aquin, E. Motta, M. Sabou, S. Angeletou, L. Gridinoc, V. Lopez, and D. Guidi. Toward a new generation of semantic web applications. *IEEE Intelligent Systems*, 23(3):20–28, 2008.
6. M. d'Aquin, A. Schlicht, H. Stuckenschmidt, and M. Sabou. Criteria and evaluation for ontology modularization techniques. In *Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization*, pages 67–89. Springer-Verlag, 2009.
7. Q. Ji, P. Haase, G. Qi, P. Hitzler, and S. Stadtmuller. RaDON-repair and diagnosis in ontology networks. In *Proc. of (ESWC)*, 2009.
8. D. Maynard, A. Funk, and W. Peters. SPRAT: a tool for automatic semantic pattern based ontology population. In *Proc. of (ICSD)*, 2009.
9. Stefano Mizzaro. Relevance: the whole history. *Journal of the American Society for Information Science archive*, 48(9):810–832, 1997.
10. N. F. Noy and M. A. Musen. Specifying ontology views by traversal. In *The Semantic Web (ISWC)*, pages 713–725. 2004.
11. K. Ottens, N. Hernandez, M. P. Gleizes, and N. Aussenac-Gilles. A Multi-Agent system for dynamic ontologies. October 2008.
12. I. Palmisano, V. Tamma, L. Iannone, T.R. Payne, and P. Doran. Dynamic change evaluation for ontology evolution in the semantic web. In *Proc. of (WI-IAT)*, 2008.
13. V. Pammer, L. Serafini, and M. Lindstaedt. Highlighting assertional effects of ontology editing activities in OWL. In *Proc. of the ISWC International Workshop on Ontology Dynamics (IWOD)*, 2009.
14. M. Sabou, M. d'Aquin, and E. Motta. Exploring the semantic web as background knowledge for ontology matching. *Journal on Data Semantics*, (XI), 2008.
15. E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 5(2):51–53, 2007.
16. D. Sperber and D. Wilson. *Relevance.* 1986.
17. R. J. Sternberg. *Metaphors of mind.* 1990.
18. F. Zablith, M. d'Aquin, M. Sabou, and E. Motta. Investigating the use of background knowledge for assessing the relevance of statements to an ontology in ontology evolution. In *Proc. of the ISWC International Workshop on Ontology Dynamics (IWOD)*, 2009.
19. F. Zablith, M. Sabou, M. d'Aquin, and E. Motta. Using background knowledge for ontology evolution. In *Proc. of the ISWC International Workshop on Ontology Dynamics (IWOD)*, 2008.