Generalizing precision and recall for evaluating ontology matching*

Marc Ehrig
University of Karlsruhe
ehrig@aifb.uni-karlsruhe.de

Jérôme Euzenat
INRIA Rhône-Alpes
Jerome.Euzenat@inrialpes.fr

Abstract

We observe that the precision and recall measures are not able to discriminate between very bad and slightly out of target alignments. We propose to generalise these measures by determining the distance between the obtained alignment and the expected one. This generalisation is done so that precision and recall results are at least preserved. In addition, the measures keep some tolerance to errors, i.e., accounting for some correspondences that are close to the target instead of out of target.

1 Problem statement

Ontology matching is an important problem for which many algorithms (see ISWC2005 proceedings) have been provided. In this short presentation we consider the result of matching, called alignment, as a set of pairs of supposedly equivalent entities $\langle e, e' \rangle$ from two ontologies O and O'.

In order to evaluate the performance of these algorithms it is necessary to confront them with ontologies to match and to compare the results based on some criterion. The most prominent criteria are precision and recall originating from information retrieval and adapted to the matching task. Precision and recall are based on the comparison of the resulting alignment A with another standard alignment R, effectively comparing which correspondences are found and which are not. Precision and Recall are the ratio of the number of true positive ($|R \cap A|$) on that of the retrieved correspondences ($|A|$) and those expected ($|R|$) respectively.

Definition 1 (Precision, Recall)**.** *Given a reference alignment R, the precision and recall of some alignment A are given by*

$$P(A,R) = \frac{|R \cap A|}{|A|} \text{ and } R(A,R) = \frac{|R \cap A|}{|R|}.$$

These criteria are well understood and widely accepted. However, they have the drawback that whatever correspondence has not been found is definitely not considered. As a result, they do not discriminate between a bad and a better alignment and they do not measure the user effort required to correct alignments.

*This work has been partially supported by the Knowledge Web European network of excellence (IST-2004-507482)

Indeed, it often makes sense to not only have a decision whether a particular correspondence has been found or not, but somehow measure the proximity of the found alignments. This implies that "near misses" are also taken into consideration instead of only the exact matches.

2 Generalizing precision and recall

As precision and recall are easily explained measures, it is good to extend them. This also ensures that measures derived from precision and recall (e.g., F-measure) still can be computed easily.

In fact, if we want to generalize precision and recall, we should be able to measure the proximity of alignment sets rather than the strict size of their overlap. Instead of taking the cardinal of the intersection of the two sets ($|R \cap A|$), the natural generalizations of precision and recall measure their proximity ($\omega(A,R)$).

Definition 2 (Generalized precision and recall)**.** *Given a reference alignment R and an overlap function ω between alignments, the generalized precision and recall of some alignment A are given by*

$$P_\omega(A,R) = \frac{\omega(A,R)}{|A|} \text{ and } R_\omega(A,R) = \frac{\omega(A,R)}{|R|}.$$

2.1 Basic properties

In order, for these new measures to be true generalizations, we would like ω to share some properties with $|R \cap A|$. In particular, the measure should be positive:

$$\forall A, B, \omega(A,B) \geq 0 \qquad \text{(positiveness)}$$

and should not exceed the minimal size of both sets:

$$\forall A, B, \omega(A,B) \leq min(|A|,|B|) \qquad \text{(maximality)}$$

Further, this measure should only add more flexibility to the usual precision and recall so their values cannot be worse than the initial evaluation:

$$\forall A, B, \omega(A,B) \geq |A \cap B| \qquad \text{(boundedness)}$$

Hence, the main constraint faced by the proximity is the following:

$$|A \cap R| \leq \omega(A,R) \leq min(|A|,|R|)$$

2.2 Designing overlap proximity

There are many different ways to design a proximity between two sets satisfying these properties. The most obvious one, that we retain here, consists of finding correspondences matching each other and computing the sum of their proximity. This can be defined as an overlap proximity:

Definition 3 (Overlap proximity)**.** *The overlap proximity ω between two sets A and R is defined by:*

$$\omega(A,R) = \sum_{\langle a,r\rangle \in M(A,R)} \sigma(a,r)$$

in which $M(A,R)$ is a matching between the elements of A and R and $\sigma(a,r)$ a proximity function between two elements.

The standard measure $|A \cap R|$ used in precision and recall is such an overlap proximity which provides the value 1 if the two correspondences are equal and 0 otherwise.

There are two tasks to fulfill when designing such an overlap proximity function:

- the first one consists of designing the correspondence matching M;
- the second one is to define a proximity measure σ on correspondences.

We consider these two issues below.

2.3 Matching correspondences

A matching between alignments is a set of correspondence pairs, i.e., $M(A,R) \subseteq A \times R$. However, if we want to keep the analogy with precision and recall, it will be necessary to restrict ourselves to the matchings in which an entity from the ontology does not appear twice, i.e., $|M(A,R)| \leq min(|A|,|R|)$. This is compatible with precision and recall for two reasons: (i) in these measures, any correspondence is identified only with itself, and (ii) appearing more than once in the matching would not guarantee that the resulting measure is bounded by 1 . The natural choice is to select the best match because this guarantees that this function generalizes precision and recall.

Definition 4 (Best match)**.** *The best match $M(A,R)$ between two sets of correspondences A and R, is the subset of $A \times R$ in which each element of A (resp. R) belongs to only one pair, which maximizes the overall proximity:*

$$M(A,R) \in Max_{\omega(A,R)}\{M \subseteq A \times R\}$$

As defined here, this best match is not unique. This is not a problem for our purpose because we only want to find the highest value for ω and any of these best matches will yield the same value.

Of course, the definition M and ω are dependent of each other, but this does not prevent from computing them. They are usually computed together but presenting them separately is clearer.

2.4 Correspondence proximity

In order to compute $\omega(A,R)$, we need to measure the proximity between two matched correspondences (i.e., $\langle a,r\rangle \in M(A,R)$) on the basis of how close the result is to the ideal one. Each element in the tuple $a = \langle e_a, e'_a, \rangle$ will be compared with its counterpart in $r = \langle e_r, e'_r\rangle$. If elements are identical, correspondence proximity has to be 1 (maximality). If they differ, proximity is lower, always according to the chosen strategy. In contrast to the standard definition of similarity, the mentioned proximity measures do not necessarily have to be symmetric. We will only consider normalized proximities, i.e., measures whose value ranges within the unit interval [0 1], because this is a convenient way to guarantee that

$$\sigma(A,R) \leq min(|A|,|R|)$$

From this simple set of constraints, we have designed several concrete measures:

symmetric is a simple measure of the distance in the ontologies between the found entities and the reference one;

edit measures the effort necessary to modify the errors found in the alignments;

oriented is a specific measure which uses different ω for precision and recall depending on the impact an error has on these measures, e.g., when one wants to retrieve instances of some class, a subclass of the expected one is correct but not complete, it thus affects recall but not precision.

3 Discussion

In order to overcome the lack of discrimination affecting precision and recall, we provided a framework properly generalising these measures (in particular, precision and recall can be expressed in this framework). We here presented the general principles that guide the design of such generalisations.

This framework has been instantiated and tested by hand against some examples. Due to space constraints, we refer to [1], but all the measures that we designed were having the expected results:

- they keep precision and recall untouched for the best alignment;
- they help discriminating between irrelevant alignments and not far from target ones;
- specialized measures are able to emphasize some characteristics of alignments: ease of modification, correctness or completeness.

The measures have been implemented in the Alignment API [2], which has been used for evaluation at the OAEI.

References

[1] M. Ehrig and J. Euzenat. Relaxed precision and recall for ontology matching. In *Proc. K-Cap workshop on Ontology integration*, pages 25–32, Banff (CA), 2005.

[2] J. Euzenat. An API for ontology alignment. In *Proc. 3rd international semantic web conference, Hiroshima (JP)*, pages 698–712, 2004.