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 measuring the distance between the obtained alignment and the expected one. This generalisation is made so that the results given by precision and recall are at worse preserved, but the measure keep some tolerance to errors (i.e., counting some correspondences are close to the target instead of out of target).

1 Problem statement

Ontology matching is an important problem for which many algorithms (see ISWC-2005 proceedings) have been provided. In this short presentation we will consider that the result of matching, called alignment, is a set of pairs of entities $\langle e, e' \rangle$ of two ontologies O and O' supposed to be equivalent.

In order to evaluate the performance of these algorithms it is necessary to confront them with ontologies to match and to compare them 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 to be retrieved ($|R|$) respectively.

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

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

These criteria are well understood and widely accepted. However, they are often criticized because they have the drawback of considering that whatever correspondence that has not be found is definitely not worth of attention. As a result, they do not discriminate between a bad and a better alignment and they do not measure user effort to adapt the alignment.

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 also "near misses" are taken into consideration instead of only the exact matches.

2 Generalizing Precision and Recall

Because precision and recall are easily explained measures, it is good to extend them. It also brings the benefit that measure derived from precision and recall can be computed from the generalisation.

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 the taking the cardinal of the intersection of the two sets ($|R \cap A|$), the natural generalization of precision and recall will measure of their proximity (ω).

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

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

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 not exceeding the minimal size of both sets:

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

The next interesting property is that this measure should only add more flexibility to the usual precision and recall so their values cannot be worse than the initial evaluation:

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

So the main constraint faced by the proximity is the following:

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

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

This is indeed a true generalization because, $|A \cap R|$ satisfies all these properties. One more property satisfied by precision and recall that we will not enforce here is symmetry:

$$\forall A, B, \omega(A, B) = \omega(B, A) \qquad \text{(symmetry)}$$

This guarantees that the precision and recall measure are true normalized similarities.

2.2 Designing Overlap Proximity

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

Definition 3 (Overlap proximity)**.** *A measure that would generalize precision and recall is:*

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

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

Again, 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 main problems to face for designing such a overlap proximity function:

- the first one consists of finding the correspondences to be compared M.
- the second one is to define a proximity measure on correspondences σ;

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 Overlap proximity under 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 us from computing them. They are usually computed together but it is better to present them separately.

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 from 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, proximity has to be one (maximality). If they differ proximity is lower, always according to the chosen strategy. Please note that in contrast to the standard definition of similarity, the mentioned proximity measures do not necessarily have to be symmetric. We will only consider here 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 (standard correspnds to standard precision and recall):

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

edit mesure 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 harm that the error can do to these measures (e.g., a subclass of the expected on is correct but not complete).

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 presented here the general principles that guide the design of such generalisations.

This framework has been instantiated and tested by hand against some examples. Two pages are to short to show this but all the measures that we designed were having the expected result :

- 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.

These are only results based on one example, but they show that the approach goes in the good direction. The measures will be implemented and run on a realistic test bench before the conference.