A modest proposal for data interlinking evaluation*

Jérôme Euzenat

INRIA & LIG (Jerome.Euzenat@inria.fr)

Data interlinking, i.e., finding links between different linked data sets, is similar to ontology matching in various respects and would benefit from widely accepted evaluations such as the Ontology Alignment Evaluation Initiative. Instance matching evaluation has been performed as part of OAEI since 2009. Yet, there has been so far few participants to IM@OAEI and there is still no largely acknowledged benchmark for data interlinking activites.

In order to secure more participation, we analyse the specificities of data interlinking and propose diverse modalities for evaluating data interlinking.

1 The problem

The data interlinking problem could be described in the same way as the ontology matching problem was:

Given two linked data sets, usually tied to their vocabularies (or ontologies),
Generate a link set, i.e., a set of sameAs links between entities of the two data sets.

We concentrate on sameAs links because these are by far the most important links to be retrieved when interlinking.

In terms of evaluation, the same kind of procedure as in ontology matching may be used by comparing the provided link set to a reference link set. Measures such as precision, recall or time and memory consumption may be used.

Given the size of the data, it is difficult to provide correct and more specifically complete reference link sets. The OAEI 2011 IM task solved the problem by using links already provided by data providers. This is valuable when these links are curated manually. However, the quality of these links may still be questioned.

2 Specific interlinking features

We present some features of the way data interlinking is performed nowadays, that distinguish it from ontology matching.

Blocking vs. matching Ontology matching, when confronted to large ontologies, cannot start upfront comparing instances with a similarity measure. The same applies to data interlinking. Hence many data interlinking procedures are designed in two steps:

blocking divides the sets of pairs of resources into subsets called blocks in which matching resources should be part;

*This work has been partially supported by the French National Research Agency (ANR) under grant ANR-10-CORD-009 (Datalift). A longer version of the proposal is available at `ftp://ftp.inrialpes.fr/pub/exmo/publications/euzenat2012b.pdf`

matching compares entities in the same block in order to decide if they are the same.

Since these two steps are clearly different, well identified, and it is possible to use different matching methods with different blocking techniques, it is useful to evaluate independently the capacities of these two techniques.

In particular, from a reference link set, it is possible to determine how many pairs in the link set a particular blocking technique misses (blocking recall) and how many non necessary pairs a blocking technique imposes to compare (blocking precision). Similarly, a matching technique could be evaluated with a given block structure if this is necessary: the evaluation can be achieved by comparing the part of the reference alignments which can be found from the given blocks.

Scalability Linked data has to deal with large amounts of data. Even if this is also the case in ontology matching, automatic data interlinking is really useful with large amounts of data. So, besides qualitative evaluation, it is critical to assess the behaviour of interlinking tools when data sizes get larger.

Learning Another effect of the size of linked data is that learning is more relevant, mainly for two reasons:

- The size of the data makes it difficult to study it for choosing the best approach and after extracting a training sample, much work remains to be done;
- The regularity of the data facilitates machine learning efficiency.

So, it is not surprising that learning methods are successful in data interlinking. This provides incentive to evaluate data interlinking techniques using machine learning. For that purpose, it is necessary to provide tests in which a part of the reference link set is provided as training set to the systems which have then to deliver a complete link set.

Instance and ontology matching Data interlinking is dependent on the vocabularies used in data sets. This vocabulary may be described by a schema or an ontology or not described explicitly. Some matchers may be specialised for some of these situations and it may be useful to recognise this by providing different evaluation tasks. In particular, it seems useful to test these configurations:

- without published vocabulary (or ontology),
- with the same vocabulary or, alternatively, with aligned vocabularies,
- with different vocabularies, without alignment (as in the IIMB data set of IM@OAEI).

3 Conclusions

In conclusion, it seems that in addition to or in combination with existing tasks provided in IM@OAEI, such benchmarks should consider including:

- scalability tests (retaining one tenth, one hundredth, one thousandth of the data);
- training sets for tools using machine learning;
- separating the evaluation of blocking and matching for users who specifically consider one of these aspects only;
- tests with no ontology, the same ontology and different ontologies.