SERIMI Results for OAEI 2011

Samur Araujo[1], Arjen de Vries[1], and Daniel Schwabe[2]

[1] Delft University of Technology, PO Box 5031, 2600 GA Delft, the Netherlands
{S.F.CardosodeAraujo, A.P.deVries}@tudelft.nl

[2]Informatics Department, PUC-Rio Rua Marques de Sao Vicente, 225, Rio de Janeiro, Brazil
dschwabe@inf.puc-rio.br

Abstract. This paper presents the results of SERIMI in the Ontology Alignment Evaluation Initiative (OAEI) 2011. We participate in the track IM@OAEI2011 (IMEI) of the campaign. We first describe the basic interlinking process and interlinking strategies in SERIMI, and then we present specific techniques used in this track. We conclude with a discussion of our results, and possible directions to improve SERIMI in future work.

Keywords: data integration, RDF interlinking, instance matching, linked data, entity recognition, entity search.

1 Presentation of the System

The interlinking of datasets published in the Linked Data Cloud (LDC) [1] is a challenging problem and a key factor for the success of the Semantic Web. Given the heterogeneity of the LDC, techniques aimed at supporting interlinking should ideally operate agnostic of a specific domain or schema.

In this context, ontology matching [2, 3, 4, 5, 6] and instance matching [9] are the two most-studied sub-problems of interlinking. The former refers to the process of determining correspondences between ontological concepts. The latter often refers to the process of determining whether two descriptions refer to the same real-world entity in a given domain. In this paper we focus on the problem of instance matching.

1.1. State, purpose, general statement

Our solution for the instance-matching problem is composed of two phases: the selection phase and the disambiguation phase. In the selection phase we apply traditional information retrieval strategies to generate a set of candidate resources for interlinking. For each instance **r** in a source dataset A, we extract its label (its identifier) and we search for instances in a target dataset B that may have a similar label. The problem that multiple distinct instances in B may share the same label is addressed in the second, disambiguation phase. Here, we attempt to filter among the instances found in B, those that actually refer to the same entity in the real world as **r**.

SERIMI uses existing traditional information retrieval and string matching algorithms for solving the selection phase; our contribution is the novel similarity

measure used in the disambiguation phase. This function is designed to operate even when there is no direct ontology alignment between the source and target datasets being interlinked. For example, SERIMI is able to interlink a dataset A that describes social aspects of countries with a dataset B that describes geographical aspects of countries. The SERIMI software is available for download as an interlinking tool at GitHub[1]. Fig. 1 and Fig. 2 show an overview of SERIMI's architecture.

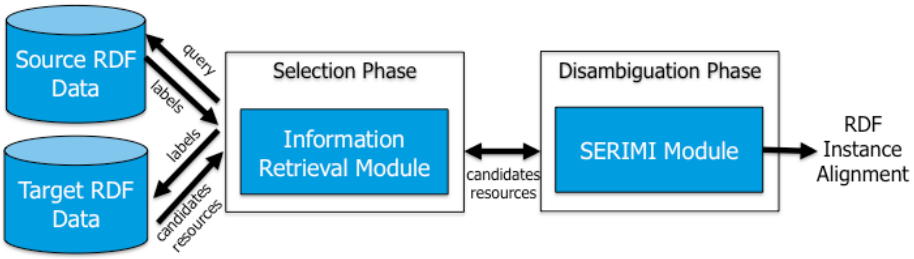

Fig. 1 – Overview of SERIMI's architecture.

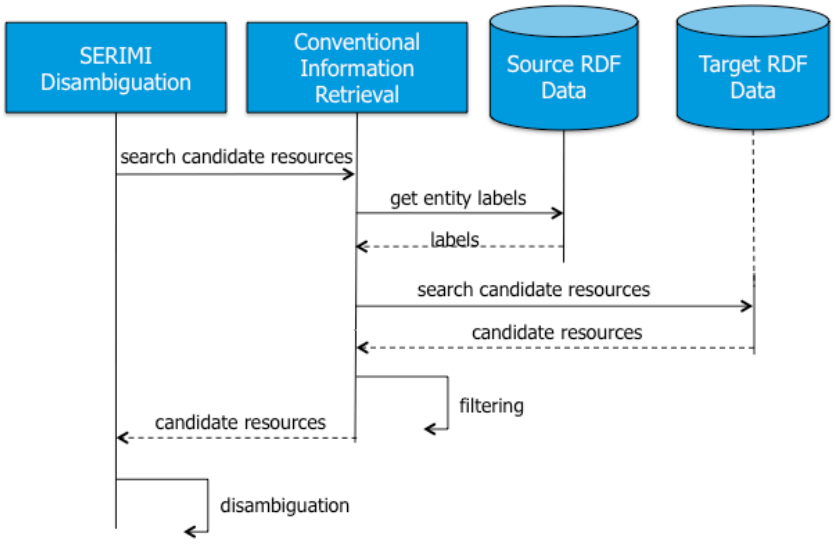

Fig. 2 – Overview of SERIMI's information flow.

1.2. Specific Technique Used

Fig.3 shows an overview of the SERIMI interlinking process.

[1] https://github.com/samuraraujo/SERIMI-RDF-Interlinking

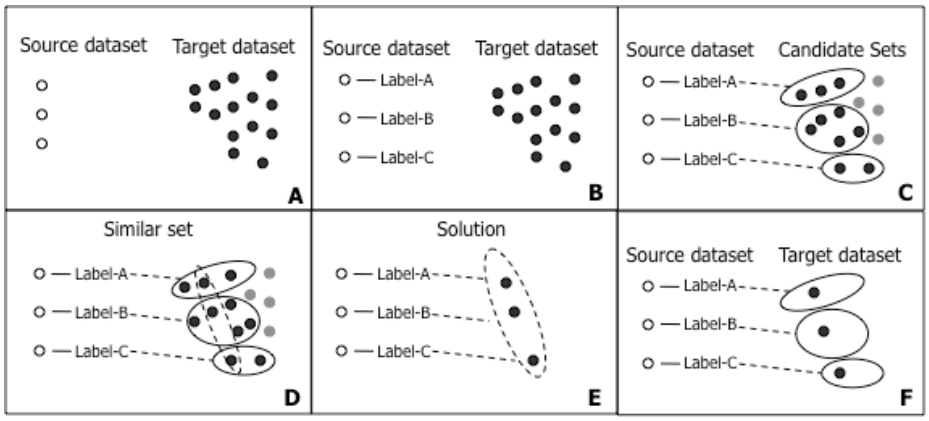

Fig. 3 – Overview of SERIMI interlinking process. (A) Given a source and target dataset and a set of source resources (instance of a class), (B) SERIMI obtains the label of these source resources and retrieves candidate resources from the target dataset that share a similar label. (C) For each source resource, SERIMI retrieves a set of candidate resources. (D) In order to disambiguate a set of candidate, SERIMI applies a novel function of similarity that selects the resources that are the most similar between all candidate sets (E). These selected resources are the solutions for the interlinking (F). The determination of this optimal cross section is a sophisticated process based on an underlying assumption that the source resources belong to a homogeneous class of interest (e.g. musician, drugs, country, etc.)

1.2.1. Selection Phase

In SERIMI's selection phase we first select the class of resources in the source dataset that we want to interlink. For each class (an rdfs:type object) found in the source dataset, SERIMI selects its instances and applies the approach below to select candidate resources in the target dataset.

Entity label property selection: in order to select resources in the target dataset that can match a specific source resource, we first select the labels that represent these source resources. We call *entity label properties*, the properties where these labels occur. We consider as entity label properties, all RDF predicates that have a literal, including numbers, but we eliminate long text values. We assume that we do not know the entity label properties in advance, and apply an automatic approach to select those. Considering that predicates with higher entropy are more discriminative than predicates with lower entropy, we select predicates with entropy $\Omega \geq \Omega_{threshold}$, where $\Omega_{threshold}$ is obtained by averaging the entropy of all predicates of the resources that we want to interlink. Those selected predicates compose the list of entity label properties. The procedure above is applied over the set of source instances selected for interlinking.

Pseudo-homonym resource selection: once we have determined the entity label properties in the source, we can use their labels for searching for resources in the target dataset that share the same or similar labels. We refer to the set of target candidate resources that share a similar label as the *pseudo-homonym set*. For each

resource to be interlinked, we use this source entity label for searching for candidate resources in the target dataset. We apply the same step described in the previous paragraph over the target dataset, to obtain the set of entity label properties in the target dataset. Then we search for the source entity label only on triples that contain such selected properties. For each source entity label, we normalize the string (by removing non alphanumeric characters), tokenize it, and then we apply a set of conjunctive Boolean queries (expressed in SPARQL) for retrieving target candidate resources. Afterwards, we select from the retrieved resources those with a maximum string similarity with respect to the searched source entity label. If the maximum score is below 70% we discard it. As a string matching algorithm, we used a variation of the RWSA[7] algorithm. By selecting only those resources with maximum relative similarity measure, we reduce the number of resources in the pseudo-homonym set, thereby improving the chance of true positive matches. If no resource is retrieved, then we select the next entity label property with the highest entropy and repeat the same procedure. This process ends forming a set of pseudo-homonym resources for each source resource. Then the task is to select from each set the resource(s) that one which is (are) more similar to the source resources. We do this selection during the disambiguation phase.

1.2.2. Disambiguation Phase

Pseudo-homonyms resource disambiguation: in some cases, a pseudo-homonym set may have instances of different classes or instances of the same classes that share the same label. As we do not know the class of the resources that we are looking for in the target dataset, we try to leverage this class of interest from the pseudo-homonym resources. Once the class of interest in determined, we can disambiguate the pseudo-homonym resources, by selecting the resources that belong to the class of interest. Notice that the concept of class of interest is understood as a set of attributes that instances may share in common. To solve this ambiguity problem, we propose an innovative model called *Resource Description Similarity*, or *RDS*. RDS uses the intuition that if we select two or more resources that are similar in the source dataset, and for each of them there is a set of pseudo-homonym resources in the target dataset, then the solutions for each pseudo-homonym set should be similar among themselves. In other words, the solution to the problem is the set of resources that are the most similar among pseudo-homonym sets, which implicitly defines the class of interest. The main requirement to apply this method is that we have to have at least two sets of pseudo-homonyms. Fig.3d and Fig. 4 illustrate this intuition.

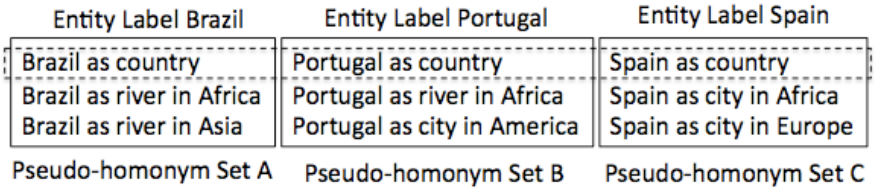

Fig. 3 – A simple example of pseudo-homonym sets for three labels that represent countries.

Disambiguating candidate resources: Given S as a set of all sets of pseudo-homonyms and $R \in S$, for each resource r in R, we generate a score $\delta = CRDS(r, R, S)$. As solution for a pseudo-homonym set R, we select all resources with a score $\delta \geq \delta_{threshold}$. Details about the function CRDS is given in [8].

1.3. Adaptations made for the evaluation

SERIMI operates directly over SPARQL Endpoints. For that reason, we have loaded the RDF version of the datasets Geonames, Freebase and NYTimes into an open-source instance of Virtuoso Universal server[2] installed on a local workstation, summing up millions of RDF triples. An exception was the DBPedia dataset, which we accessed online via its SPARQL endpoint. Then we run our method over these endpoints.

1.4. Link to the system and parameters file

SERIMI can be found at: https://github.com/samuraraujo/SERIMI-RDF-Interlinking

1.5. Link to the set of provided alignments (in align format)

The alignments for OAEI2011 campaign should be available at the official web-site: http://www.instancematching.org/oaei/imei2011.html. Alternatively, these can also be found at: https://github.com/samuraraujo/SERIMI-RDF-Interlinking.

2 Results

We now provide an analysis of the results obtained with SERIMI on the Instance Matching track (IM), on the subtask of data integration (Interlinking New-York Times Data) of the OAEI 2011 campaign. We use SERIMI to resolve RDF instance interlinking between the pairs of datasets, namely NYT-People vs. DBpedia, NYT-Locations vs. DBpedia, NYT-Organization vs. DBpedia, NYT-People vs. Freebase, NYT-Locations vs. Freebase, NYT-Organization vs. Freebase, NYT-Location vs. Geonames. Table 1 shows the results for each pair of dataset above.

[2] http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/

Table 1. SERIMI's precision and recall.

Dataset Pairs	Precision	Recall	F1
NYT-People vs. DBpedia	0.943	0.942	0.943
NYT-Locations vs. DBpedia	0.693	0.670	0.681
NYT-Organizations vs. DBpedia	0.887	0.870	0.878
NYT-People vs. Freebase	0.923	0.911	0.920
NYT-Locations vs. Freebase	0.922	0.904	0.913
NYT-Organizations vs. Freebase	0.921	0.895	0.908
NYT-Locations vs. Geonames	0.787	0.807	0.797

As we can see in Table 1, SERIMI performed quite well in all cases.

Although SERIMI was designed to perform over RDF datasets where the instances are organized in fine-grained homogenous classes, it performed quite well in average in the NYT scenario, where the instances are grouped in four heterogeneous classes (organization, locations, people, and, descriptors). This heterogeneity on the data was the main reason that we obtained a poor performance in the pair NYT-Locations vs. DBpedia. The NYT-Locations instances are very ambiguous and the class is too heterogeneous, representing cities, countries, lakes, etc. For instance, this class does not distinguish a city from a neighborhood, and for that reason SERIMI's disambiguation phase could not work properly. Even if the results are far from perfect, in this specific case of NYT-Locations vs. DBpedia, the results show that the proposed disambiguation phase leads to an approximated gain of 64% over the accuracy of the selection phase on its own (which produced a F1 of 44%).

Regarding SERIMI's selection phase, the set of boolean queries used in SERIMI failed in selecting resources where the label of the source resource was an abbreviation or acronym of the target resource, or vice-versa (e.g. source: "Minnesota Mining & Manufacturing Co", target: "3M Company"). SERIMI has also failed due to distinct string formatting between the source and target datasets labels. For instance, SERIMI selected the resource labeled "Jackson Michael" for the searched label "Jackson, Michael", instead of the resource labeled "Michael Jackson", which was the correct answer. These two problems are known issues in the literature, and we intent to investigate them as future work.

We noticed that the use of ontological knowledge could have improved the precision in the Geonames case, since the instances of the class NYT-Location have the properties longitude and latitude that also occur in the Geonames, with exactly the same values. For instance, the use of both label and longitude in the search process of the selection phase would have improved the precision for this case. Nevertheless, as we aim to provide a fully automated approach agnostic of ontology, we did not consider the use of ontological knowledge as a solution. However, the use of two attributes in the search process will be investigated as future work.

We observed that the fully automatic approach for detecting the entity labels using entropy performed satisfactorily in all dataset pair compared. No wrong label was selected in our evaluation.

We noticed that the accuracy of the NYTimes alignment is quite good, since it was manually curated, but it is not perfect. We encountered a few inconsistencies, and evidences of incorrect alignment in almost all pair of datasets. This fact led SERIMI to reach a non-optimal performance in this challenge. Below we show some examples of the inconsistence, incorrect and arbitrary judgment found in the reference alignment.

Label: Expedia Inc

•http://rdf.freebase.com/ns/en.expedia (reference alignment)

•http://rdf.freebase.com/ns/en.expedia_inc (SERIMI)

Label: USG Corporation

•http://dbpedia.org/resource/United_States_Gypsum (reference alignment)

•http://dbpedia.org/resource/USG_Corporation (SERIMI)

Label: Kirov Ballet

•http://rdf.freebase.com/ns/en.mariinsky_ballet (reference alignment)

•http://rdf.freebase.com/ns/en.kirov_ballet (SERIMI)

SERIMI took 40 minutes in average to compute the interlinking of an individual pair of dataset when it was performed under a controlled environment. In the case of DBPedia, its performance varied a lot due to the remote server avaiability.

3 General Comments

RDF instance matching is a challenging problem and the community has only recently started to develop a systematic framework to evaluate approaches to tackle this problem: the IMEI track of the OAEI initiative. We have however also encountered some problems in applying this framework to understand our results.

1. The accuracy of the reference alignment is critical point for the participants. Its quality prevents participants try to improve their precision, in cases where the reference alignment lacks in accuracy, or can be considered quite arbitrary, since there are dual interpretation in the alignment. We wasted a plenty of time to realize that the reference alignment was not 100% accurate, since we trusted on it beforehand. Therefore, we propose the organizers to warn the participants of lack of accuracy in the reference alignment, or whether possible, to publish some statistics about its accuracy.
2. Since DBPedia and Freebase contain a lot of duplicate entities associated to different URIs (e.g. http://rdf.freebase.com/ns/en.expedia_inc and http://rdf.freebase.com/ns/en.expedia), we propose the organizers to take this into consideration while computing the precision and recall of the participant results. Two participants may send distinct alignment results that are both correct.

Finally, we see an opportunity to ease the participation in the track. The preparation of the datasets is a non-trivial task, especially because they are large and available in different formats. Since all participants face the same problem here, it would be huge improvement whether the OAEI initiative could provide a SPARQL endpoint for all datasets mentioned in the challenge. All participants would work

exactly over the same datasets, consequently increasing the credibility of the results. RDF database engines exist that allow text search via SPARQL endpoint with a quite high performance; and when used properly can support a large amount of requests, as demanded by a challenge of this scale.

4 Conclusion

In this paper, we present the results of SERIMI in OAEI 2011 Campaign's IMEI-DI track. We have presented the architecture of SERIMI system and described specific techniques used in this campaign. SERIMI matches instances between a source and target datasets, without prior knowledge of the data, domain or schema of these datasets. SERIMI solves the instance-matching problem in two phases. In the selection phase, it uses traditional information retrieval and string matching algorithms to select candidate resources for interlinking. In the disambiguation phase, it uses a novel approach to measure similarity between RDF resources and disambiguate the resources. The results illustrates that SERIMI can achieve good accuracy in instance matching track.

References

1. Bizer, C., Heath, T. and Berners-Lee, T. (2009) Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst., 5 (3). pp. 1-22.
2. Tejada, S.; Knoblock, C. A.; and Minton, S. (2001). Learning object identification rules for information integration. Information Systems 26(8): 607–633.
3. L. A. P. P. Leme, M. A. Casanova, K. K. Breitman, and A. L. Furtado. (2008). Evaluation of similarity measures and heuristics for simple RDF schema matching. Technical Report 44/08, Dept. Informatics, PUC-Rio.
4. Isaac A., Meij L. V. D., Schlobach S., and Wang S. (2007). An empirical study of instance-based ontology matching. In Proceedings of the 6th international semantic web and 2nd Asian conference on Asian semantic web conference (ISWC'07/ASWC'07), Springer-Verlag, Berlin, Heidelberg, 253-266.
5. K. K. Breitman, D. Brauner, M. A. Casanova, R. Milidi, A. Gazola, and M. Perazolo. (2008). Instance-Based Ontology Mapping. In Fifth IEEE Workshop on Engineering of Autonomic and Autonomous Systems (ease 2008), Belfast, Northern Ireland, pp. 67-74.
6. N. Choi, I.-Y. Song, and H. Han,. (2006). A survey on ontology mapping," ACM SIGMOD Record, vol. 35, no. 3, pp. 34-41.
7. Branting, L. K. (2003) A Comparative Evaluation of Name-Matching Algorithms. ICAIL '03, June 24-28, 2003, Edinburgh, Scotland, UK.
8. Araújo S., Hidders J., Schwabe S., and Vries A. P. de. (2011). SERIMI - Resource Description Similarity, RDF Instance Matching and Interlinking. CoRR, vol. abs/1107.1104.
9. Kopcke H., Thor A., and Rahm E. (2010). Evaluation of entity resolution approaches on real-world match problems. In Proceedings of the 3rd VLDB Endowment. Pp. 484-493.