STRIM Results for OAEI 2015 Instance Matching Evaluation

Abderrahmane Khiat[1] , Moussa Benaissa[1] and Mohammed Amine Belfedhal[2]

[1] LITIO Laboratory, University of Oran1 Ahmed Ben Bella, Oran, Algeria
abderrahmane_khiat@yahoo.com , moussabenaissa@yahoo.fr
[2] Evolutionary Engineering and Distributed Information Systems Laboratory (EEDIS), Djillali LIabes University of Sidi Bel Abbes, Algeria
Mohammed.belfedhal@gmail.com

Abstract. The interest of instance matching grows everyday with the emergence of linked data. This task is very necessary to interlink semantically data together in order to be reused and shared. In this paper, we introduce STRIM, an automatic instance matching tool designed to identify the instances that describe the same real-world objects. The STRIM system participates for the first time at OAEI 2015 in order to be evaluated and tested. The results of the STRIM system on instance matching tracks are so far quite promising. In effect, the STRIM system is the top system on SPIMBENCH tracks.

Keywords: String-Based Similarity, Instance Mapping, Instance Matching, Linked Data, Web of Data, Semantic Interoperability, Semantic Web.

1 Introduction

The current Web, contains *documents* in various formats (PDF, Excel, HTML file, etc.) *connected* by *hypertext links*, also known as the *Web of Documents*. Note that, we mean by document, if the content is unstructured and not exploitable i.e. the semantic the content is not presented. Contrary to data, where the content is structured and exploitable i.e. the semantic of the content is presented using RDF for example.

The *inadequacy* of the *Web of Documents* resides in the fact that the *content* of *these documents* is probably *unstructured* and *its semantic is not presented* which means that it *is not exploitable* and *untreatable automatically* in different applications, either by the *machine* or by *expressive queries*.

In order to deal with these problems, and especially for the re-use and sharing of content, the *transition* from the *document* to the *data* is very necessary. This involves the *use of semantic web technologies* in order to (a) publish structured data on the Web, (b) make possible, the links between data from one data source to data within other data sources. These two points are very important to ensure *semantic interoperability*.

These data should be expressed using the RDF language (Resource Description Framework [see section 2.1]) to achieve the two major points that we have mentioned before in order to enable the semantic interoperability, which led to the emergence of the Web of Data. The data presented and structure in this form (RDF) can be easily interpreted by the computer, re-used in applications and easily linked with other data. If the data are easily linked the computer can work through relationships with other data and in this case the interoperability will be ensured. Other advantages of Linked Data among others are: improving the data quality, less human intervention and processing and short development cycles (quicker and save time).

With the effort of Linked Data Community to publish existing open license datasets as Linked Data on the Web and interlink things between different data sources, the Web of Linked Data has seen remarkable increase over the past years. In terms of statistics, in 2007, over 500 million RDF triples published on the web with around 120,000 RDF links between data sources. In 2010, over the 28.5 billion triples, in 2011 over 31.6 billion triples and in 2013 over 50 billion triples. According to these statistics, the Linked Data seems to be increasing drastically [6].

Linked Data, by definition, links the instances of multiple sources. A common way to link these instances to others is to use the owl:sameAs property. The enormous volume of data already available on the web and its continuity to increase, requires techniques and tools capable to identify the instances that describe the same real-world objects automatically.

With the OAEI evaluation campaign which distinguishes between matching systems that have participated in the category of ontology matching and those that have participated in the category of instance matching, these tools can be tested and evaluated. However, few systems[3] [10] namely InsMT, LogMap and RiMOM-IM have participated to test their performance at instance matching track of OAEI 2014.

In this paper we deal with two challenges namely:

1. How to link the distributed and heterogeneous data which are described with instances.
2. How to deal with the huge volume of data available on the web and its continuity to increase [14].

Indeed, the Solution to this problem consists to provide techniques and tools capable to identify the instances that describe the same real-world objects automatically.

In this paper, we describe the STRIM system in order to resolve automatically the instance matching problem. The STRIM system, extracts first all information about the two instances to be matched and normalizes them using NLP techniques. Then, it applies edit distance as a matcher to calculate the similarities between the normalized information. Finally, the approach selects the equivalent instances based on the maximum of shared information between the two instances.

The STRIM system has participated for the first time at OAEI evaluation campaign and it provides very good results in terms of precision, recall and f-measure.

The rest of the paper is organized as follows. First, the preliminaries on instance matching are presented in section 2. In the Section 3, we presented the related work on instance matching systems that participated in Instance Matching Track of OAEI 2014. In the Section 4, we describe our system by giving a detailed account of our approach. The experimentation results is presented in Section 5. The Section 6 contains concluding remarks and sets directions for future work.

---

[3] The declaration of OAEI 2014 evaluation campaign about instance matching systems Again, given the high number of publications on data interlinking, it is surprising to have so few participants to the instance matching track, although this number has increased.

2 Preliminaries

In this section, we present the basic notions related to Instance Matching by explaining the linked data and instance matching definition.

2.1 Linked Data Principle

The Linked Data consist to relate data with typed links across the Web using URIs, HTTP and RDF. The linked Data principles are defined by Tim Berners-Lee in 2007 [11]. These principles are as follow:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful RDF information.
- Include RDF statements that link to other URIs so that they can discover related things.

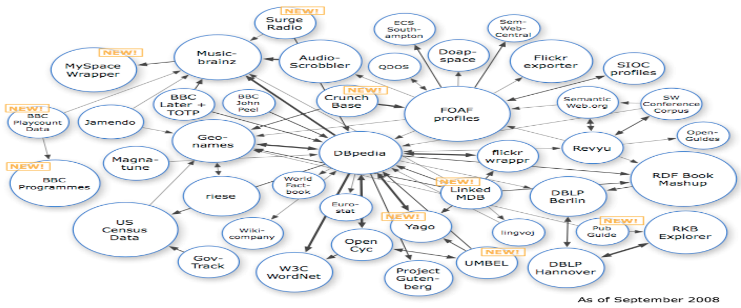

Fig. 1: Linked Data

The Linked Data (Fig. 1), by definition [12], links the instances of multiple sources. A common way to link the instances in these sources to others, is the use of the owl:sameAs property. Instance matching is required to interlink these data.

2.2 RDF Language

These data should be expressed using the RDF language (Resource Description Framework) to achieve the two major points that we have mentioned before in order to enable the semantic interoperability. The RDF language is a graph model to formally describe Web resources and metadata, in order to allow automatic processing of such descriptions [13][1][2]. An RDF file thus formed is a labeled directed multi-graph. Each triplet corresponds to a directed arc whose label is the predicate, the source node is the subject and the target node is the object.

2.3 Instance Matching Definition

The Instance Matching (Fig.2) is a process that starts from collections of data as input and produces a set of mappings (simple or complex) between entities of the collections as output [5].

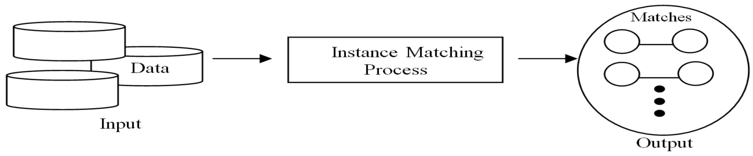

Fig. 2: Instance Matching Process

2.4 Entity Resolution Notion

Definition [5]: Let D1 and D2 be represent two datasets, each one contains a set of data individuals T_i which are structured according to a schema O_i. Each individual $I_i j$ T_i describes some entity w_j.

Two individuals are said to be equivalent I_j I_k if they describe the same entity w_j = w_k according to a chosen identity criterion. The goal of the entity resolution task is to discover all pairs of individuals ($I_1 i$, $I_2 j$) — $I_1 i$ T_1, $I_2 j$ T_2 such that $w_1 i$ = $w_2 j$.

In the context of linked data, datasets D_i are represented by RDF graphs. Individuals I_i T_i are identified by URIs and described using the classification schema and properties defined in the corresponding ontology O_i.

Example of Instance Matching We give below an example that shows how to link data from DBpedia with other data sources using the owl:sameAs property.

<http://dbpedia.org/resource/Berlin>
owl:sameAs
<http://sws.geonames.org/2950159>

3 Related Work

We present and discuss in this section the major works relevant to instance matching that participated at OAEI 2014 evaluation campaign. Only two systems succeed to finish all sub-tracks of instance matching track of OAEI 2014, namely RiMOM-IM and our previous InsMT system. We cite in exhaustive way only the instance matching systems that have participated in OAEI 2014 evaluation campaign and which are the object of comparison with our system STRIM.

1) LogMap [7]: The LogMap family participated with four different versions namely LogMap, LogMap-Bio, LogMap-C and LogMapLite in OAEI 2014. Only two versions (LogMap and LogMap-C) of them have participated at instance matching track. The LogMap-family system is a highly scalable ontology matching system with built-in reasoning and inconsistency repair capabilities. The two versions of LogMap systems identifies mappings between instances. The LogMap and LogMap-C systems finish only the first sub-track of instance matching of OAEI 2014 which is Identity Recognition.

2) RiMOM-IM [9] [3] [4]: is an acronym of Risk Minimization based Ontology Mapping Instance Matching. The principle of RiMOM-IM is to construct a document from the dataset by extracting the instances information. Then, it uses cosine-similarity to compare documents. The version of RiMOM-IM system that participated in OAEI 2014 for instance matching is developed based on ontology matching system RiMOM with some changes in objective. The objective of RiMoM-IM is to solve the challenges in large-scale instance matching by proposing a novel blocking method.

3) InsMT(L) [8]: is an acronym of Instance Matching at Terminological (Linguistic) level. InsMT(L) has participated for the first time in OAEI 2014. The principle of InsMT(L) is to use String-based algorithms (and WordNet as matcher at linguistic level) in order to calculate similarities between instances after the annotation step. The similarities calculated by each matcher are aggregated using the average aggregation strategy after a local filtering. Finally InsMT(L) system operates a global filtering in order to identify the alignment. The InsMT(L) system shows good results in terms of recall on different sub-tracks of instance matching of OAEI 2014. The InsMT(L) system finishes all sub-tracks of instance matching of OAEI 2014 which is Identity Recognition and Similarity Recognition.

4) Other Approaches:

There are several other instance matching approaches like HMatch [18], FBEM [17], SILK [16] and the works proposed in [15] which are not covered by this paper due to minor importance for our approach. These instance matching approaches have not participated in instance matching track of OAEI 2014. With respect to these approaches, we did not take them in consideration because we do not have their official results for the experimental protocol of OAEI in 2014.

As we have mentioned before, with the high number of publications about interlinking approaches only a few systems have participated at OAEI 2014. These systems are LogMap, RiMoM-IM and our previous InsMT(L) system.

4 STRIM: STRing based algorithm for Instance Matching

We summarize the process of our approach to provide a general idea of the proposed solution. It consists in the following successive phases:

4.1 Extraction and Normalization

The system extracts from each individual Ii P_1 m_1; P_2 m_2,... a set of information m_1, m_2, ... using different properties P_1, P_2, ... Then, NLP techniques are applied to normalize these infrmation. In particular, three pre-processing steps are performed: (1)

case conversion (conversion of all words in same upper or lower case) (2) lemmatization stemming and (3) stop word elimination. Since String based algorithm is used to calculate the similarities between information, these steps are necessary.

4.2 Similarity Calculation

In this step, the system calculates the similarities between the normalized informations using edit distance as string matcher. Our system selects the maximum similarity values calculated between different informations by edit distance. If two informations are the same (based on maximim similarity values) the counter is incremented to 1, etc.

4.3 Identification

Finally, we apply a filter on maximum counter values in order to select the correspondences which mean that the selected correspondences (equivalent individuals) are those who share maximum informations.

5 Experimentation

In this section, we present the results (Tab. 1) obtained by running our STRIM system on instance matching tracks of OAEI 2015 evaluation campaign.

Table 1: The Results of STRIM System

System	Track	Precision	F-measure	Recall
STRIM	sandbox val-sem task	0.90	0.95	0.99
LogMap	sandbox val-sem task	0.99	0.92	0.86
STRIM	mainbox val-sem task	0.91	0.95	0.99
LogMap	mainbox val-sem task	0.99	0.92	0.85
STRIM	sandbox val-struct task	0.99	0.99	0.99
LogMap	sandbox val-struct task	0.99	0.90	0.82
STRIM	mainbox val-struct task	0.99	0.99	0.99
LogMap	mainbox val-struct task	0.99	0.90	0.82
STRIM	sandbox val-struct-sem task	0.91	0.95	0.99
LogMap	sandbox val-struct-sem task	0.99	0.88	0.79
STRIM	mainbox val-struct-sem task	0.91	0.95	0.99
LogMap	mainbox val-struct-sem task	0.99	0.88	0.79

Only two systems have participated at SPIMBNNCH tracks namely the LogMap and STRIM systems. The SPIMBENCH consists of the following three different tasks: val-sem, val-struct and val-sem-struct. Each task has two tests (1) the Sandbox which contains two datasets in small scale and (2) the Mainbox which contains two datasets in large scale. The goal of three tasks consists to determine when two OWL instances describe the same Creative Work. However, the three tasks have been produced by altering a set of original data. In other words, the datasets of the val-sem task have been produced by using value-based and semantics-aware transformations. For the datasets of the val-struct task have been produced by using value-based and structure-based transformations. Finally the datasets of the val-sem-struct task have been produced by using value-based, structure-based and semantics-aware transformations.

We have evaluated the results of STRIM system based on the results obtained on Mainbox tests. The reason is that these tests were blind (i.e. the reference alignment is not given to the participants) during the evaluation of Instance matching systems by the OAEI evaluation campaign. On the other side, in the Sandbox tests, the reference alignment were available to help the instance matching systems to configure theirs parameters.

Regarding F-measure results, the STRIM system seems to achieve the best results before the LogMap system. The F-measure is always more than 95%. we can remark that STRIM system achieve high recall for the three tasks. It always equal to 99%.

* As conclusion, the result proves that our STRIM system is effective and efficient for the three tasks of SPIMBENCH track of OAEI 2015.

6 Conclusion

In this article, we have introduced STRIM, an instance matching system which consists in identifying the instances that describe the same real-world objects automatically. Our approach is useful, especially when the instances contain terminological information.

The STRIM system is composed of three steps: the first step consists in extracting and normalizing all information about the two instances to be matched. The second step consists in applying an edit distance as a matcher to calculate the similarities between the normalized information. The final step, consists in selecting the equivalent instances based on the maximum of shared information between the two instances.

The STRIM system has participated for the first time at OAEI evaluation campaign and it provides very good results in terms of precision, f-measure and recall at Instance Matching of OAEI 2015.

As future perspective, we attempt to apply STRIM to link data on could computing environment and develop other approaches.

References

1. J. Euzenat and P. Shvaiko Ontology Matching, Second Edition, Springer-Verlag, Heidelberg, pp. 1-511, 2013.
2. M. Ehrig Ontology Alignment: Bridging the Semantic Gap, Semantic Web And Beyond Computing for Human Experience 4, Springer, pp. 1-250, 2007.
3. Z. Wang, X. Zhang, L. Hou,Y. Zhao, J. Li, Y. Qi and J. Tang RiMOM results for OAEI 2010, In The Proceedings of the 4th International Workshop on Ontology Matching co-located with the 9th International Semantic Web Conference (ISWC 2010), pp. 195-202. CEUR-WS.org, Vol. 689, Shanghai, China, 2010.

4. J. Li, J. Tang, Y. Li and Q. Luo RiMoM: a Dynamic Multistrategy Ontology Alignment Framework, Journal IEEE Transactions on Knowledge and Data Engineering, vol. 21, No. 8, pp. 1218-1232, 2009.
5. A. Ferrara, A. Nikolov, J. Noessner and F. Scharffe Evaluation of instance matching tools: The experience of OAEI, Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 21 pp. 49-60, 2013.
6. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 2009.
7. E. Jimnez-Ruiz, B. C. Grau, W Xia, A. Solimando, X. Chen, V. Cross, Y. Gong, S. Zhang and A. Chennai-Thiagarajan LogMap family results for OAEI 2014. In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (ISWC 2014), October 20, pp. 126-134. CEUR-WS.org, Trentino, Italy, 2014.
8. A. Khiat, M. Benaissa, InsMT / InsMTL results for OAEI 2014 instance matching. In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (ISWC 2014), October 20, pp. 120-125. CEUR-WS.org, Trentino, Italy, 2014.
9. C. Shao, L. Hu and J. Li, RiMOM-IM results for OAEI 2014. In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference (ISWC 2014), October 20, pp. 149-154. CEUR-WS.org, Trentino, Italy, 2014.
10. Z. Dragisic, K. Eckert, J. Euzenat, D. Faria, A. Ferrara, R. Granada, V. Ivanova, E. Jimnez-Ruiz, A. O. Kempf, P. Lambrix, S. Montanelli, H. Paulheim, D. Ritze, P. Shvaiko, A. Solimando, C. Trojahn, O. Zamazal, B. C. Grau, Results of the Ontology Alignment Evaluation Initiative 2014. In Proceedings of the 9th International Workshop on Ontology Matching co-located with the 13th International Semantic Web Conference, pp. 61-104. CEUR-WS.org, Trentino, Italy, 2014.
11. Tim Berners-Lee. Linked Data - Design Issues, 2006. http://www.w3.org/DesignIssues/LinkedData.html. 7, 26, 82.
12. R. Parundekar, C.A. Knoblock, J.L. Ambite, Linking and building ontologies of linked data. In: Proceedings of the 9th International Semantic Web Conference (ISWC 2010). Shanghai, China, 2010.
13. G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation, http://www.w3.org/TR/rdf-concepts/, 2004.
14. P. Shvaiko and J. Euzenat. Ten challenges for ontology matching. In R. Meersman and Z. Tari, editors, On the Move to Meaningful Internet Systems: OTM 2008, volume 5332 of Lecture Notes in Computer Science, pp. 11641182. 2008.
15. D. Engmann and S. Mamann. Instance matching with COMA++. In Proceedings of Datenbanksysteme in Business, Technologie and Web(BTW 07), pages 2837, 2007.
16. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data, In The Proceedings of 8th International Semantic Web Conference (ISWC 2009), A. Bernstein, D. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, editors, The Semantic Web - ISWC 2009, volume 5823 of Lecture Notes in Computer Science, pp. 650665. Springer Berlin / Heidelberg, 2009.
17. H. Stoermer and N. Rassadko. Results of okkam feature based entity matching algorithm for instance matching contest of oaei 2009, 2009.
18. S. Castano, A. Ferrara, S. Montanelli, and D. Lorusso. Instance matching for ontology population. In Proceedings of the 16th Italian Symposium on Advanced Database Systems, pages 121132, 2008.