ISSN 2348 – 9928
IJAICT Volume 6, Issue 4, April 2019 Doi:01.0401/ijaict.2015.07.05 Published on 05 (4) 2019

A APPROACH FOR CLASS BASED MATCHING OVER HETEROGENOUS DATA SETS

Ms. S. Revathi
PG Scholar,
Sasurie Academy of Engineering,
Coimbatore, Tamilnadu, India

Ms. N. Poongothai
Assistant Professor,
Sasurie Academy of Engineering,
Coimbatore, Tamilnadu, India

Abstract– ***In heterogeneous datasets while used matching instances state-of-the-art instance matching approaches do not perform well. From the core operation on direct matching these drawbacks should be derived. The direct matching involves a direct comparison between instances from the source dataset and instances in the target dataset. If the overlap between the datasets is small direct matching is not suitable. The big aim of this survey is resolving this problem by proposing a new paradigm called class-based matching. The class of interest is defined as a class of instances from the source dataset. The class-based matching is defined as a set of candidate matches retrieved from the target. The candidate refining process could be done by filtering out those that do not belong to the class of interest. For this type of refinement, only data in the target dataset is used which states that no direct comparison between source and target is involved. Based on the public benchmarks in the difficult matching tasks this approach greatly improves the quality of state-of-the-art systems.***

Keywords– ***Data Integration, Class-Based Matching, Direct Matching, Instance Matching, Semantic Web.***

I. INTRODUCTION

1.1 RDF

In a web a large number of datasets has been available which internally contains more initiatives such as Linking Open Data. In a general graph-structured data model, RDF1 is widely used in publishing Web datasets. An entity termed an instance is represented via triples format. They are subject; predicate; object statements. Predicates captures attributes and objects capture values of an instance respectively

1.2 OWL2

In addition to RDF, OWL2 is another standard language for knowledge representation. It should be widely used for capturing the "same-as" semantics of instances. Using OWL scheme data providers can make explicit call. The two distinct URIs actually refer to the same real world entity. The entity resolution and instance matching is the task of establishing the same-as links.

1.3 Semantic-Driven Approach:-

Semantic-driven approaches use specific OWL semantics they termed as explicit owl same as statements. It allows the same as relations to be inferred via logical reasoning.

1.4 Data Driven Approach

This approach is opposed to the semantic-driven approach which derives same-as relations mainly vary with respect to the selection and weighting of features. The data-driven approaches are built upon the same paradigm of direct matching (DM).If the two instances have many attribute values in common they are considered the same. If the sufficient overlap between instance representations is occurred means they can produce only high quality results. If the Overlap is small in heterogeneous datasets means the same instance represented in two distinct datasets may not use the same schema. In an instance matching across heterogeneous datasets, direct matching alone cannot be expected to deliver high quality results .Contributions [1] provides detailed analysis of many datasets and matching tasks. These tasks greatly vary in their complexity. There are difficult tasks with a small overlap between datasets that cannot be effectively solved using state-of-the-art direct matching approaches. The big Aim of these tasks is to propose a direct matching in combination with [2] class-based matching (CBM).

1.5 Class Based Matching

In this paper following class notion should be employed. A class is considered as a set of instances where each instance in this set must share at least one feature in common with any other instance in this set.CBM aims to purify the set of candidates by filtering out those that do not match the class of interest. This matching is however not assumed that the class semantics are explicitly given. Direct matching at the class level is possible between the source (e.g. Nations) and target (e.g. Countries). CBM is based on the idea that if the instances have some features in common means they are known to form a class and their matches should also form a class in the target dataset i.e. matches should also have some features in common.

By computing the subset of candidates the correct matches can be found in this members have the most features in common. According to the direct matching method these candidates may form source instances. The class of interest should be created by the class they form correspond to the source instance, i.e. the instances found by CBM belong to a class, which matches the class of interest. During the candidate selection step the source and target instances are compared.

In a class-based matching, only data from the target dataset is needed. This is the main difference with direct matching, which compares the source and the target data. Ref[3] evaluated this approach called SERIMI using data from OAEI 2010 and 2011 based on the two reference benchmarks in the field.

Class-based matching achieved competitive results with direct matching method. Most importantly if the direct matching's performance was bad the improvements are complementary, achieving good performance. The simple combination of the DM and CBM this approach greatly improve the results of existing systems.

1.6 Instance Matching

Instance matching across datasets involves similarity functions, thresholds and comparable attributes. By using a matching scheme they should be captured. While the majority of approaches use a flat representation of instances based on attribute values, other features might be applied. RDF-based graph-structured model used to accommodate different kinds of structured data. The combination of direct Matching and class based matching produces good quality. In SERIMI, those combined components are treated as black boxes that yield two scores considered independent. SERIMI multiplies, normalizes and on and off these scores to obtain a value in form of 0s and 1s.

II. EXISTING SYSTEM

2.1 Boolean Matching

Simple Boolean matching could be used to generate candidates in this work. The Boolean queries are constructed using tokens extracted from candidate labels. Standard preprocessing is applied to lowercase tokens and to remove stop words. These queries derive candidates, which have values that share at least one token with the values of the corresponding source instance.

Advantage

- This method is primarily twisted towards quickly finding all matches, i.e. high recall,

Disadvantage

- May produce many incorrect candidates.
- Other techniques known in literature [3] achieve higher precision compared with Boolean matching.

Matching features and Instance features are derived from flat attributes. Structure information (e.g. relations between RDF resources) [7], [8], [9] or semantic information extracted from ontologies. Object Coref [5] for instance, exploits the semantics of OWL properties namely OWL: INVERSEFUNCTION ALPROPERTY and OWL: FUNCTIONALPROPERTY. Also, the combination of instance-level and schema-level features have been explored by PARIS [1], which jointly solves the problem of instance and schema matching.

If the SERIMI targets the heterogeneous scenario means no structure, semantic or schema information is available in the worst case. It is based on a simple flat representation, where instances are captured as a set of attribute values. This representation is employed for single instances as well as for class of instances, which are needed for CBM.

2.1 Similarity Functions

The choice of similarity functions depends on the nature of the features. For an string, character-based, token-based and document-based functions (e.g. cosine similarity) were used. In addition with syntactic information, special similarity functions have also been used to exploit different kinds of (lexical) semantic relatedness [10], [11].In addition to this dimension, a simple approach should be pursued where only tokens are employed. The new problem of CBM involves comparing sets of instances for this we propose a set-based similarity function that take the token overlaps between sets into account.

2.2 Matching Schemes

Based on approaches relying on a flat representation of instances, i.e., attribute values, the matching schemes con1tain the similarity functions, thresholds and comparable attributes. Comparable attributes are either computed via automatic schema matching or assumed to be manually defined by experts [12]. Then, techniques with different degrees of supervision are employed for learning the scheme.

In 2011, Knofuss+GA proposed [13],it is an unsupervised approach that applys a genetic algorithm for learning process. In 2011 SIFIet.al proposed[2] and In 2007 OPTrees et.al proposed [6] which represent supervised approaches that learn the schemes from a given set of examples. Others approaches such as 2011-Zhishi.links [12] and 2010-RIMON [14] and Song and

2009-Heflin [4] assume matching schemes that for the most part, were manually engineered, i.e., the similarity functions and thresholds were defined manually. They focus on the problem of learning the best comparable parts.

The above solutions focus on direct matching which is oppose to the, class-based matching does not rely on a complex scheme. It uses a special similarity function specifically designed for this matching task. The problem of finding the threshold is cast as the one of detecting outliers, for this unsupervised solution should be proposed.

For 2011 data, SERIMI also greatly improves the results of recently proposed approaches (2011-PARIS [1] and 2011-SIFI-Hill [2]). Compared to the best system proposed in OAEI 2011, SERIMI achieved the same performance. However, while that system leverages domain knowledge and assumes manually engineered mappings, our approach is generic, completely automatic and does not use training data.

Overall this solution can be characterized as an unsupervised, simple, yet effective solution, which employs a novel class-oriented similarity function, matching technique and threshold selection method to exploit the space of class related features never studied before. He Fast-Join method described in 2011-[15] studied the problem of string similarity join that finds similar string pairs between two string sets. This concept focused on the entire problem of matching two distinct instances of data. An instance should be understood as a structured representation of a real world entity, containing specific semantic attributes that cannot be trivially reduced to a set of tokens. Therefore, representative direct matching approaches for instance matching where particularly selected in our evaluations.

III. PROPOSED SYSTEM

The process of instance matching performed by SERIMI .It focuses on the problem of instance matching across heterogeneous datasets. Since the direct overlap at the level of predicates (or values) between instances may be too small to perform matching in the heterogeneous setting. This proposes class-based matching. Class-based matching can be applied in combination with direct-matching, on top of the candidate selection step. The equal-weight strategy to give a greater emphasis on commonalities. This is because the goal of class-based matching is to find whether some instances match a class., For deciding whether an instance belongs to a class or not, the common features are thus, by definition, more crucial. Not only that, the special treatment of common features also makes sense when considering that common features are scarcer. That is, the number of features shared by all instances in a class is typically much smaller than features that are not.

3.1 Advantages of Proposed System

- SERIMI reported the best performance in the benchmark that they participated compared to the other state of art approaches.
- SERIMI achieved considerable performance gain for the life science collection.
- SERMI present a type of features which represent a large part of all features used.
- Hence, processing was much faster without them.
- In general, the results suggest that all proposed features are useful as they contributed to higher accuracy.
- CBM produced similar scores for all candidates. For this DM performed better because the overlap between the source and target instances is sufficiently high to identify the correct matches.
- Overall, the results show that SERIMI achieved the best accuracy results.
- Further, there is room for improvement as SERIMI so far neither uses training data nor exploits domain knowledge.
- Training data could be exploited to fine tune the threshold (as done by SIFI).
- SERIMI yields superior quality.

IV. CONCLUSION

This survey proposes an unsupervised instance matching approach. This type of matching combines direct-based matching with a novel class-based matching technique to know Same as relation over heterogeneous data. This method using two public benchmarks namely OAEI 2010 and 2011. The matching achieved good and competitive quality compared to representative systems focused on instance matching over heterogeneous data.

References

[1] F. M. Suchanek, S. Abiteboul, and P. Senellart, “Paris: Probabilistic alignment of relations, instances, and schema,” Proc. VLDB Endowment, vol. 5, no. 3, pp. 157–168, 2011.

[2] J. Wang, G. Li, J. X. Yu, and J. Feng, “Entity matching: How similar is similar,” Proc. VLDB Endowment, vol. 4, no. 10, pp. 622–633, 2011.

[3] A. Arasu, S. Chaudhuri, and R. Kaushik, "Learning string transformations from examples," Proc. VLDB Endowment, vol. 2, no. 1,pp. 514–525, 2009.

[4] D. Song and J. Heflin, "Automatically generating data linkages using a domain-independent candidate selection approach,"in Proc. Int. Semantic Web Conf., 2011, pp. 649–664.

[5] W. Hu, Y. Qu, and X. Sun, "Bootstrapping object coreferencing on the semantic web," J. Comput. Sci. Technol., vol. 26, no. 4, pp. 663–675, 2011.

[6] S. Chaudhuri, B.-C. Chen, V. Ganti, and R. Kaushik, "Exampledriven design of efficient record matching queries," in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 327–338.

[7] S. Melnik, H. Garcia-Molina, and E. Rahm, "Similarity flooding: A versatile graph matching algorithm and its application to schema matching," in Proc. Int. Conf. Data Eng., 2002, pp. 117–128.

[8] P. Shvaiko and J. Euzenat, "A survey of schema-based matching approaches," in J. Data Semantics IV, vol. 4, pp. 146–171, 2005.

[9] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Discovering and maintaining links on the web of data," in Proc. Int. Semantic Web Conf., 2009, pp. 650–665.

[10] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," Comput. Linguistics, vol. 32, no. 1,pp. 13–47, 2006.

[11] X. Han and J. Zhao, "Structural semantic relatedness: A knowledge-based method to named entity disambiguation," in Proc.48th Annu. Meeting Assoc. Comput. Linguistics, 2010, pp. 50–59.

[12] X. Niu, S. Rong, Y. Zhang, and H. Wang, "Zhishi.links results for oaei 2011," in Proc. 6th Int. Workshop Ontology Matching, 2011,pp. 220–227.

[13] A. Nikolov, M. d'Aquin, and E. Motta, "Unsupervised learning of link discovery configuration," in Proc. 9th Int. Conf. Semantic Web:Res. Appl., 2012, pp. 119–133.

[14] Z. Wang, X. Zhang, L. Hou, Y. Zhao, J. Li, Y. Qi, and J. Tang, "Rimom results for oaei 2010," in Proc. 4th Int. Workshop Ontology Matching, 2010, pp. 195–202.

[15] J. Wang, G. Li, and J. Feng, "Fast-join: An efficient method for fuzzy token matching based string similarity join," in Proc. Int. Conf. Data Eng., 2011, pp. 458–469.