ISSN 2278-3091
Volume 8, No.1.4, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse2481.42019.pdf
https://doi.org/10.30534/ijatcse/2019/2481.42019

AIBD-A Data-Driven Model for Ontology Property Alignment

Mansir Abubakar, Hazlina Hamdan, Norwati Mustapha, Teh Noranis Muhd Aris
Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang
Selangor Darul Ehsan
abubakar.mansir@auk.edu.ng, {hazlina, norwati, noranis}@upm.edu.my

ABSTRACT

The unprecedented increase in the web data sources provides numerous opportunities to the organizations that require data manipulation in their daily businesses. However, organizations are faced with a serious challenge associated to this rate of increase as a result of distinct data sources. This challenge is known by the community of data science as heterogeneity. Many researchers proposed feasible ways of addressing data heterogeneity using both traditional and classical approaches. Ontology instance based matching is one of popular approach of addressing data heterogeneity and improvement of interoperability between different sources of data. However, one serious issue with instance based approach is the inability to matching all potential property attributes in generating alignments between these data sources using their URIs. This paper addresses this challenge by proposing a property alignment and similarity generation algorithms that effectively used generated training samples to align properties and also generate similarity using both the training samples and the generated alignments. The experimental result obtained on benchmark data sets demonstrates significant improvement of our method over state of the art approaches in minimizing the heterogeneity and generation of complete alignment between interrelated data sources.

Key words: Data Interoperability, Knowledge Sharing, Linked Data, Ontology Matching, Semantic Web

1. INTRODUCTION

Knowledge sharing over linked data is an important aspect for many web applications. Most importantly, if the data to be shared across two or more RDF graphs, or to be replicated in the same source of data (i.e. self-match) [1] may link to different domains, such as health, publication, people, production, agriculture and so on. For instance, with the growing number of online shopping websites, there is a need for accurate and all-inclusive price unification among similar products so that a shopping site (such as AMAZON) can be able to precisely identify similar products alongside their prices. This can be achieved through an appropriate technique of aligning ontology properties where the semantics resides. Figure 1 depicts the alignment abstraction.

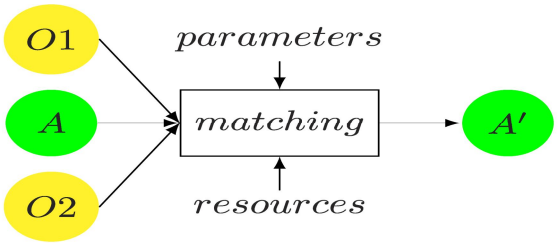

Figure 1: Abstract View of Ontology Alignment Process through Resources Matching

Ontology matching is the process of finding relationships in different ontologies of the same domain to represent a single real-world entity. Although in some situations conflict may occur only when the ontologies represent the same real word. Ontology matching process can be seen as a function *f*, in which from a two given ontologies *O1* and *O2*, the input alignment *A*, a set of parameters *P* and a resources r returned the matching result as A'. The outcome of this process is popularly known as *ontology alignment* [2]. The process can be represented as:

$$A' = f(O1, O2, A, P, R) \quad (1)$$

Most of the approaches to ontology alignment relies heavily in matching the basic schema of ontologies to generate alignments [3] [4] and so on. Many kind of these approaches were reviewed in the work of [5] in their article of review on ontology matching applications and challenges. Recently, researchers began to address ontology matching problems at instance level rather than the traditional schema matching approach. [6] proposes an instance matching benchmark in the era of linked data. The main goal of this survey is to present the state-of-the-art instance matching benchmarks for Linked Data. It describes in details how ontology matching plays important role for the success of linked data. However, these approaches are characterized by many challenges, especially in reducing the amount of heterogeneity [7] due to rapid expansion in the volume and variety of web data. Property matching also depend on intuitions in many approaches [8]. Therefore, feasible solution is required to match ontologies at instance level with due consideration to

all necessary ontology's property (predicate) which is the core aspect of N-Triples. In this paper, we proposed (1) a data-driven alignment algorithm for generating set of aligned properties using the generated training samples to facilitate knowledge sharing over linked data environment. The aligned properties will be used subsequently to extract semantic features within the properties of ontologies. And (2) a dual phase (learning and retrieval) algorithms for similarity generation that will execute the property matching based on the extracted feature values, which can scale to the growing volume of linked data. This method would facilitate an effective generation of alignments that can support efficient knowledge sharing over linked data environment.

The main contribution of this paper is the proposed property alignment algorithm and similarity generation algorithm that can aid the generation of complete alignment and to ensure scaling to the linked data expansion. The algorithm uses the generated training samples to align properties according to their type semantics. This is an important improvement in data and information matching process that was always based on intuition rather than data-driven in most of the best existing matching approaches.

2. RELATED WORKS

Instance matching is one of the ontology matching techniques proposed to be the future of ontology matching and liked data success. In this technique, the target is on instance that is the sub entities or subclasses of an ontology [9]. They are characterized by matching between two ontologies that share the same set of property's instances or individuals. Many existing works addresses semantic heterogeneity through ontology alignment [10]. Most of these approaches concentrated heavily on schema matching to generate alignments, for example the work in [11] use Markov model approach to improve the similarity flooding method for ontology alignment. However, this approach lack ability to generate complete alignment as few or no properties of the instances is considered for the matching. In [12], a methodology for knowledge representation using ontology concepts is proposed. this approach employ bond graph model (BGM) to produce a function structure of equipment related to fault propagation in part-component levels in order overcome the problems of heterogeneity and inconsistency in maintenance records, which are attributable to abbreviations, noisy data, non-generic data structures, and ambiguous technical words in textual maintenance records. However, the methodology has limited concern on the ontology properties rather effectively match ontology concepts (schema). In an attempt to perform property alignment, [13] proposes an iterative framework for instance matching called RIMOM-IM. The main concept of this framework is to utilize the distinctive and available matching information to improve the efficiency and control the propagation of error in generating alignments. Although, this framework demonstrated performance as intended but the error propagation control has limitation when more properties are the target for a matching. The work of [6] identified some feasible approach to solving property alignment limitations in their survey. they explained in detail the principles of benchmark design for instance matching systems, discuss the dimensions and characteristics of an instance matching benchmark, provide a comprehensive overview of existing benchmarks, as well as benchmark generators, discuss their advantages and disadvantages, as well as the research directions that should be exploited for the creation of novel benchmarks, to answer the needs of the Linked Data paradigm. Following the survey, [14] proposed a property based solution to ontology matching by modeling instance matching as a binary pattern classification problem to solve the challenge ontology matching in relation to e-learning educative content in a Knowledge Society context. To sum up the investigation conducted on the literature, it is found that reducing heterogeneity to the minimal level is a continual process that requires non-stop research attention. This is due to the fact that the volume and variety of web data is non-deterministic as everyone can be data publisher, especially with the emergence of social media and other modern publishing platforms.

3. PROPERTY ALIGNMENT (PA) APPROACH

The core idea behind the property alignment is to address the impact of semantic heterogeneity due to the differences of its sources. One important function of the training set generator [15] is to train the property aligner. In this paper, we assumed training sets already generated. The essence is to detect and address the effect of irregular data in the initial point of the generated training set, which pose a challenge in the process of learning. The possible encounter to this issue is to design a property alignment algorithm in order to accommodate these kinds of risen challenges. A property alignment is a set of semantically related or equivalent property (attributes) pairs. Property alignment is inadequately defined in most of the instance matching systems and always based on intuition rather than being data-driven [13]. In principle, Property alignment has a role similar to type alignment but with additional functionality such as its implication in the whole instance matching functionality, more specifically, in its effectiveness and complexity.

In this paper, alignment algorithm is designed to accommodate the two requirements stated above in a domain-independence way. This algorithm will carry out alignment between a set of attributes in ontology O_1 and ontology O_2 represented in the two property tables P_1 and P_2. The input to this algorithm is not the whole data sets but the training sets generated using our training set generating algorithm, termed UTSG. This algorithm utilizes information

signals to achieve high recall without affecting the precision in all data domain. The two single-type RDF graph data serialized as logical property tables, these logical table used as working example to illustrate the property alignment.

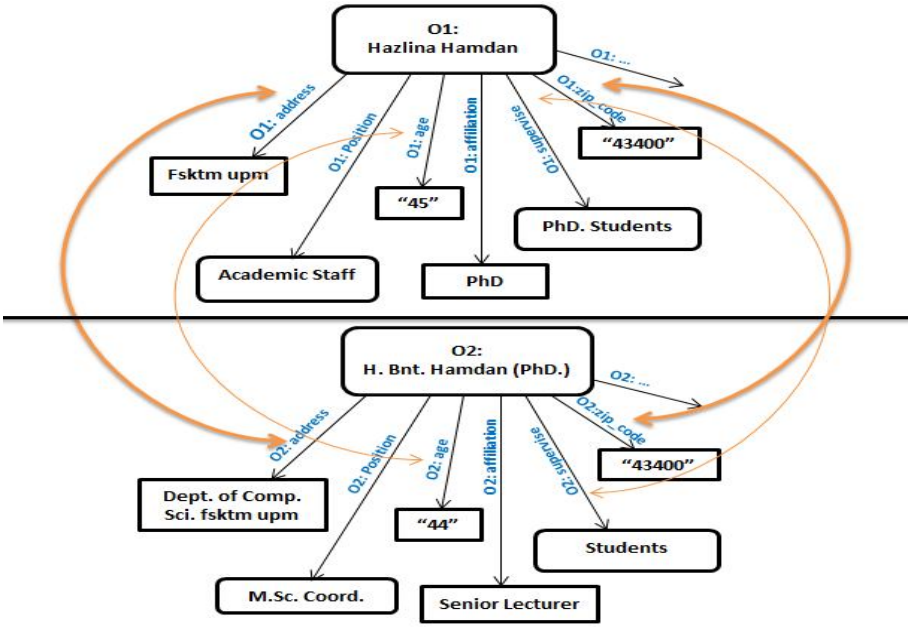

Figure 2: Intuition behind Property Alignment

3.1 Intuition of Property Alignment

In linked data sharing, property alignment substitutes the schema alignment convention. This is due to the fact that schema alignment is applied for data serialization purpose. RDF schema is formally defined by vocabularies like RDFS and OWL [16]. Property alignment also depends on the existing semantic relation found within the given data to match. Therefore, property alignment is important enough to be data-driven rather always on intuitions or assumptions. Property alignment between two given property schemas is described in Figure 2 to illustrate the intuition behind the property alignment. The property alignment is shown by the curve arrows between the vertices labels of the graph. The alignment between these properties indicates that the alignment cannot rely heavily on the strings but also through the triples properties.

A very good property alignment algorithm can be integrated with training set generator component or be an independent component in an instance based matching pipeline. To perform this alignment, Machine Learning (ML) based function for similarity will transform the entity pairs into numeric feature vectors. The PA proposed in this paper is shown in Figure 3.

The algorithm describes the property alignment step-by-step. Stage 1 of the algorithm map the prefix (URI) of columns in the property table and applies the exact match indexing method on remaining URIs in order to determine 1:1 mapping. The possible issue here is that $\acute{E}$ is poised to involve subject-subject pair. The columns score over records used in the algorithm is described in the two given examples below.

Input: Sets D and N of positive and negative training samples respectively Attribute sets A1 and A2

Output: Property Alignment $\acute{E}$

	Initialize empty set $\acute{E}$
	Initialize numeric variable avg: = 0
Stage 1	Initialize empty $\|A1\| \times \|A2\|$dimensional matrix M
	for all attribute pairs (a_i^1, a_j^2) in $A1 \times A2$ ***do***
	if URI of a_i^1 and a_j^2 exactly match **then**
	add (a_i^1, a_j^2) to $\acute{E}$
	end if
	end for
	for all attribute pairs (a_i^1, a_j^2) in $A1 \times A2$ ***do***
Stage 2	$M[i,j] :=$ $ColumnSim(D, a_i^1, a_j^2) - ColumnSim(N, a_i^1, a_j^2)$
	end for
	for all pairs $(a_i^1, a_j^2) \in \acute{E}$ **do**
Stage 3	$Avg+ = \frac{ColumnSim(D, a_i^1, a_j^2) - ColumnSim(N, a_i^1, a_j^2)}{\|\acute{E}\|}$
	end for
	for all inputs in M **do**
	If inputs $M[i,j] \geq Avg$ **then**
Stage 4	add (a_i^1, a_j^2) to $\acute{E}$
	end if
	end for
Stage 5	Output $\acute{E}$

Figure 3: Property Alignment Algorithm

Ex1: let n = 3 and the sorted list in D' in the algorithm is given as $Q' = [(i1, j3), (i2, j5), (i1, j7), (i6, j1)]$, where i and j represents the records *P1 and P2* in the property tables respectively. The matrix of the above list can be represented as:

	i_1	i_2	i_3	i_4	i_5	i_6	i_7
j_1						$i6, j1$	
j_2							
j_3	$i1, j3$						
j_4							
j_5		$i2, j5$					
j_6							
j_7	$i1, j7$						

With $n = 3$, the positive training set will be $D = [(i1, j3), (i2, j5), (i6, j1)]$, since the record i1 appeared most in the scoring pairs.

Ex2: proceeding from ex1, the generated set $D = [(i1, j3), (i2, j5), (i6, j1)]$, and the possible non-match set *G* by permuting *D* may be derived from the table below:

	i_1	i_2	i_3	i_4	i_5	i_6	i_7
j_1	$i1, j1$	$i2, j1$					
j_2							
j_3		$i2, j3$					
j_4							
j_5	$i1, j5$						
j_6					$i6, j5$		
j_7							

The possible set $D = [(i1, j1), (i2, j3), (i6, j5)]$, in practice these permutation yields a near perfect result in terms of accuracy on the generated sets.

Stage 2 of the algorithm was introduced to protect the happening of unintended matches. In this stage, *M* matrix is equipped with the matrix cell *i* and *j* which contains the value obtained by substituting the score of *columns* of corresponding elements a_1 and a_j on *D* and *N*. Applying the initial alignment *É* obtained via the URI stem matching, it calculates the average score of corresponding attributes in the matrix cells to the attributes in *É*, (stage 3) of the algorithm. It is also used as a threshold to select property alignment, (stage 4) of the algorithm. Finally, the algorithm will output the resulting alignment *É* of the properties in the final stage of the algorithm (stage 5). This property alignment algorithm runs in less than a minute even with a high number of properties. The parameter-free behavior of this algorithm gave it merits that it can run in similar to a readily available item in stock as it does not include parameter search. Similar to this approach was applied to populate linked data [17] and to the best of our knowledge, there is no existing hybrid approach which is parameter-free, specific to instance matching in the available literature.

4. ALIGNMENT GENERATION BASED ON ALIGNED PROPERTIES

This component covers the final stage of the matching as presented in our framework. Every candidate set generated in the previous stages is transformed into a feature vector. These feature vectors will then be classified by the machine learning model trained earlier. Since machine learning dominated the similarity paradigm within AI research community, this similarity stage is considered classification step in this work. The candidate attribute pairs are converted to vectors at this step by machine learning classifier described earlier in this thesis. Machine learning method of classification dominated the paradigm of similarity generation among many research communities [18], [19]. Due to this reason, the similarity generating algorithm proposed in this thesis takes as input the auto-generated training samples produced by our training set generation algorithm and the property alignment generated by our PA algorithm to generate the final similarity between the ontologies thereby complete the process shown in Figure 4. The training set generator in conjunction of property alignment model enable the automatic creation of *semantic mapping* which is the core idea behind any ontology matching. With this approach, similarity can be determined with high degree of accuracy (high precision and high recall) as all concepts and instances of the candidate ontologies are fully considered for the matching. Generating mappings semantically can be greatly subjective and be contingent to target application or choice of the user.

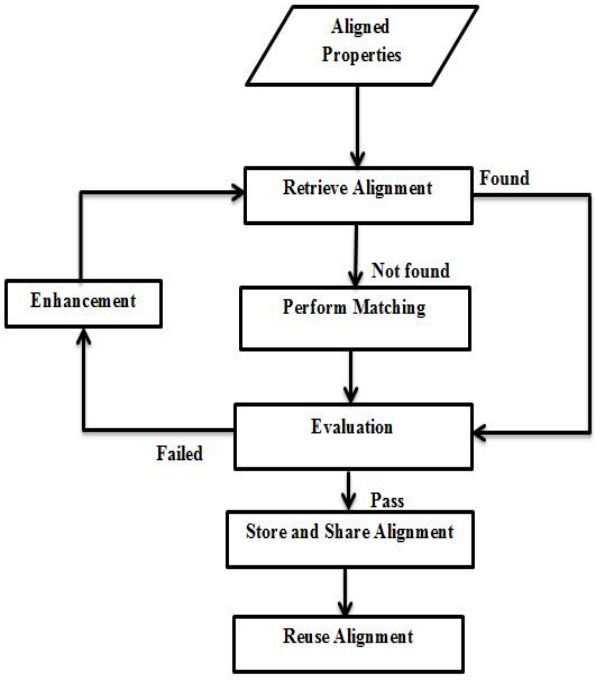

Figure 4: Similarity Generation Process Flow

4.1 Experimental Setup

The input data sets in our method are ontologies represented in *Resource Description Framework (RDF)*[1] data format. These ontologies are categorized as source and target ontologies. The source ontology is ontology of a given domain that satisfied to be mapped into the different ontology of the same domain. The later is the one described as the target ontology. We pre-processed our data by applying basic text filters: parenthesis and punctuations were eliminated; all words are converted to lower case. Sentences and phrases were tokenized by spaces so as the generated similarity between sets of words. Therefore, sentence or phrase *p* of *m* words defined as:

$$w = [wi] \quad (2)$$

$$where, w = a\,word\ and\ i = 1, 2, 3 \ldots, m$$

The training sets generated by the UTSG are the input to the property alignment component. We evaluate our proposed property alignment algorithm against two baseline property aligners to explore the potential of the algorithm. The first one is the column matcher proposed by [20]. This algorithm can classify and identify corresponding instance attributes by comparing feature values within the generated duplicate values. This matcher does not apply the training set; all values above threshold are outputted as the match provided the similarity matrix is generated. The second approach is the algorithm proposed by [17]. To the best of our knowledge, these are the only technique that presented an instance matcher which is able to aligned properties of graph data with generated training samples, even though it demonstrated poor performance on a large-scale data and requires little user participation.

The measurement metrics for this experiment is the precision, the recall and the F-Measure of the generated alignment. The precision measures the effectiveness of the algorithm in matching the candidate training sets and the recall measures the efficiency of the algorithm in generating the true alignments while the F-measure determines the trade-off between the precision and recall by calculating their harmonic mean.

5. RESULTS AND DISCUSSION

Most of the existing property aligners works outside the context of instance matching rather applied similar approach based on the extensional RDF property, [1], [21], [22]. In order to address this issue, we propose a parameter-free property alignment algorithm which is hybrid in nature. This algorithm will consider both the property name and column aggregation of the training sets generated by UTSG. Applying this combination, the system can achieve better performance with regard to both recall and precision measures without the loss of any piece of useful information.

Learning Effect: The goal of this experiment is to evaluate the performance of our unsupervised approach algorithm against the supervised SVM classifiers. In this experiment, two SVM (20%, 40%) classifiers are evaluated against our unsupervised classifier. The goal is to determine the level of supervision and noise effect on the general classification performance. The evaluation is carried out against two supervised configurations, 20% and 40% respectively. These approach have recently been successfully utilized in the existing literature for both schema and instance matching [26],[27] and [25].

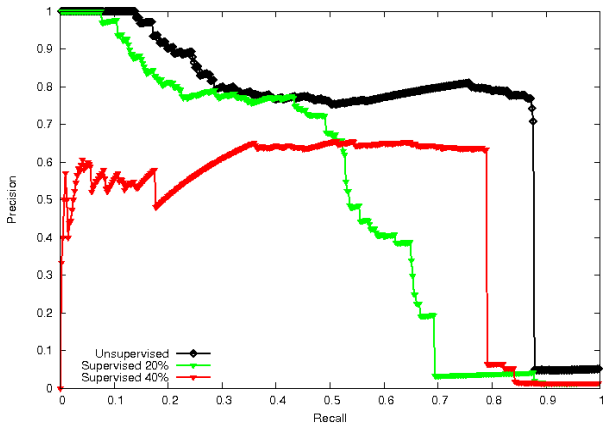

Figure 5: Unsupervised against SVM Supervised Configurations on PR Data

The precision-recall curve (Figure 5) shows that our unsupervised method outperforms the supervised ones. The increase in the level of supervision indicates the decrease in the performance of the classifier, this can be clearly sported from the figure where 40% supervision led to the decrease in the performance of the classifier compared to 20% supervision and unsupervised supervision as well. Thus, applying the unsupervised classifier yields a better result in generating a complete alignment. The usual metrics of precision and recall are also used to evaluate the supervision effect in this similarity generation process.

6. CONCLUSIONS AND FUTURE WORK

The volume and variety of data in the present day web give organizations opportunities as well as the challenges in their organization's management. Data integration and unification methods provides promising approach to addressing heterogeneity issues and improve the ability of data interoperability and sharing across organizations. In this

[1] RDF is "Standard model for data interchange on the web": (https://www.w3.org/RDF/)

paper, a formal data-driven technique of addressing heterogeneity through data properties alignment is proposed. The method utilizes the generated training samples in performing the alignment. Previous works assumes property alignments by intuition in matching data instances, especially with regard to their underlying ontological presentation. We evaluated our method on standard data set synthesized from benchmark ontologies from semantic web ontology matching web sites. The experimental results describes that our method can significantly improve and bootstrap the matching performance by considering all ontology instances to be a matching candidate. We also designed a similarity generation method that can effectively generate complete alignment using both generated training samples and aligned ontologies. There are many direction of work intended to undertake in a near future. This includes MapReduce implementation of this method to ensure real time scalability evaluation.

ACKNOWLEDGEMENT

This work is supported by University Putra Malaysia grant (Putra Grant No: 9569200) and Al-Qalam University Katsina (AUK), Katsina State Nigeria under the University's Staff Development Unit.

REFERENCES

[1] N. DuyHoa, Z. Bellahsene, and R. Coletta, **A flexible system for ontology matching**, CEUR Workshop Proc., vol. 734, pp. 73–80, 2011.

[2] A. H. Nejhadi, B. Shadgar, and A. Osareh, **Ontology Alignment Using Machine Learning Techniques**, Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 2, pp. 139–150, 2011. https://doi.org/10.5121/ijcsit.2011.3210

[3] S. Link, D. Nikovski, A. Esenther, and X. Ye, **Matcher Composition Methods for Automatic Schema Matching**, Enterp. Inf. Syst., vol. 141, pp. 108–123, 2013. https://doi.org/10.1007/978-3-642-40654-6_7

[4] F. Duchateau and Z. Bellahsene, **Designing a Benchmark for the Assessment of Schema Matching Tools**, Open J. Databases, vol. 1, no. 1, pp. 3–25, 2014.

[5] S. Anam, Y. S. Kim, B. H. Kang, and Q. Liu, **Review of Ontology Matching Approaches and Challenges**, Int. J. Comput. Sci. Netw. Solut., vol. 3, no. 3, 2015.

[6] E. Daskalaki, G. Flouris, I. Fundulaki, and T. Saveta, **Instance matching benchmarks in the era of Linked Data**, Web Semant. Sci. Serv. Agents World Wide Web, vol. 39, pp. 1–14, 2016. https://doi.org/10.1016/j.websem.2016.06.002

[7] B. Araujo and L. Zhao, **Data heterogeneity consideration in semi-supervised learning**, Expert Syst. Appl., vol. 45, pp. 234–247, 2016. https://doi.org/10.1016/j.eswa.2015.09.026

[8] S. Cerón-Figueroa et al., **Instance-based ontology matching for e-learning material using an associative pattern classifier,** Comput. Human Behav., vol. 69, pp. 218–225, 2017. https://doi.org/10.1016/j.chb.2016.12.039

[9] M. Gawich, A. Badr, H. Ismael, and A. Hegazy, **Alternative Approaches for Ontology Matching**, Int. J. Comput. Appl., vol. 49, no. 18, pp. 29–37, 2012.

[10] Zhu and Junwu, **Survey on Ontology Mapping**, Phys. Procedia, vol. 24, pp. 1857–1862, 2012.

[11] M. E. Cotterell and T. Medina, **A Markov Model for Ontology Alignment**, arXiv Prepr. arXiv1304.5566, 2013.

[12] V. Ebrahimipour and S. Yacout, **Ontology-based schema to support maintenance knowledge representation with a case study of a pneumatic valve**, IEEE Trans. Syst. Man, Cybern. Syst., vol. 45, no. 4, pp. 702–712, 2015. https://doi.org/10.1109/TSMC.2014.2383361

[13] C. Shao, L. M. Hu, J. Z. Li, Z. C. Wang, T. Chung, and J. B. Xia, RiMOM-IM: **A Novel Iterative Framework for Instance Matching**, J. Comput. Sci. Technol., vol. 31, no. 1, pp. 185–197, 2016.

[14] S. Cerón-Figueroa et al., **Instance-based ontology matching for e-learning material using an associative pattern classifier,** Comput. Human Behav., vol. 69, p. 53, 2017. https://doi.org/10.1016/j.chb.2016.12.039

[15] P. F. Patel-Schneider, **Using Description Logics for RDF Constraint Checking and Closed-World Recognition**, Proc. Twenty-Ninth AAAI Conf. Artif. Intell., pp. 247–253, 2014.

[16] M. Kejriwal and D. P. Miranker, **An unsupervised instance matcher for schema-free RDF data**, J. Web Semant., vol. 35, pp. 102–123, 2015.

[17] M. del C. Legaz-García, M. Menárguez-Tortosa, J. T. omás Fernández-Breis, C. G. Chute, and C. Tao, **Transformation of standardized clinical models based on OWL technologies: from CEM to OpenEHR archetypes**, J. Am. Med. Inform. Assoc., vol. 22, no. 3, pp. 536–544, 2015.

[18] D. Vatsalan, P. Christen, and V. S. Verykios, **A taxonomy of privacy-preserving record linkage techniques**, Inf. Syst., vol. 38, no. 6, pp. 946–969, 2013.

[19] A. Bilke and F. Naumann, **Schema Matching Using Duplicates**, 21st Int. Conf. Data Eng., pp. 69–80, 2005.

[20] K. Todorov, P. Geibel, and K.-U. Kuehnberger, **Extensional Ontology Matching with Variable Selection for Support Vector Machines**, Complex, Intell. Softw. Intensive Syst. (CISIS), 2010 Int. Conf., 2010. https://doi.org/10.1109/CISIS.2010.59

[21] S. Gherbi and M. T. Khadir, **Inferred Ontology Concepts Alignment Using Instances and an External Dictionary**, Procedia Comput. Sci., vol. 83, no. Ant, pp. 648–652, 2016.

[22] W. E. Winkler, **Improving EM Algorithm Estimates for Record Linkage Parameters**, Research report series, Stat. #2002-05, pp. 1–29, 2003.

[23] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, **Locality-sensitive hashing scheme based on p-stable distributions**, Proc. Twent. Annu. Symp. Comput. Geom. - SCG '04, p. 253, 2004.

[24] S. Duan, A. Fokoue, O. Hassanzadeh, A. Kementsietsidis, K. Srinivas, and M. J. Ward, **Instance-based matching of large ontologies using locality-sensitive hashing**, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7649 LNCS, no. PART 1, pp. 49–64, 2012. https://doi.org/10.1007/978-3-642-35176-1_4

[25] A. L. Dallora, S. Eivazzadeh, E. Mendes, J. Berglund, and P. Anderberg, **Prognosis of Dementia Employing Machine Learning and Microsimulation Techniques: A Systematic Literature Review**, Procedia Comput. Sci., vol. 100, pp. 480–488, 2016. https://doi.org/10.1016/j.procs.2016.09.185

[26] M. U. Devi and G. M. Gandhi, **An Enhanced Fuzzy Clustering and Expectation Maximization Framework based Matching Semantically Similar Sentences**, in Procedia Computer Science, 2015, vol. 57. https://doi.org/10.1016/j.procs.2015.07.406