UNIVERSITI PUTRA MALAYSIA

INSTANCE MATCHING FRAMEWORK FOR HETEROGENEOUS SEMANTIC WEB CONTENT OVER LINKED DATA ENVIRONMENT

ABUBAKAR MANSIR

FSKTM 2022 9

INSTANCE MATCHING FRAMEWORK FOR HETEROGENEOUS SEMANTIC WEB CONTENT OVER LINKED DATA ENVIRONMENT

By

ABUBAKAR MANSIR

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

June 2021

COPYRIGHT

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

INSTANCE MATCHING FRAMEWORK FOR HETEROGENEOUS SEMANTIC WEB CONTENT OVER LINKED DATA ENVIRONMENT

By

ABUBAKAR MANSIR

June 2021

Chair :Hazlina Binti Hamdan, PhD
Faculty :Computer Science and Information Technology

Over the past decade, instance matching has been the possible method of discovering relationships within heterogeneous Resource Description framework (RDF) based data that can represent the same real-word entity over Linked Data environment. The exponential growth of data being experienced in the recent times in terms of volume, variety and velocity makes existing instance matching frameworks difficult to effectively discover relationships and generate a matching output. These frameworks suffer a high amount of comparisons in discovering matching attributes at initial stage which leads to missing attributes in generating training samples, thus results to incomplete alignment generation as matching output. Manual parameter configuration is another problem associated to existing matching frameworks, which make them weak in handling data with high level of heterogeneity. Another issue caused by these problems is the time taken to generate alignment as well as maximum memory space utilization during the process.

Effective and scalable instance matching framework is needed to improve the matching performance. In this study, an instance matching framework is proposed to address the identified problems to improve the ability of generating better and accurate matching output (alignment) in a minimum running time. This framework adapted the methods used in the benchmark studies with additional components and modifications in some existing components to boost the matching performance. A proposed framework works interactively with the following components: Serialisation and pre-processing, unsupervised training set generation, property alignment and two-fold similarity generation components.

Serialisation involves translating RDF data from of N-Triples file to Comma Separated Value (CSV) file format while pre-processing performs basic text filter. In attribute discovery component, potential matching attributes are discovered by clustering attributes of matching instances into similar and non-similar clusters in order to

discover potential attribute pairs for the matching. These discovered attributes serve as input to a modified training set generation component, where training sets are generated based on the potential attributes' clusters. Property alignment check the irregular data associated to the generated sets to optimise the matching performance. The last component generates similarity with self-configuration behavior.

Experiments have been conducted to evaluate the performance of individual components and the output of the framework as whole. The evaluation is performed on real-world datasets provided in different Ontology Alignment Evaluation Initiative (OAEI) campaign as benchmark data for instance matching track evaluation. The output of each algorithm is evaluated, the results have shown that each algorithm performs well and outperforms the existing algorithms on all test cases in terms better output generation and effective handling of heterogeneity from different domains, which is a necessary concern in all data-intensive problems.

A proposed framework demonstrated a significant improvement compared to the benchmark frameworks: Agreement Maker Light (AML), RiMOM-Instance Matching (RiMOM-IM) and Unsupervised Instance Matcher in terms of accuracy of alignment generation in a minimum time frame with ability to accommodate increase in the size of Linked Data (LD) in today's web content.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

KERANGKA KERJA PEMADANAN TIKA (*INSTANCE*) UNTUK KANDUNGAN WEB SEMANTIK HETEROGEN DALAM PERSEKITARAN DATA TERPAUT

Oleh

ABUBAKAR MANSIR

Jun 2021

Pengerusi :Hazlina Binti Hamdan, PhD
Fakulti :Sains Komputer dan Teknologi Maklumat

Sepanjang dekad yang lalu, pemadanan tika telah menjadi kaedah yang mungkin untuk menemui hubungan dalam Kerangka Kerja Deskripsi Sumber (RDF) heterogen berdasarkan data yang boleh mewakili entiti perkataan sebenar yang sama dalam persekitaran Data Terpaut. Pertumbuhan data yang eksponen sejak kebelakangan ini dari segi jumlah, kepelbagaian dan halaju menjadikan kerangka kerja pemadanan tika sedia ada sukar untuk menemui hubungan dengan berkesan dan menjana output yang sepadan. Kerangka kerja ini mengalami jumlah perbandingan yang tinggi dalam menemui atribut padanan pada peringkat awal yang menyebabkan atribut hilang dalam menjana sampel latihan, sekali gus mengakibatkan penjanaan penjajaran tidak lengkap sebagai output padanan. Konfigurasi parameter manual ialah satu lagi masalah yang berkaitan dengan kerangka kerja pemadanan sedia ada, menjadikannya lemah dalam mengendalikan data dengan tahap kepelbagaian yang tinggi. Isu lain disebabkan oleh masalah ini ialah masa yang diambil untuk menjana penjajaran serta penggunaan ruang memori yang maksimum semasa proses.

Kerangka kerja pemadanan tika yang efektif dan boleh skala diperlukan untuk meningkatkan prestasi pemadanan. Dalam kajian ini, kerangka kerja padanan tika dicadangkan untuk menangani masalah yang dikenal pasti bagi meningkatkan keupayaan menjana output padanan (penjajaran) yang lebih baik dan tepat dalam masa jalan yang minimum. Kerangka kerja ini disesuaikan dengan kaedah yang digunakan dalam kajian penanda aras dengan komponen tambahan dan pengubahsuaian dalam beberapa komponen sedia ada untuk meningkatkan prestasi padanan. Kerangka kerja yang dicadangkan berfungsi secara interaktif dengan komponen berikut: *Serialisation* dan pra-pemprosesan, penjanaan set latihan tanpa penyelia, penjajaran sifat dan komponen penjanaan keserupaan *two-fold*.

Serialisation melibatkan terjemahan data RDF dari fail N-Triples ke format fail CSV manakala pra-pemprosesan melaksanakan penapisan teks asas. Di dalam komponen penemuan atribut, atribut pemadanan yang berpotensi ditemukan oleh atribut penggugusan pemadanan *instances* kepada kelompok serupa dan tidak serupa untuk menemui pasangan atribut yang berpotensi untuk pemadanan. Atribut yang ditemukan ini berfungsi sebagai input kepada komponen penjanaan set latihan yang diubah suai, di mana set latihan dijana berdasarkan kelompok atribut yang berpotensi. Penjajaran sifat menyemak data tidak teratur dikaitkan dengan set yang dijana untuk mengoptimumkan prestasi pemadanan. Komponen terakhir menjana keserupaan dengan perilaku konfigurasi-sendiri.

Eksperimen telah dijalankan untuk menilai prestasi komponen individu dan output kerangka kerja secara keseluruhan. Penilaian dilakukan terhadap dataset sebenar yang terdapat dalam kempen OAEI yang berbeza sebagai data penanda aras untuk penilaian trek pemadanan *instance*. Output setiap algoritma dinilai, keputusan telah menunjukkan bahawa setiap algoritma berfungsi dengan baik dan mengatasi algoritma sedia ada pada semua kes ujian dari segi penjanaan output yang lebih baik dan kepelbagaian pengendalian yang efektif dari domain yang berbeza, dimana ini perlu diambil perhatian dalam semua masalah data intensif.

Kerangka kerja yang dicadangkan menunjukkan peningkatan yang ketara berbanding dengan kerangka kerja penanda aras (AML, RiMOM-IM, dan Unsupervised Instance Matching) dari segi ketepatan penjanaan penjajaran dalam jangka masa minimum dengan keupayaan untuk menampung peningkatan saiz Data Terpaut (LD) dalam kandungan web hari ini.

ACKNOWLEDGEMENTS

All praises to the Almighty Allah (SWT), the Most Gracious and Merciful, for giving me the life, strength, and determination to complete this study. I sincerely wish to express my deepest gratitude to the entire members of my thesis supervisory committee who made this journey possible. Specifically, Dr. Hazlina Binti Hamdan (Committee Chair), Prof. Madya Dr. Norwati Mustapha (member), Prof. Madya Dr. Teh Noranis Moh'd Aris (member) who provided their guidance, support, and understanding during my candidature. Your scholarly guidance, good support, commitment, excellent supervision, and great encouragement are highly appreciated.

My sincere gratitude and appreciation go to the management of Al-Qalam University Katsina (AUK) under the leadership of the vibrant Vice-chancellor, Professor Shehu Garki Ado for sponsoring my study. For the entire AUK community, I say thank you for the prayers and profound good wishes.

My profound gratitude also goes Mother, Hajiya Zainab Abubakar and to my entire brothers and sisters in Gantarbi family for their prayers, encouragement as well as moral and financial support. A very big accolade to my exquisite wife Bilkisu Ibrahim and my children Maryam, Muhammad, A'isha, Abubakar (Sayyid), Naja'atu, Ibrahim (Khalil) and Zainab (Suhailat) for their overwhelming sacrifices and patience throughout my study period. May Allah bless their lives to benefit humanity, Ameen.

Immense appreciation to all my friends particularly Dr. Maharazu Mamman Danmusa and Dr. Aminu Musa whose efforts cannot be overlooking right from my admission process, upon my arrival and stay in Malaysia. Finally, I will like to acknowledge the support of colleagues in the Department of Mathematical Sciences, Al-Qalam University Katsina as well as the colleagues at FSKTM, especially Dr. Hazrina Binti Sofian (Malaysian), Dr. Abdulkareem Bello, Abdulmajeed Babangida Umar, Suleiman Sa'ad, Ma'aruf Muhammed Lawal, Babangida Lawal, Tahiru Shitu, Muath Ali (Jordanian), Ismail Mohammed (Somalian) and Muhammed Suraj (Ghanian). May you be rewarded abundantly for sharing productive ideas.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Hazlina binti Hamdan, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairperson)

Norwati binti Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Teh Noranis binti Mohd Aris, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 10 February 2022

Declaration by Graduate Student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection check.

Signature: _____ Date: _____

Name and Matric No.: Abubakar Mansir, GS47201

Declaration by Members of the Supervisory Committee

This is to confirm that:

- the research and the writing of this thesis were done under our supervision;
- supervisory responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2015-2016) are adhered to.

Signature: _____
Name of Chairman of Supervisory Committee: _____

Signature: _____
Name of Member of Supervisory Committee: _____

Signature: _____
Name of Member of Supervisory Committee: _____

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AML	*AgreementMakerLight*
BoW	Bag of Words
CSV	Comma Separated Value
DL	Description Language
GIFs	General Indexing Function
IIMB	Islab Instance Matching Benchmark
JVM	Java Virtual Machine
LD	Linked data
LibSVM	Library SVM
LOD	Linked Open Data
LSF	Link Specification Function
NLP	Natural Language Processing
OAEI	Ontology Alignment Evaluation Initiative
OWL	Web Ontology Language
OWL-API	Web Ontology Language – Application Programming Interface
RDF	Resource Description Framework
RDF-S	RDF - Schema
RIF	Rule Interchange Format
RiMOM-IM	RiMOM Instance Matching
SVM	Support Vector Machine
SW	Semantic Web
TF-IDF	Text Frequency-Inverse Document Frequency
TSG	Training Set Generator

URIs:	Resource Identifiers
UTSG	Unsupervised Training Set Generator
W3C	World Wide Web Consortium
WORA	Write one, Run Anywhere
XML	eXtensible Mark-up Language

CHAPTER 1

INTRODUCTION

1.1. Background

The exponential growth of web data in terms of its volume, variety and velocity across organizations necessitated the concern over issues of data integration in terms of linked data. Existing studies have reached a certain point to find solutions to the heterogeneity issues associated with data integration with regard to its semantics, popularly known as semantic heterogeneities (Gracia and Mena, 2012). These solutions addresses quite number of different problems, such as the application of different specification languages, details in describing the domain of interest, high level of semi-automation as well as possible matching solutions in both schema and instance level. Today, the high expressiveness of semantic information enables the automatic or semi-automatic processing of Web resources.

Industries and academia have discovered that the Semantic Web can ease the interoperability and integration of both Intra and Inter-business processes. The power of the Semantic Web lies in systems interoperability, compatibility between diverse formats, and the discovery of new relationships between different resources (Saif, 2016). Linked Data (LD) is the remarkable effort made to allow people connect and share independently generated data in the Semantic Web (SW) (Song, Luo and Heflin, 2016). Connecting these data over the LD is made by integrating all related data and information by aligning their abstractions.

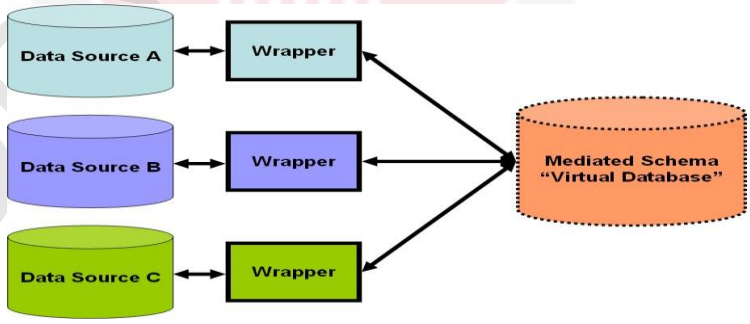

Figure 1.1: Abstract View of Data Integration Problem
(Zhao & Ichise, 2014)

Data integration is the process of providing to a user or an application, a uniform means of combining multiple data sources (Zhao and Ichise, 2014). These data unification can be accomplished through wrappers or virtual interface (Figure 1.1). For data integration to be conducted in an efficient manner it is necessary and equally important to resolve inconsistencies or conflict that exists in the heterogeneous sources. To

achieve the effectiveness of information systems, some systems have to undergo reuse via integration performed if not all but in some certain features of the data sources.

Ontology integration is the general term used to describe different operations being conducted on the ontologies, such as features sharing, merging, unifying, mapping, aligning and matching between different ontologies belonging to either same or different domain of the ontologies. Ontology integration is the process which may be done in three levels (Pinto and Martins, 2001):

1. Building a new ontology by reusing other available ontologies: this is the simplest case of ontology integration in which new ontology is built by adopting the existing ontologies.

2. Merging different ontologies about the same domain into a single one that unifies all ontologies: In this case, the ontology should be built by using knowledge from exactly the same domain of existing ontologies.

3. Introducing ontologies into the application: Here, several ontologies are introduced into an application, they are shared among different software applications which make it possible to use several ontologies to implement or identify knowledge-based applications based on distributed resources.

On the other hand, (Noy and McGuinness, 2001) identified two general approaches for ontology integration process: (1) Merging several ontologies for developing one consistent ontology and (2) Alignment of several ontologies to be identifying their references to determine the possibility of employing all the ontologies. Therefore, to determine the inconsistency or conflict among ontologies one must define and analyze several ontologies as suggested in case one above. In the second case, in every two candidate's ontology, it is required to find a mechanism which points out the relationship between the attributes of both ontologies. In this case, it is possible to apply both ontologies for the set goal without unifying them to single ontology.

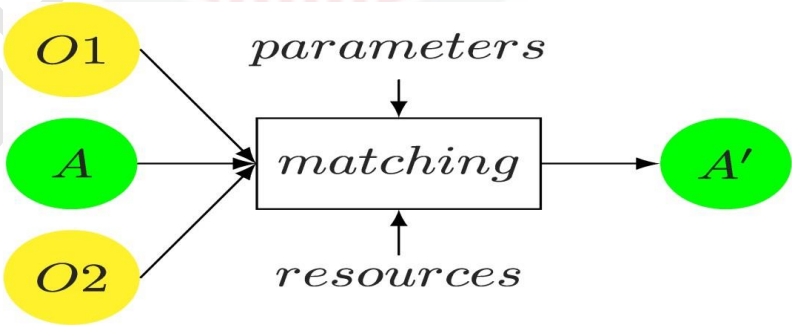

Figure 1.2: Abstract view of Ontology Matching Problem
(Source: Noy and McGuinness, 2001)

To achieve the desired objective of ontology integration, it is obvious that the process involves conflict resolution popularly known as ontology matching.

Ontology matching is the process of finding relationships in different ontologies of the same domain to represent a single real-world entity, (Figure 1.2). Although in some situations conflict may occur only when the ontologies represent the same real word. Ontology matching process be 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 process can be represented in a formula as:

$$\boldsymbol{A'} = (\boldsymbol{O1}, \boldsymbol{O2}, \boldsymbol{A}, \boldsymbol{P}, \boldsymbol{R}) \qquad (1.1)$$

In ontology matching research, ontologies conflicts in both same and different domain will be considered, different level of conflicts (*instance level*, *concept level*, and *relationlevel)*, as well as different approaches toward the resolution of the conflict at each of these levels. But the proposed idea will be limited to addressing the matching problem at instance-level. Specifically, heterogeneity and scalability issue as well as the possible solution to the general weakness of the existing matching techniques, the trade-off between effectiveness and efficiency in matching. Table 1.1 illustrates the goaldefinition of ontology matching.

Table 1.1: Ontology Matching Goal Definition

<table>
<tr><td colspan="2">DEFINITION

Ontology matching is the activity that finds relationships between two or more ontologies.</td></tr>
<tr><td colspan="2">GOAL

Designing a process for matching the ontologies using their URIs definition.</td></tr>
<tr><td>INPUT

The characteristics of the ontologies to match and context in which matching will occur.</td><td>OUTPUT

The specification of a matching process of candidate ontologies.</td></tr>
<tr><td colspan="2">END USERS

Semantic Web Application Designers (SWAD)</td></tr>
<tr><td colspan="2">PURPOSE

Use when developing intelligent applications that requires run-time matching.</td></tr>
</table>

In the recent days of semantic web technology, there is a wide interest in ontologies and the categories of issues associated with semantic heterogeneity, like ontology conflict, ontology integration, and ontology sharing (Euzenat & Shvaiko, 2007). Some notable requirements have been deliberated individually and collectively by different scholars which resulted to the expert's agreements on basic requirements that need to be satisfied any instance matching system or framework (Anam, Kim, Kang, & Liu, 2015). Despite the heterogeneity requirement, three additional requirements are identified to be necessary for instance matching to realize its full potentials of being Artificial Intelligent problem. Therefore, alignment generation will only be complete if the framework exhibits Adaptability, Scalability and domain-Independence as basic requirements. In a nutshell, ontology alignment is the output of ontology matching.

Ontology alignment being the result of the matching involves range of techniques to generate complete alignment between given ontologies. Information origin can be considered as a basis in which alignment can be generated. These information may come directly from the concept of the ontology or its properties (Euzenat, Jérôme and Shvaiko, 2012). These concepts and properties that made up ontology are referred as knowledge piece (Pinto & Martins, 2001).

Major issues associated to alignment generation are mainly as a result of inability of most existing matching systems to match ontology properties which is widely known as instance-based matching (Maree & Belkhatir, 2015), (Altnel & Ganiz, 2016), (Kejriwal & Miranker, 2014). One big issue with ontology alignment is that it is difficult task to understand semantic relations between properties of different ontologies. In order to perform the alignment, a developer is always require to make mapping definitions using semi-automated tool or manually. Another issue associated with alignment is lack of empirical validation of alignment that involves rea-world ontologies (Hu et al., 2017). Existing ontology matching systems focused on lightweight concepts of the ontologies but not its properties or instances during matching (Li, Wang, Zhang, & Tang, 2013) and (Shao et al., 2016). Lastly, lack of gold standard to serve as reference ontologies for evaluation is another drawback of existing matching frameworks (Shao et al., 2016). Therefore, property alignment is required to substitute the conventional schema or conceptual matching in order to address the challenges of ontology matching highlighted in this study.

Property Alignment. RDF schema is formally defined by vocabularies like RDFS and OWL (Patel- Schneider, 2014). Property alignment also depends on the existing semantic relation found within the given vocabularies of the data to match. Property alignment component is aimed to bootstrap the general matching process (Kejriwal and Miranker, 2015). Property alignment component controls irregular data found in an ontology. These irregular data are similar to noise in a conventional database. One objective of this study enables enhancement of property alignment component proposed in the work of (Kejriwal and Miranker, 2015) to accept input from attribute discovery component proposed in this study that provided clustering capability at the initial stage of the matching process, thereby minimizing or eliminating all irregular data as stated earlier.

1.2 Research Motivation

The basic idea of instance-based matching is that the more significant the overlaps of common instances of two concepts, the more related these concepts are. The difficult question is how to define the notion of significance for such extension overlap (Isaac *et al.*, 2008). This tricky question that motivates the research on instance-based matching. Even though it has been an artificial intelligence research area for almost five decades but still requires research attention with regards to artificial intelligence, knowledge discovery as well as intelligent information retrieval.

Furthermore, the overlap can be identified if the fourth principle of linked data as highlighted by Tim Berners Lee at W3C which says, *"Data should not continue to exist in warehouses but be connected to existing related data*" can be realized. With due consideration to the Linked Open Data (LOD), this principle has a remarkable effect. LOD is the collection of RDF graph data being published under an open license (Umbrich *et al.*, 2010). LOD has exponential nature of growth in both volume and variety. It is based on the metadata collected by different contributors accessible via their URIs (http://lod- cloud.net/).

A recent study shows that the size of LOD has risen to contain billions of triples from 1,184 datasets connected by 15,993 links as of August 2017 (Figure 1.3). These data sets are sub-clouds by domains, Cross-Domains, Geography, Government, Life Sciences, Linguistics, Media, Social Networks and Publications. LOD continue to grow in an unstructured form in both volume and variety and have attracted a significant research interest across research communities.

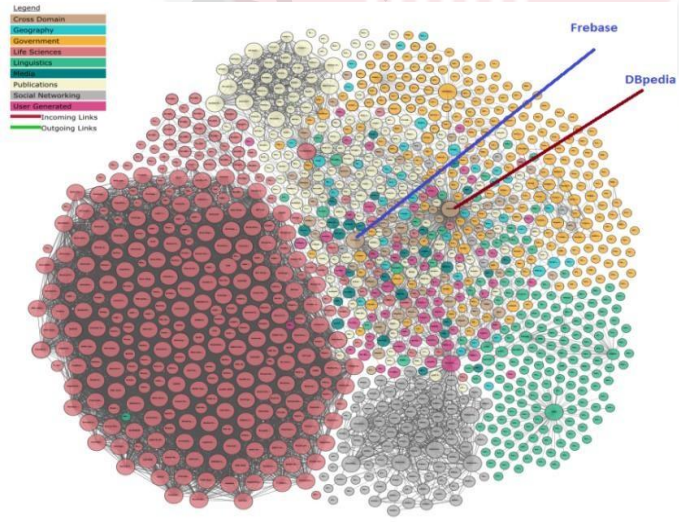

Figure 1.3: Linked Open Data: Current Update 22/08/2017
(Source: http://lod- cloud.net/)

Instance-based matching is a matching that compares sets of individuals of classes in order to decide whether they represent the same real-world object. They help in grouping together items or computing distances between the items. Instance matching plays a crucial role in semantic data integration as it interconnects all the key points of instances of the semantic world to achieve the interoperability and information integration issues. However, to cope with the demand for enabling Semantic Web technology practically, ontology instance matching is more important than schema matching (Deb Nath, Seddiqui and Aono, 2012). Another issue raised in Instance Matching (IM) is to accurately choose the subset of instances that are more likely to be like the input instance, avoiding the comparisons with impertinent instances. As the detection of mappings on schema level directly affects instance level matching, in this research, ontology schema matching and instance matching work together for discovering semantic mappings between possible distributed and heterogeneous semantic data (Deb Nath, Seddiqui and Aono, 2012).

There are many pieces of evidence on why the one real-world entity is described in different sources. In the case of instance mentioned above, in open and social data, anyone has ample right to published data and/or information, and simply adhere to representation and that best fits his application. Therefore, an effective and efficient instance matching framework is needed that can address the following challenges:

1. Resolve incomplete alignment generation by avoiding missing attributes during matching process.

2. Minimize the amount of comparisons when generating training sets by improving the computational performance of the training set algorithm.

3. Improve the performance of property alignment algorithms in handling irregular data using the improved generated training sets.

4. Resolves the manual configuration problem in generating the similarity within aligned properties. The unfortunate side of the existing methods is that they are unable to address or resolve these challenges and therefore the output of the matching frameworks still need additional and consistent research effort to achieve better performance considering the exponential growth of knowledge bases that requires to semantically interoperate in a shared environment.

These can be achieved through the incorporation of property alignment component into a matching pipeline as proposed in this study.

1.3 Problem Statement

Many studies reviewed and presented in chapter two shows that the prevailing approaches to alignment generation whether value-oriented or record-oriented that are mainly applied to enrich linked data environment used supervised learning method as presented in (Li *et al.*, 2009), (Jiménez-Ruiz and Cuenca Grau, 2011), (Khiat and Benaissa, 2015) and (Shao *et al.*, 2016) and few approaches using unsupervised

learning method as in the works of (Bilke, Alexander, 2015) and (Kejriwal and Miranker, 2015). The literature investigation shows that unsupervised methods demonstrated better performance in matching ontology properties. Therefore, unsupervised method is considered as one of the benchmark studies implemented in this study.

Although, unsupervised frameworks demonstrated a significant effort to alignment generation but did not address heterogeneity issue on many data instances as expected because they failed to consider all data attributes during mapping (Hu *et al.*, 2010), (Samur *et al.*, 2010), (Li *et al.*, 2013), (Kejriwal and Miranker, 2015) and (Shao *et al.*, 2016). In a successful matching system, all potential matching attributes need to be considered in order to generate better alignment (matching result), without which, the problem will affect the overall performance of the matching frameworks in generating final alignment (Lars C., Rafael Schimassek, Dominik Huser, Maximilian Peters, Christoph Kramer, 2018), (Müller *et al.*, 2019). This is a serious limitation in linked data settings considering the level of increase in the amount of data available in today's web and the unprecedented variety of data sources as well. Achieving a high performance in the matching model of this nature is technically impossible as alignment generation requires that all attributes of instances be potential matching attributes.

Most of these studies utilize traditional blocking methods and available matching information to improve the matching performance (Faria *et al.*, 2013), (Shao *et al.*, 2016). With blocking method, literal values are the only RDF triples used as an indexing keys which usually leads to errors of mismatched instances in each iteration. Traditional blocking methods can hardly eliminate mismatched instances, this is because instances in two different ontologies are usually defined by different set of RDF triples.

Very few studies focused on heuristic generation of training sets to improve matching effectiveness (Kejriwal and Miranker, 2015). Despite these studies, several challenges remain unresolved in achieving effective and efficient instance matching over Linked Data Environment. These include: Existing instance matching frameworks are ineffective in alignment generation (matching output) due to their inability to generate complete alignments as many attributes are missing during the process.

In (Kejriwal and Miranker, 2015), the system generates training set through direct access to serialized ontologies and then uses the generated sets to perform the property alignment. However, the system resulted to high amount of comparison in generating the training set, which in turn requires maximum search space and high running time to perform the matching task. Many attributes that are potential for the matching are also missingduring the process. Property alignment component is aimed to bootstrap the general matching process. The primary objective of Training Set Generator (TSG) is to provide input to the Property Aligner (PA) which is also a new component proposed by (Kejriwal and Miranker, 2015) to the traditional matching system.

The RiMOM-IM framework (Shao *et al.*, 2016) utilizes available matching instances to improve the matching performance and error control. It also generates candidate set via direct access to instances at pre-processing step and then perform the blocking. However, the system lack self-configuration behaviour which denied many attributes to be considered for a matching. The property alignment which control the irregular data found in the training sets is assumed by intuition rather than data driven, thus the framework cannot handle irregular data associated to the generated training sets. Irregular data is similar to noise found in traditional database. It is any unwanted exploited in the RDF based data that may affect the performance of the matching. Example of irregular data in heterogeneous semantic web data include variation in the ages of a particular person in different knowledge bases.

Therefore, these frameworks are impractical to be used in semantic web applications that requires a complete alignment between inputs data, self-configuration behaviour and run- time matching. Self-configuration allows automatic parameter tuning in order to maximize and optimize system performance in ensuring high level of automation in generating output (Konen *et al.*, 2011). This limitation is clearly identified in the benchmark studies as crucial to Linked Open Data in real-time, robust deployment.

1.4 Research Objectives

The goal of this study is a propose Instance Matching Framework that can generate complete alignment (matching result) as ***sameAs*** links without missing attributes during the matching process with due consideration to heterogeneity and scalability requirements. This can be achieved through the following objectives:

1. To propose an attribute discovery method based on clustering approach in order to minimize the attributes comparison when generating training set. Existing frameworks accept inputs directly from the schema of the ontologies which make their running time higher as a result of too much comparison within the instances of source and target ontologies. Integration of a component that can enable discovery of potential matching attributes at the initial stage of the matching will make matching process more efficient in terms of execution time when generating final alignments between instances of given ontologies.

2. To propose an effective property alignment algorithm that can control irregular data found in the generated training sets that can support self-configuration during similarity generation. Property alignment component controls irregular data found in an ontology. These irregular data are similar to noise found in traditional databases. However, the heterogeneous nature of semantic web content makes control of these irregular data difficult in a linked data settings. Therefore, a powerful method is required when matching instances of ontologies with maximum accuracy. This method may be powerful if it incorporate self-configuration behavior. By self-configuration, it means that the matching framework can automatically perform parameter adjustment without user intervention and can scale to the increase in data size.

1.5 Research Scope

This study cantered around developing an instance matching framework for the generation of alignment in large-scale RDF-based data without loss of attributes during the process, which is the limitation of most existing matching frameworks. The framework introduces the attributes discovery component that uses clustering technique to accommodate serialized input data to determine the potential matching attributes from the candidate matching ontologies. The framework improves the process of training set generation using no-primitive recursive *Ackermann function* to improve performance of both training set generator and property aligner. The framework also includes self-configuration algorithm to improve the performance of similarity execution model via a two-fold phases: training phase and retrieval phase. Self-configuration behavior of the made it capable of adjusting parameter automatically within the constraints of the system's functionality. The output of this framework are ***sameAs*** links, which are the alignment between two candidate ontologies that can represent the same real-world entity. Effectiveness and efficiency of the outputs are measured at each level of development in terms of accuracy and running time respectively to determine its performance to the heterogeneity and scalability requirements of instance matching.

1.6 Research Contributions

The contributions of this study are summarized as follows:

1. Introduction of attribute discovery method to resolve the problem of missing attributes during the entire matching process. This is done via a proposed clustering method (K-Medoid) that partitions the input data into similar and non- similar clusters during the initial process in a framework. This is a novel component integrated into instance matching framework.

2. Minimize the amount of comparisons when generating training sets by improving the computational performance of the training set algorithm to accommodate all potential matching attributes produced in objective one. The training set generation (TSG) algorithm works in an unsupervised manner to effectively handle unstructured RDF-based data.

3. Improve the performance of property alignment algorithm in handling irregular data using the generated training sets as input. This component bootstraps the performance of proposed matching framework by efficiently handled any unwanted data that may unintentionally be in the generated training sets.

4. Introducing self-configuration behavior in detecting similarities when generating final alignments to guarantee auto-adjustment of the defined parameters that changes the state of the system within its functionality constraints. To the best of our knowledge, this is a novel approach used for similarity detection in an instance matching framework. In all the methods so

far reviewed, the parameters are defined to be manually configured which in turn affects automation.

5. Provides alignment generation algorithms for Semantic Web developments that requires run-time data matching over Linked Data environment. The algorithm proposed in this framework can independently be used by semantic web designers to establish link between both internal and external knowledge-bases at run-time over LD environment.

The contributions of this study on linked data environment are highlighted as blue and red colours as shown in Figure 1.4. The contributions labeled with red colour signifies the main contributions while the blue colour ones are auxiliary contributions of the study.

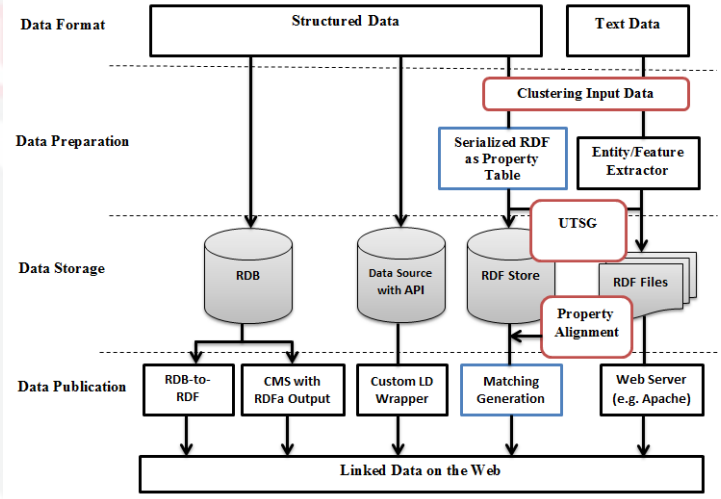

Figure 1.4: Research Contributions on Linked Data Environment

1.7 Organization of the Thesis

The rest of this thesis is organized as follows: Literature review on the background of this research is presented in Chapter 2. This chapter describes the background of ontology, ontology matching and instance-based ontology matching with respect to semantic webs and linked data. Chapter 3 provides the methodology used in conducting this research. Chapter 4 discusses the general instance matching approach while Chapter 5 presents the results of the experiments conducted to evaluate the approach. The thesis is concluded in Chapter 6 with possible future directions of this study.

1.8 Chapter Summary

This chapter presents the general background of this thesis, the motivations behind the conduct of this research, some research questions addressed, statement of the problem as well as the research objectives achieved. It also presents the scope within which the research is limited, its significance to the research community and the general public as well, the contribution of the work over linked data environment and lastly the organisational structure of the remaining parts of the thesis.

REFERENCES

Abanda, F. H., Tah, J. H. M. and Keivani, R. (2013) 'Trends in built environment Semantic Web applications: Where are we today?', *Expert Systems with Applications*. Elsevier Ltd, 40(14), pp. 5563–5577. doi: 10.1016/j.eswa.2013.04.027.

Altnel, B., & Ganiz, M. C. (2016). A new hybrid semi-supervised algorithm for text classification with class-based semantics. *Knowledge-Based Systems*, *108*, 50–64. https://doi.org/10.1016/j.knosys.2016.06.021

Anam, S., Kim, Y. S., Kang, B. H., & Liu, Q. (2015). Review of Ontology Matching Approaches and Challenges. *International Journal of Computer Science & Network Solutions*, *3*(3). Retrieved from http://www.ijcsns.com/March.2015-Volume.3-No.3//Article01.pdf

Araujo, B. and Zhao, L. (2016) 'Data heterogeneity consideration in semi-supervised learning', *Expert Systems with Applications*, 45, pp. 234–247. doi: 10.1016/j.eswa.2015.09.026.

Bailey, J. *et al.* (2005) 'Web and Semantic Web Query Languages: A Survey', pp. 35–133. doi: 10.1007/11526988_3.

Balaji, J. and Sunderraman, R. (2015) 'Scalable storage structure for pattern matching on big graph data', *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 1848–1855. doi: 10.1109/BigData.2015.7363958.

Bao, J. *et al.* (2015) 'RIMM: A novel map matching model with rotational invariance', *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*. doi: 10.1109/DSAA.2015.7344891.

Berners-lee, T. I. M., Hendler, J. and Lassila, O. R. A. (2001) 'The Semantic Web will enable machines to', *Scientific American*, (May), pp. 1–4.

Bilke, Alexander, F. N. (2015) 'Semi-supervised instance matching using boosted classifiers', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 388–402. doi: 10.1007/978-3-319-18818-8_24.

Bilke, A. and Naumann, F. (2005) 'Schema Matching using Duplicates', *21st International Conference on Data Engineering (ICDE'05)*, pp. 69–80. doi: 10.1109/ICDE.2005.126.

Cardoso, J., Hepp, M. and Lytras, M. (2004) 'Chapter 1 REAL WORLD APPLICATIONS OF SEMANTIC', *Information Systems Research*, (Rdf 2002).

Castano, S. *et al.* (2011) 'Ontology and instance matching', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6050, pp. 167–195. doi: 10.1007/978-3-642-20795-2_7.

Choi, D. W. and Chung, C. W. (2017) 'A K-partitioning algorithm for clustering large-scale spatio-textual data', *Information Systems*. Elsevier, 64(June 2016), pp. 1–11. doi: 10.1016/j.is.2016.08.003.

Cruz, I. F. *et al.* (2013) 'Building linked ontologies with high precision using subclass mapping discovery', *Artificial Intelligence Review*, 40(2), pp. 127–145. doi:

10.1007/s10462-012-9363-x.

Datar, M., Immorlica, N., Indyk, P., & Mirrokni, V. S. (2004). Locality-sensitive hashing scheme based on p-stable distributions. *Proceedings of the Twentieth Annual Symposium on Computational Geometry - SCG '04*, 253. https://doi.org/10.1145/997817.997857

Dallora, A. L. *et al.* (2016) 'Prognosis of Dementia Employing Machine Learning and Microsimulation Techniques: A Systematic Literature Review', *Procedia Computer Science*. Elsevier Masson SAS, 100, pp. 480–488. doi: 10.1016/j.procs.2016.09.185.

Daskalaki, E. *et al.* (2016) 'Instance matching benchmarks in the era of Linked Data', *Web Semantics: Science, Services and Agents on the World Wide Web*. Elsevier B.V., 39, pp. 1–14. doi: 10.1016/j.websem.2016.06.002.

Deb Nath, R. P., Seddiqui, H. and Aono, M. (2012) 'Resolving scalability issue to ontology instance matching in Semantic Web', *Proceeding of the 15th International Conference on Computer and Information Technology, ICCIT 2012*, pp. 396–404. doi: 10.1109/ICCITechn.2012.6509778.

Devi, M. U. and Gandhi, G. M. (2015) 'An Enhanced Fuzzy Clustering and Expectation Maximization Framework based Matching Semantically Similar Sentences', in *Procedia Computer Science*. doi: 10.1016/j.procs.2015.07.406.

Diallo, G. (2014) 'An effective method of large scale ontology matching.',*Journal of biomedical semantics*, 5(1), p. 44. doi: 10.1186/2041-1480-5-44.

Duan, S. *et al.* (2012) 'Instance-based matching of large ontologies using locality-sensitive hashing', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7649 LNCS(PART 1), pp. 49–64. doi: 10.1007/978-3-642-35176-1-4.

DuyHoa, N., Bellahsene, Z. and Coletta, R. (2011) 'A flexible system for ontology matching', *CEUR Workshop Proceedings*, 734, pp. 73–80. doi: 10.1007/978-3-642-29749-6_6.

Euzenat, Jérôme and Shvaiko, P. (2012). Ontology matching. In *Springer-Verlag Berlin Heidelberg* (pp. 73–85). https://doi.org/10.1007/978-3-642-38721-0

Euzenat, J., & Shvaiko, P. (2007). Ontology matching. *Heidelberg: Springer.*, (i), 333. https://doi.org/10.1007/978-3-540-49612-0

Fan, Z., Euzenat, J., & Scharffe, F. (2014). Learning concise pattern for interlinking with extended version space. *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*, *1*(3), 189–204. https://doi.org/10.1109/WI-IAT.2014.18

Faria, D. *et al.* (2013) 'The AgreementMakerLight ontology matching system', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8185 LNCS, pp. 527–541. doi: 10.1007/978-3-642-41030-7_38.

Faria, D. *et al.* (2016) 'OAEI 2016 Results of AML'.

Farooq, A., Ahsan, S. and Shah, A. (2010) 'An Efficient Technique for Similarity Identification between Ontologies', *Computing*, 2(6), pp. 147–155.

Gherbi, S. and Khadir, M. T. (2016) 'Inferred Ontology Concepts Alignment Using Instances and an External Dictionary', *Procedia Computer Science*. Elsevier Masson SAS, 83(Ant), pp. 648–652. doi: 10.1016/j.procs.2016.04.145.

Gherbi, S. and Khadir, M. T. (2017) 'ONTMAT: Results for OAEI 2017', *CEUR Workshop Proceedings*, 2032, pp. 166–170.

Gracia, J. and Mena, E. (2012) 'Semantic heterogeneity issues on the web', *IEEE Internet Computing*, 16(5), pp. 60–67. doi: 10.1109/MIC.2012.116.

Gruber, T. R. (1993) 'A translation approach to portable ontology specifications' *Knowledge Acquisition*, pp. 199–220. doi: http://dx.doi.org/10.1006/knac.1993.1008.

Gu, L. and Baxter, R. (2003) 'Record linkage: Current practice and future directions', *Cmis*, p.03/83. Available at: http://festivalofdoubt.uq.edu.au/papers/record_linkage.pdf.

Hadj Taieb, M. A., Ben Aouicha, M. and Ben Hamadou, A. (2014) 'Ontology-based approach for measuring semantic similarity', *Engineering Applications of Artificial Intelligence*. Elsevier, 36, pp. 238–261. doi: 10.1016/j.engappai.2014.07.015.

Hakimpour, F. and Geppert, A. (2001) 'Resolving semantic heterogeneity in schema integration: An ontology based approach', *Formal Ontology in Information Systems: Collected Papers from the Second International Conference*, pp. 297–308. doi: 10.1145/505168.505196.

Hertling, S., Portisch, J., & Paulheim, H. (2020). Supervised ontology and instance matching with MELT. *CEUR Workshop Proceedings*, *2788*(September), 60–71.

Hopke, P. K. (1990) 'The Use of Sampling to Cluster Large Data Sets', *Chemometrics and Intelligent Laboratory Sytems*, 8, pp. 195–204.

Hu, W. *et al.* (2010) 'ObjectCoref & Falcon-AO : Results for OAEI 2010'.

Hu, W., Qiu, H. and Dumontier, M. (2015) 'Link Analysis of Life Science Linked Data', *The Semantic Web - ISWC 2015 SE - 29*, 9367, pp. 446–462. doi: 10.1007/978-3-319-25010-6_29.

Huang, Z. (1998) 'Extensions to the k -Means Algorithm for Clustering Large Data Sets with Categorical Values', *Data Mining and Knowledge Discovery*, 304(2), pp. 283–304.

Isaac, A. *et al.* (2008) 'An empirical study of instance-based ontology matching', *Belgian/Netherlands Artificial Intelligence Conference*, pp. 317–318. doi: 10.1007/978-3-540-76298-0_19.

Jean-Mary, Y. R., Shironoshita, E. P. and Kabuka, M. R. (2009) 'Ontology matching with semantic verification', *Journal of Web Semantics*, 7(3), pp. 235–251. doi: 10.1016/j.websem.2009.04.001.

Jiménez-Ruiz, E. and Cuenca Grau, B. (2011) 'LogMap: Logic-based and scalable ontology matching', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7031 LNCS(PART 1), pp. 273–288. doi: 10.1007/978-3-642-25073-6_18.

Joachims, T. (1999) 'Transductive Inference for Text Classification using Support Vector Machines', *16th International Conference on Machine Learning (ICML-99)*, pp. 200–209. doi: 10.4218/etrij.10.0109.0425.

Kejriwal, M., & Miranker, D. P. (2014). A two-step blocking scheme learner for scalable link discovery. *CEUR Workshop Proceedings*, *1317*, 49–60.

Kejriwal, M., & Miranker, D. P. (2015a). An unsupervised instance matcher for schema-free RDF data. *Journal of Web Semantics*, *35*, 102–123. https://doi.org/10.1016/j.websem.2015.07.002

Kejriwal, M., & Miranker, D. P. (2015b). Decision-making bias in instance matching model selection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9366*, 392–407. https://doi.org/10.1007/978-3-319-25007-6_23

Khiat, A. and Benaissa, M. (2014) 'InsMT / InsMTL Results for OAEI 2014 Instance Matching', in *CEUR Workshop Proceedings (Vol. 1545,). CEUR-WS.*, pp. 158–161.

Khiat, A. and Benaissa, M. (2015) 'InsMT + Results for OAEI 2015 Instance Matching', (1).

Khiat, A., Benaissa, M. and Belfedhal, M. A. (2015) 'STRIM Results for OAEI 2015 Instance Matching Evaluation'.

Khune, R. R. (2015) 'Mapping of Semantic Web Ontology in User Query System', *International Journal of Computer Applications*, 111(14), pp. 975–8887. doi: 10.5120/19604-1456.

Kolyvakis, P., Kalousis, A. and Kiritsis, D. (2018) 'DeepAlignment : Unsupervised Ontology Matching With Refined Word Vectors', in, pp. 787–798.

Konen, W. (2011) 'Self-configuration from a Machine-Learning Perspective', (May 2011). Available at: http://arxiv.org/abs/1105.1951.

Konen, W. *et al.* (2011) 'Tuned data mining: A benchmark study on different tuners', *Genetic and Evolutionary Computation Conference, GECCO'11*, (January), pp. 1995–2002. doi: 10.1145/2001576.2001844.

Lars C., Rafael Schimassek, Dominik Huser, Maximilian Peters, Christoph Kramer, M.

C. and S. D. (2018) 'SchemaTree: Maximum-Likelihood Property Recommendation for Wikidata', *Springer Nature Switzerland*, 1(May), pp. 0–16. doi: 10.1007/978-3-030-49461-2.

Lee, S. *et al.* (2016) 'An Efficient Parallel Graph Clustering Technique Using Pregel', pp.370–373.

Lee, W. P. and Ma, C. Y. (2016) 'Enhancing collaborative recommendation performance by combining user preference and trust-distrust propagation in social networks', *Knowledge-Based Systems*. Elsevier B.V., 106, pp. 125–134. doi: 10.1016/j.knosys.2016.05.037.

Legaz-García, M. del C. *et al.* (2015) 'Transformation of standardized clinical models based on OWL technologies: from CEM to OpenEHR archetypes', *Journal of the American Medical Informatics Association : JAMIA*, 22(3), pp. 536–544. doi: 10.1093/jamia/ocu027.

Li, J. *et al.* (2009) 'RiMOM : A Dynamic Multistrategy Ontology Alignment Framework', 21(8), pp. 1218–1232.

Li, J. *et al.* (2013) 'Large scale instance matching via multiple indexes and candidate selection', *Knowledge-Based Systems*. Elsevier B.V., 50, pp. 112–120. doi: 10.1016/j.knosys.2013.06.004.

Li, L. *et al.* (2016) 'Entropy-Weighted instance matching between different sourcing points of interest', *Entropy*, 18(2), pp. 1–15. doi: 10.3390/e18020045.

Liu, L. *et al.* (2012) 'SVM-based ontology matching approach', *International Journal of Automation and Computing*, 9(3), pp. 306–314. doi: 10.1007/s11633-012-0649-x.

Lyu, X. *et al.* (2017) 'NjuLink: Results for instance matching at OAEI 2017', *CEUR Workshop Proceedings*, 2032, pp. 158–165.

Maree, M., & Belkhatir, M. (2015). Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies. *Knowledge-Based Systems*, *73*, 199–211. https://doi.org/10.1016/j.knosys.2014.10.001

Mccallum, A. and Ungar, L. H. (2013) 'Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching', 64(213–223).

Müller, L. *et al.* (2019) 'An open access medical knowledge base for community driven diagnostic decision support system development', *BMC Medical Informatics and Decision Making*. BMC Medical Informatics and Decision Making, 19(1), pp. 1–7. doi: 10.1186/s12911-019-0804-1.

Nentwig, M., Soru, T. and Ngomo, A. N. (2015) 'The Semantic Web: ESWC 2012 Satellite Events', 7540(August). doi: 10.1007/978-3-662-46641-4.

Ngo, D., Bellahsene, Z. and others (2012) 'Yam++-a combination of graph matching and machine learning approach to ontology alignment task', *Journal of Web Semantics*, 16(August 2016).

Noy, N. F. and McGuinness, D. L. (2001) 'Ontology Development 101: A Guide to Creating Your First Ontology', *Stanford Knowledge Systems Laboratory*, p. 25. doi: 10.1016/j.artmed.2004.01.014.

Patel-Schneider, P. F. (2014) 'Using Description Logics for RDF Constraint Checking and Closed-World Recognition', *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 247–253. Available at: http://arxiv.org/abs/1411.4156.

Pinto, H. and Martins, J. (2001) 'Ontology integration: How to perform the process', *IJCAI2001 Workshop on Ontologies and Information Sharing*.

Pirró, G., & Talia, D. (2010). UFOme: An ontology mapping system with strategy prediction capabilities. *Data and Knowledge Engineering*, *69*(5), 444–471. https://doi.org/10.1016/j.datak.2009.12.002

Rattanasawad, T., Saikaew, K. R., Buranarach, M., & Supnithi, T. (2013). A review and comparison of rule languages and rule-based inference engines for the Semantic Web. *2013 International Computer Science and Engineering Conference, ICSEC 2013*, (September), 1–6. https://doi.org/10.1109/ICSEC.2013.6694743

Rosenthal, A., & Seligman, L. (2001). Scalability issues in data integration. *Proceedings of the AFCEA Federal*, (January 2001). Retrieved from

http://wwwsrv2.mitre.org/work/tech_papers/tech_papers_01/rosenthal_scalability/rosenthal_scalability.pdf

R, M.Sai Baba, N.Madurai Meenachi, B. P. (2016) 'Instance Based Matching System for Nuclear Ontologies', 4(1), pp. 10–13.

Rong, S. *et al.* (2012) 'A Machine Learning Approach for Instance Matching Based on Similarity Metrics', *The 11th International Semantic Web Conference (ISWC'12)*, pp. 1–16.

Sa, F. (2010) 'LN2R – a knowledge based reference reconciliation system : OAEI 2010 Results'.

Saif, R. A. A. & A. (2016) 'Comparative Study on Cloud Portability and Interoperability using Semantic Representation', in, pp. 1–6.

Samur, A. *et al.* (2010) 'SERIMI – Resource Description Similarity, RDF Instance Matching and Interlinking', (Oaei), pp. 3–4. Sanfilippo, E. M., & Borgo, S. (2016). What are features? An ontology-based review of the literature. *CAD Computer Aided Design*, *80*, 9–18. https://doi.org/10.1016/j.cad.2016.07.001

Shao, C. *et al.* (2016) 'RiMOM-IM: A Novel Iterative Framework for Instance Matching', *Journal of Computer Science and Technology*, 31(1), pp. 185–197. doi: 10.1007/s11390-016-1620-z.

Song, D., Luo, Y. and Heflin, J. (2016) 'Linking Heterogeneous Data in the Semantic Web Using Scalable and Domain-Independent Candidate Selection', *IEEE Transactions on Knowledge and Data Engineering*, PP(99), pp. 1–14. doi: 10.1109/TKDE.2016.2606399.

Sowmya Kamath, S. *et al.* (2014) 'Similarity analysis of service descriptions for efficient Web service discovery', *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*, pp. 142–148. doi: 10.1109/DSAA.2014.7058065.

Sun, S., & Iglesias, C. A. (2019). Instance matching in the context of the Semantic Web : A systematic review Instance matching in the context of the Semantic Web : A systematic review, (April).

Taaposa, A., & Abdullah, M. S. (2011). Goal-ontology approach for modeling and designing ETL processes. *Procedia Computer Science*, *3*, 942–948. https://doi.org/10.1016/j.procs.2010.12.154

Taheri, A. and Shamsfard, M. (2013) 'Instance Coreference Resolution in Multi-ontology Linked Data Resources', pp. 129–145.

Tian, A., Kejriwal, M., & Miranker, D. P. (2014). Schema matching over relations, attributes, and data values, 1–12. https://doi.org/10.1145/2618243.2618248

Todorov, K., Geibel, P. and Kuehnberger, K.-U. (2010) 'Extensional Ontology Matching with Variable Selection for Support Vector Machines', *Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on*. doi: 10.1109/CISIS.2010.59.

Tunis, U. De *et al.* (2015) 'EXONA Results for OAEI 2015 and Sami Zghal'.

Umbrich, J. *et al.* (2010) 'Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources', *Ldow 2010*.

Vatsalan, D., Christen, P. and Verykios, V. S. (2013) 'A taxonomy of privacy-preserving record linkage techniques', *Information Systems*. Elsevier, 38(6), pp. 946–969. doi: 10.1016/j.is.2012.11.005.

Wagstaff, K., Rogers, S. and Schroedl, S. (2001) 'Constrained K-means Clustering with Background Knowledge', pp. 577–584.

Wang, Y. *et al.* (2017) 'An efficient semi-supervised representatives feature selection algorithm based on information theory', *Pattern Recognition*. Elsevier, 61, pp. 511– 523. doi: 10.1016/j.patcog.2016.08.011.

Yang, X. S. *et al.* (2013) 'A framework for self-tuning optimization algorithm', *Neural Computing and Applications*, 23(7–8), pp. 2051–2057. doi: 10.1007/s00521-013-1498-4.

Zhao, L. and Ichise, R. (2014) 'Ontology Integration for Linked Data', *Journal on Data Semantics*, 3(4), pp. 237–254. doi: 10.1007/s13740-014-0041-9.

Zheng, J. G. *et al.* (2015) 'SEM+: tool for discovering concept mapping in Earth science related domain', *Earth Science Informatics*, 8(1), pp. 95–102. doi: 10.1007/s12145- 014-0203-1.