Ontology alignment

Simulated annealing-based system, statistical evaluation, and application to logistics interoperability

Mohammadi, Majeed

DOI
10.4233/uuid:7d8ac519-f3f7-425f-82ce-1df481bc1c34

Publication date
2020

Document Version
Final published version

Citation (APA)
Mohammadi, M. (2020). *Ontology alignment: Simulated annealing-based system, statistical evaluation, and application to logistics interoperability* . [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:7d8ac519-f3f7-425f-82ce-1df481bc1c34

ONTOLOGY ALIGNMENT

SIMULATED ANNEALING-BASED SYSTEM, STATISTICAL EVALUATION, AND APPLICATION TO LOGISTICS INTEROPERABILITY

ONTOLOGY ALIGNMENT

SIMULATED ANNEALING-BASED SYSTEM, STATISTICAL EVALUATION, AND APPLICATION TO LOGISTICS INTEROPERABILITY

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. T. H. J. J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Friday 28 February 2020 at 10:00 o'clock

by

Majeed MOHAMMADI

Master of Science in Computer Engineering - Artificial Intelligence,
Ferdowsi University of Mashhad, Iran
born in Mashhad, Iran.

This dissertation has been approved by the promotors.

promoter: Prof. dr. Y. H. Tan
co-promoter: Dr. ir. W. J. Hofman

Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof. dr. Y. H. Tan	Delft University of Technology
Dr. W. J. Hofman	The Netherlands Organisation for Applied Scientific Research (TNO)

Independent members:

Prof. dr. ir. Jan van den Berg	Delft University of Technology and Leiden University
Prof. dr. J. R. Franklin	Kühne Logistics University
Dr. J. Rezaei	Delft University of Technology
Dr. M. Cheatham	Wright State University
Dr. E. Jimenez-Ruiz	City, University of London and University of Oslo
Prof. dr. ir. M.F.W.H.A. Janssen	Delft University of Technology (reserve member)

Keywords: Ontology alignment, simulated annealing, logistics, interoperability, comparison, evaluation, Bayesian, MCDM, SANOM.
Printed by: Gilderprint
Front & Back: Mahdi Ariani (Instagram:@mahdi_ariani) and Gahan Wilson

ISBN 978-94-6402-110-3

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

The research in this dissertation was funded by the Netherlands Organisation for Scientific Research (NWO) under grant number 438-13-601 of the project Scalable Interoperability in Information Systems for Agile Supply Chains (SIISASC).

To my mother
For her endless love and unsparing support
And to my F^2
My only known in a pile of equations

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

SUMMARY

The primary motivation of this dissertation is to investigate how to enable interoperability in the logistics domain by the aid of ontology alignment. More in detail, the primary research objective of this dissertation is

> *To address interoperability between heterogeneous IT systems in logistics by using ontology alignment.*

To accomplish the objective, we first look into the literature of ontology alignment using a quantitative approach to get a thorough understanding of the available literature and its progress. We particularly identify several research gaps that are studied in the subsequent chapters that serve the objective of this dissertation. An important lesson learned from the literature analysis is that there are two segregated communities that form ontology alignment but work independently and with minimal interactions with each other.

Based on the identified research gaps, we develop a new system, SANOM (simulated annealing-based ontology matching), that addresses the non-deterministic polynomial-time (NP) ontology alignment problem based on the well-known evolutionary algorithm, simulated annealing. SANOM is equipped with an extended Soft TF-IDF (term frequency-inverse document frequency) string similarity metric that can also detect linguistic similarity among the names of entities in two ontologies in question. Structural similarity metrics are also taken into account that increase the alignment performance for more complex ontologies. Simulated annealing with a warm initialization is used as the matching strategy to find an optimal solution to the ontology alignment problem. The experiments show that SANOM has a very competitive performance with the best systems that participated in the ontology alignment evaluation initiative (OAEI) and is particularly faster than other evolutionary algorithm-based alignment systems.

To come to a better understanding of which alignment system is preferred, we develop several methods for evaluation and comparison of alignment systems using different statistical techniques and multi-criteria decision-making methods (MCDM). We first study the frequentist approach for comparing alignment systems. More in detail, we compare different statistical tests for comparing alignment systems over single or multiple benchmarks and propose a proper test based on the number of benchmarks and alignment systems. While these techniques are more reliable than those being currently in use, it suffers from the drawbacks of making decisions based on p-values, which can be addressed by Bayesian statistics.

If only the performance scores of alignment systems for different benchmarks are available, then the Bayesian tests counterpart to those recommended in the frequentist approach can be used for comparison. However, if the alignments generated by systems are available, then Bayesian statistics has more flexibility to take into account the alignments (and not performance scores) to compare and evaluate the associated systems.

In this regard, we develop a Bayesian model to evaluate the alignment systems based on a user-defined error function, which is a function of false positives and false negatives. According to this evaluation, a new Bayesian test is developed to compare the systems and compute the extent to which one alignment system is superior to another.

Despite the effectiveness of the proposed Bayesian model, it bases the comparison on one performance score. For comparing based on multiple performance scores, two classes of MCDM methods are used. First, the preferences of ontology alignment experts are elicited for the performance scores of different matching tasks, especially the OAEI tracks. The preferences are then translated into the priorities that calibrate the importance of each metric for each OAEI track. Second, the priorities are used to rank the alignment systems by using MCDM outranking methods. Since these methods rank alignment systems in a different and potentially conflicting way, a new ensemble method is proposed to aggregate the rankings of outranking methods and compute final rankings for the alignment systems in each OAEI track.

We finally apply ontology alignment to interoperability in the logistics domain, which is characterized by numerous stakeholders, each with their own ontology implemented by a database scheme. Although these ontologies have a relatively low number of concepts compared to large ontologies in domains like biomedicine, the large number of alignments (millions) that have to be obtained is a challenge. Another challenge for applying ontology alignment is that the domain has several standards without semantics, since they are not developed by using Semantic Web technologies. To experiment with ontology alignment, we create two ontologies from shipping information (SI) and electronic CMR (eCMR) data models and subject them to ontology alignment systems to find the shared entities of ontologies. Since the created ontologies use distinct terminologies from each other, the direct matching of ontologies with the most top available alignment systems is not satisfactory. Hence, we conduct indirect matching through an existing upper ontology with some annotations. The indirect matching with annotated upper ontology significantly improves the outcome of the alignment systems. The results of this experiment show that ontology alignment can address the interoperability in logistics provided that a proper logistics background knowledge, e.g., a proper logistics upper ontology or dictionary, is used. Overall, ontology alignment can enable interoperability in logistics if one of the following conditions holds:

- A proper background knowledge for logistics is developed manually.
- An existing upper ontology is used and manually annotated by an expert with the terminologies of the given ontologies for alignment.
- Many alignment experiments are conducted by different logistics standards and data models, and an ontology is annotated based on the generated alignments.

The methodologies for comparing and evaluating ontology alignment systems can be used in any domain with some standard benchmarks with known reference alignment. In logistics, however, we learned that the direct alignment of ontologies does not bear satisfactory outcome, while the indirect matching with annotated background knowledge can generate acceptable alignments, regardless of the matching system being used. In addition, more logistics benchmarks for ontology alignment must be created so

that the methodologies for evaluating and comparing alignment systems can be used for selecting the most appropriate alignment systems for enabling interoperability in logistics.

SAMENVATTING

De primaire motivatie voor dit proefschrift is het onderzoeken naar mogelijke interoperabiliteit in het logistieke domein met behulp van ontology alignment. Preciezer gezegd: het primaire onderzoeksdoel van dit proefschrift is::

> Om interoperabiliteit tussen heterogene IT-systemen in de logistiek aan te adresseren met behulp van ontologie afstemming.

Om het doel te bereiken kijken we eerst kwantitatief naar de literatuur van ontology alignment, om een grondig inzicht te krijgen in de beschikbare literatuur en de vooruitgang daarin. In het bijzonder wijzen we een aantal onderzoekslacunes aan, die in de volgende hoofdstukken worden bestudeerd ten behoeve van de doelstellingen van dit proefschrift. Een belangrijke les uit de literatuuranalyse is dat er twee gescheiden community's aan ontology alignment werken, maar dat zij dit onafhankelijk van elkaar doen en met nauwelijks enige interactie.

Op basis van de geïdentificeerde onderzoekslacunes ontwikkelen we SANOM (simulated annealing-based ontology matching), een nieuw systeem dat het probleem van niet-deterministische polynomiale-tijd (NP) ontology alignment aanpakt op basis van het bekende evolutionaire algoritme van gesimuleerd uitgloeien (simulated annealing). SANOM wordt voorzien van een uitgebreide Soft TF-IDF (term frequency-inverse document frequency) metriek voor de gelijkenis tussen strings, een metriek die ook taalkundige gelijkenis tussen de namen van entiteiten in twee ontologieën kan detecteren. Ook kijken we naar metrieken voor structurele gelijkenis, die zorgen dat de afbeelding beter werkt bij complexere ontologieën. Gesimuleerd uitgloeien met 'warme initialisatie' wordt gebruikt als matchstrategie om een optimale oplossing te vinden voor het ontology-alignmentprobleem. De experimenten laten zien dat SANOM zeer goed presteert vergeleken met de beste systemen die aan bod zijn gekomen in het Ontology Alignment Evaluation Initiative (OAEI), en dat SANOM in het bijzonder sneller is dan andere op evolutionaire algoritmen gebaseerde afbeeldingssystemen.

Om beter te begrijpen welk afbeeldingssysteem de voorkeur geniet, ontwikkelen we verschillende methoden voor de evaluatie en vergelijking van afbeeldingssystemen. Hiervoor gebruiken we verschillende statistische technieken en multicriteria-besluitvormingmethoden (multi-criteria decision-making methods, MCDM). Eerst bestuderen we de frequentistische benadering van het vergelijken van afbeeldingssystemen. We vergelijken in detail verschillende statistische tests voor het vergelijken van afbeeldingssystemen op basis van een of meer benchmarks en stellen een geschikte test voor die gebaseerd is op het aantal benchmarks en afbeeldingssystemen. Hoewel deze technieken betrouwbaarder zijn dan degene die nu worden gebruikt, ondervinden ze nadelen van hun besluitvorming op basis van p-waarden. Daar kan iets aan worden gedaan met Bayesiaanse statistiek.

Als alleen de prestatiescores van afbeeldingssystemen voor verschillende benchmarks beschikbaar zijn, kunnen voor de vergelijking Bayesiaanse tests worden gebruikt in plaats van de tests die daarvoor worden aanbevolen in de frequentistische aanpak. Als er echter door systemen gegenereerde afbeeldingen beschikbaar zijn, biedt Bayesiaanse statistiek meer flexibiliteit om rekening te houden met de afbeeldingen (en niet de prestatiescores) om de bijbehorende systemen te vergelijken en te evalueren. Met dit op het oog ontwikkelen we een Bayesiaans model om de afbeeldingssystemen te evalueren op basis van een zelf gedefinieerde foutfunctie, die een functie is van foutpositieven en foutnegatieven. Op grond van deze evaluatie wordt een nieuwe Bayesiaanse test ontwikkeld om de systemen te vergelijken, en om te berekenen in hoeverre het ene afbeeldingssysteem beter is dan het andere.

Hoewel het voorgestelde Bayesiaanse model effectief is, baseert het de vergelijking op slechts één prestatiescore. Voor een vergelijking op basis van meerdere prestatiescores worden twee klassen van MCDM-methoden gebruikt. Om te beginnen zoeken we uit wat de voorkeuren van de deskundigen op het gebied van ontology alignment zijn wat betreft de prestatiescores van verschillende matchtaken, met name voor de diverse OAEI-trajecten. De voorkeuren worden vervolgens vertaald naar de prioriteiten die het belang van elke score voor elk OAEI-traject bepalen. Ten tweede worden de prioriteiten gebruikt om de afbeeldingssystemen te rangschikken aan de hand van MCDM-methoden. Aangezien bij deze methoden afbeeldingssystemen op verschillende en mogelijk tegenstrijdige manieren worden gerangschikt, stellen we een nieuwe ensemblemethode voor om de ranglijsten van de afbeeldingssystemen samen te voegen en zo voor elk OAEI-traject eindranglijsten voor de afbeeldingssystemen te berekenen.

Tot slot passen we ontology alignment toe op interoperabiliteit in de logistiek. De logistiek kenmerkt zich door een groot aantal stakeholders, elk met hun eigen ontologie, geïmplementeerd door middel van een databaseschema. Hoewel deze ontologieën een relatief laag aantal concepten bevatten vergeleken met grote ontologieën in bijvoorbeeld de biogeneeskunde, vormt het grote aantal afbeeldingen (miljoenen) dat moet worden verkregen een uitdaging. Een andere uitdaging voor het toepassen van ontology alignment is dat het domein verschillende standaarden heeft zonder semantiek, omdat deze niet ontwikkeld zijn met behulp van semantische-webtechnologieën. Om te experimenteren met ontology alignment creëren we twee ontologieën van datamodellen van verzendgegevens en elektronische CMR (eCMR), en laten we hier ontology-alignmentsystemen op los om de gedeelde entiteiten van ontologieën te vinden. Aangezien deze twee ontologieën verschillende terminologieën gebruiken, is directe matching van ontologieën met de beste beschikbare afbeeldingssystemen hiervoor onvoldoende. Daarom voeren we indirecte matching uit via een bestaande basisontologie (upper ontology) die licht geannoteerd is. Het experiment met indirecte matching met een geannoteerde basisontologie leidt ertoe dat de afbeeldingssystemen een aanzienlijk beter resultaat geven. De resultaten van dit experiment tonen aan dat ontology alignment kan worden toegepast op interoperabiliteit in de logistiek, op voorwaarde dat er een goede logistieke achtergrondkennis wordt gebruikt, bijvoorbeeld een geschikte logistieke basisontologie of een goed woordenboek. Algemeen gesproken kan ontology alignment interoperabiliteit in de logistiek mogelijk maken als aan een van de volgende voorwaarden wordt voldaan:

- Er wordt handmatig goede achtergrondkennis voor de logistiek ontwikkeld.
- Er wordt een bestaande basisontologie gebruikt die door een deskundige handmatig wordt geannoteerd met de terminologieën van de gegeven ontologieën die op elkaar moeten worden afgebeeld.
- Er worden veel afbeeldingsexperimenten uitgevoerd met verschillende logistieke standaarden en datamodellen, en een ontologie wordt geannoteerd op basis van de gegenereerde afbeeldingen.

De methodologieën voor het vergelijken en evalueren van ontology-alignmentsystemen kunnen worden gebruikt in elk domein dat enkele standaard benchmarks met een bekende referentieafbeelding heeft. We hebben echter geconstateerd dat in de logistiek een directe afbeelding van ontologieën geen bevredigend resultaat oplevert, terwijl een indirecte afbeelding met geannoteerde achtergrondkennis acceptabele afbeeldingen kan genereren, ongeacht het gebruikte matchsysteem. Daarnaast moeten er meer logistieke benchmarks voor ontology alignment worden gecreëerd, zodat de methodologieën voor het evalueren en vergelijken van afbeeldingssystemen kunnen worden gebruikt voor het selecteren van de beste afbeeldingssystemen die interoperabiliteit in de logistiek mogelijk maken.

1

INTRODUCTION

Your Reality Might Not be Mine.

Poppy Crum

1

1.1. HUMAN PERCEPTION AND ONTOLOGY DESIGN

Humans constantly receive different information (e.g., sounds, pictures) from sensory receptors. The way that humans *perceive* the information is not necessarily identical that affects their interaction with the world. Perception entails the organization, interpretation, and conscious experience of the information received by sensory receptors [1]. Since this concept is more psychological, compared to sensation which is a sheer physical process, different people have distinct perceptions about the similar information received by their receptors. As an instance, a determining factor for how we perceive information is *attention*. In fact, attention plays an essential role in the human perception of information. The perception can also be influenced by many other factors, including, inter alia, beliefs, values, experiences, and expectations. The idea of *how* we perceive information and *what* influences the perception are also extensively discussed in the so-called *ladder of inference* [2].

Whether human perception is correct or not has been a philosophical question for which many philosophers have tried to provide an answer. Descartes is one of the pioneers who questioned human perception by the idea of *radical doubt*. In particular, he verified all the beliefs by radical questions to distinguish *the things we think we are certain about* from *things we are justified in being certain about*. The result of this radical doubt was the well-known *cogito argument* [3], that can be summarized in the phrase *I think, therefore, I am*. The aim of this thesis is neither about the perception nor the philosophical understanding of perception veracity, but rather about one of its influences in computer science. In fact, we accept that the perception of identical phenomena is distinct from one person to another and we seek a solution to deal with it.

We particularly consider an influence of perception on the Semantic Web, an active research area in computer science. The main aim of the Semantic Web is to represent information in a way that machines can also process them, ideally as well as humans do, so that machines are able to seek for information they require for a particular, allocated task. In this regard, *ontologies* provide a formal representation of the real world, abstract or scientific concepts [4]. An ontology is formally defined as a set of concepts with their reciprocal relationships, while in practice, they are usually a host of vocabularies or thesauri with relatively weak semantics [5]. Ontologies have manifold functionalities: They have been used to model the content message of agents in agent-based modeling [6, 7], they mapped the structure of different vocabularies in the biomedical domain so that one can simply translate among these vocabularies [8], and last but not least, they have been employed to model the goods flow and their corresponding information in the logistics domain [9]. The primary focus of this dissertation is on data sharing between ontology-based systems, where an ontology is used to model the underlying information of a particular information system.

Ontology-based modeling is subjective and is reliant on the *perception* of a designer. Since perceptions of humans are often sharply distinguished, similar concepts in one particular domain could be constructed in entirely distinct ways. The discrepancy in models is referred to as *heterogeneity*, which is a major impediment to the path to *interoperability* [10]. Interoperability refers to the interactions among different information systems in such a way that they can simply exchange data and information among each other. This difference among information systems prevents such interactions un-

less a pre-processing strategy is employed to reconcile their differences. One solution is to ask experts to find the common vocabulary and thesauri between two ontologies of heterogeneous information systems. The manual alignment of ontologies is, however, costly and time-consuming. For example, the alignment of Chinese Agricultural Thesaurus (CAT) with around 60,000 concepts to the Food and Agriculture Organization of the United Nation thesaurus (AGROVOC) with around 25,000 concepts took seven man-year of manual labor [5]. As a result, it is required to reduce the manual efforts by designing an automatic solution to find the shared concepts of two given ontologies. These automatic solutions are commonly referred to as ontology matching or ontology alignment.

1.2. ONTOLOGY ALIGNMENT: AN EXAMPLE FROM LOGISTICS

In order to explain the basic concepts of ontology alignment, a simple example from logistics is provided. The supply and logistics sector includes millions of enterprises, each with a specific data model or database schema. The discrepancy among the data models of different logistics enterprises puts a serious obstacle in their way of exchanging data and conducting business. The discrepancy between the data models of logistics companies can be addressed by the aid of ontology alignment. As a simple example, we consider two different perspectives, i.e., consignor[1] and carrier, on moving/delivering cargo or products from one location to another. Assume there are five products, shown by $P1, P2, ..., P5$, that need to be transferred from one of five locations (shown by $L1$, $L2$,..., $L5$) to one another (see Figure 1.1). The left panel of this figure shows a list of products moving from one location to another, while the right panel plots the graphical model of the same movements. The nodes of this graph are the locations, and each edge $Li \xrightarrow{P_i} Lj$ indicates that product P_i must move from Li to Lj. Since this example has been shown from the viewpoint of a person who sends the shipment, it is called the *consignor* model. We can now go to a higher level of abstraction by modeling this process from the consignor perspective. Figure 1.2 displays a potential model representing the overall transport process. In this model, a *consignment* is simply a concept repre-

[1] Consignor is typically the sender of the goods.

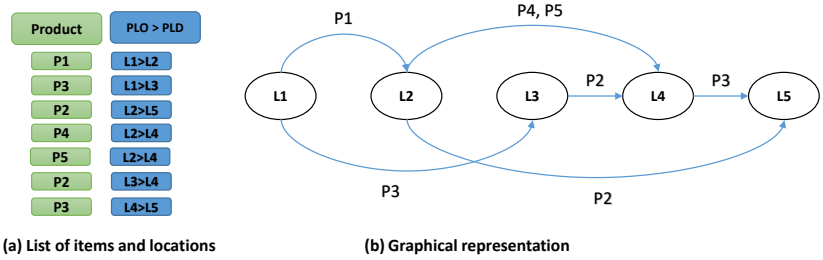


Figure 1.1: The logistics example from the consignor perspective. (a) The list of products $P1, P2, ..., P5$ that needs to move from a place of origin (PLO) to a place of destination (PLD); (b) The graphical representation of the same process, where nodes (i.e., ovals) are the locations, and each arrow with the label of a product represents the movement of that product from one end to the other.

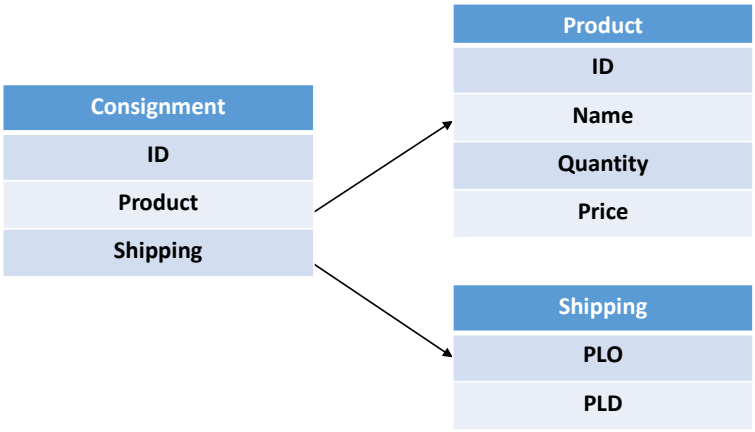


Figure 1.2: The abstractions of the logistics example from the consignor viewpoint.

senting the movement of products from one location to another at the same time (a directed edge in the graphical representation in Figure 1.1). Thus, the *consignment* entity requires a *product* and shipping information. For the product entity, we can assume several straightforward properties such as name, weight, and quantity. The shipping entity also contains the place of origin (PLO) and place of destination (PLD). Thus, the three entities in Figure 1.2 are a model to represent the instances in Figure 1.1.

The carrier view, on the other hand, is distinct from the consignor. In this perspective, there are various locations, in each of which some products might be picked up and some others may be dropped off. In this view, as a result, there is no PLO or PLD. Figure 1.3 illustrates the same example as in Figure 1.1 from the carrier viewpoint. In this figure, there are different locations in each of which two operations might happen: Picking up (PU) and/or dropping off (DO) some of the cargoes $M1, M2, ..., M5$. We deliberately use different terminology for product and cargo, since nomenclature is not essentially the same in different perspectives. In Figure 1.3-(b), the nodes of the graph are the locations, and the cargoes to pick up and drop off are shown at the top and the bottom of the nodes, respectively.

The process of transport from the carrier view can be modeled as in Figure 1.4. In this figure, the main entity is *carrier*, which has a location and two lists of picked-up and dropped-off cargoes. These lists are of the type *cargo*, which stores the basic properties of each cargo item.

We can now present the definition of *ontology* that can model logistics perspectives. The following definition is a simple yet sufficient definition of ontology for the aim of this dissertation.

Definition 1 *[11] An ontology O could be defined as a 4-tuple:*

$$O = (C, DP, OP, I),$$

where

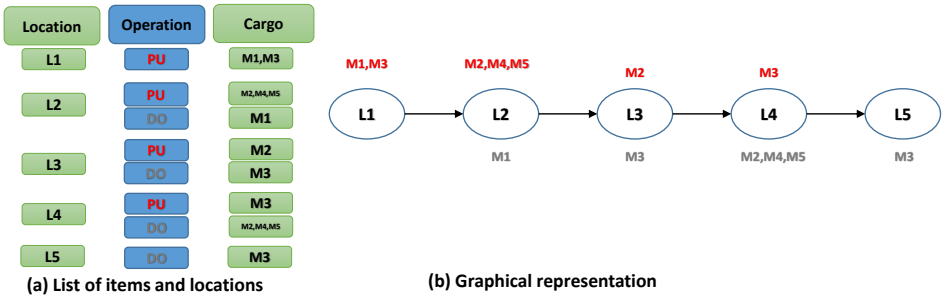


Figure 1.3: The logistics example from the transportation perspective. (a) The list of locations $L1, L2, ..., L5$, in each of which some of the cargoes are picked up (PU) or dropped off (DO); (b) The graphical representation of the same process, where nodes (i.e., ovals) are the locations, and the cargoes shown above or under them are the picked-up or dropped-off cargoes, respectively.

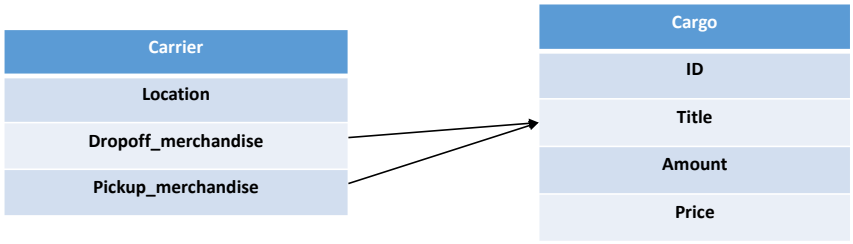


Figure 1.4: The abstractions of the logistics example from the carrier viewpoint.

- *C is a set of classes, which are the principal concepts in a domain;*
- *DP is a set of data properties explaining the characteristics of the classes;*
- *OP is a set of object properties, defining the relation of two classes;*
- *I is a set of instances, which instantiate the modeled concepts.*

In the consignor view in the previous example, *consignment* and *product* are two classes, and ID is a data property. Since each consignment contains a set of *products*, then we can define object property *contain* that relates *consignment* to *product*. The examples of movements in the left panel of Figure 1.1 are the instances in an ontology developed based on the consignor view.

The simple logistics example shows that the perception, or the view, can lead humans to model the same process in totally-distinct ways, each of which serves different purposes: One involves supplying product to locations, while the other involves the logistics operation with, for instance, a truck. Ontology alignment is used to resolve this heterogeneity by aligning the concepts of one model to those in the other. A matching of a concept from the first ontology to one in the other is called a *correspondence*, and a set of correspondences between two given ontologies is called an *alignment*. The following definitions present these two important concepts in ontology alignment.

Definition 2 (Correspondence [11]) *A correspondence between two ontologies O and O' is defined as a set of 4-tuples:*

$$< e, e', r, d >,$$

where

- *e is an entity, e.g., class, property, or instance, from the first ontology;*
- *e' is an entity form the second ontology;*
- *r is the type of relation between two entities, e.g., equivalence, subsumption;*
- *$d \in [0,1]$ is the confidence of the matching.*

Definition 3 (Alignment [11]) *An alignment is the typical outcome of an ontology matching system and consists of a set of correspondences between different entities of two given ontologies.*

Figure 1.5 displays an alignment of two ontologies in the logistics example. After identifying the alignment between two ontologies, the instances of one ontology can be transformed into the instances of the other. This process is called data transformation [12]. The alignments generated by systems should be first inspected by an expert for having a reliable transformation, especially because the correspondences in alignments have a *confidence* degree as stated in Definition 2. There is no principled approach to determine the extent to which a *confidence* is acceptable for data transformation, and determining the proper correspondences for data transformation is the expert's decision. However, for repetitive experiments, it is possible to devise a method to learn the experts' preferences and make the data transformation automatic.

Generally speaking, an ontology alignment system is an application that takes two ontologies as the inputs and uses matching techniques to generate an alignment between ontologies in question. Besides, it often uses some *resources* such as a dictionary and requires some parameters for generating final correspondences. Figure 1.6 shows the general inputs and output to a matching system in the alignment process. In this figure, O and O' are two ontologies that are matched by a matching system and A is the alignment generated by the system.

In some domains like biomedical, there are some ontologies that contain general terms in the domain. These ontologies that are called upper ontologies (also known as a top-level ontology, upper model, or foundation ontology) [13] can help increase the quality of matching between two ontologies from a domain. One way to use such upper ontologies is to first match each of the ontologies to the upper ontology, and then finding their related correspondences based on their matching with the upper ontology. This type of alignment is called *indirect matching*, that is visualized in Figure 1.7. In this figure, ontologies O and O' are first matched to an upper ontology and generate two alignments A' and A''. Then, a composition module is used to identify the alignment between O and O' based on A' and A''.

There are several matching systems available in the literature, such as LogMap [14], AML [15], and XMap [16], which can solve the ontology alignment problem efficiently.

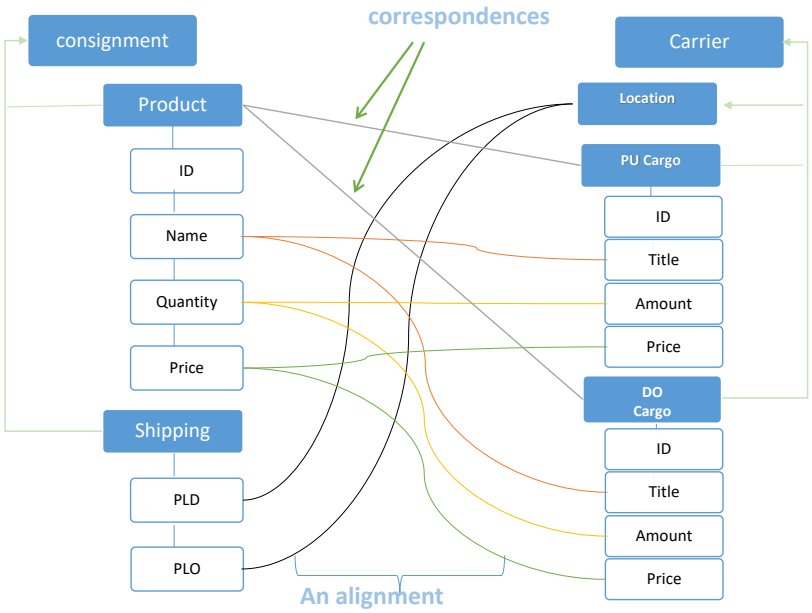


Figure 1.5: An alignment of the two simple ontologies in the logistics domain.

One way to model the ontology alignment problem is to translate it into a zero-one non-convex programming [17] that is very difficult to solve in polynomial time, if possible at all, making an approximation solution desired. One promising way is to use evolutionary algorithms (EAs) that have been extensively studied in the literature of ontology alignment [18, 19].

After identifying an alignment between two ontologies in question, we need to evaluate the generated alignment. To that end, we need to have benchmarks, which usually include pairs of ontologies with known reference alignment that represent the actual correspondences between two ontologies. Therefore, the evaluation of alignment systems is typically made by using several *performance metrics* (also referred to the score of a metric for an alignment system as *performance score*) that are a ratio directly related to true positives and true negatives of the generated alignment. The true positives and true negatives are identified by comparing the correspondences in the generated alignment to those in the reference. For example, one of the most common performance metrics is precision that is computed as the ratio of *true positives* to *true positives plus false positives*. The comparison among alignment systems are also made by juxtaposing a performance score like precision, or its average in the case of having multiple benchmarks.

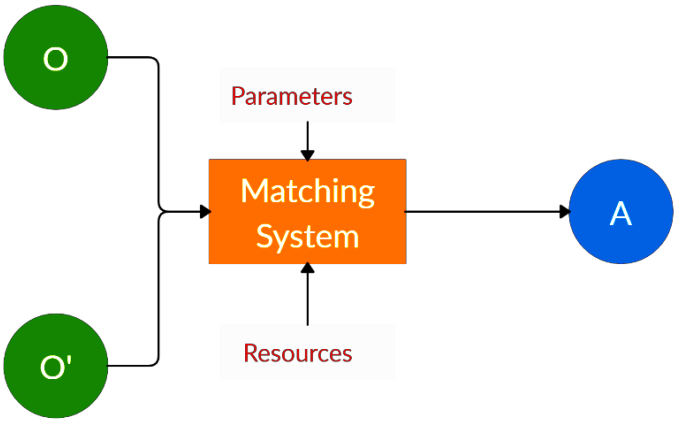


Figure 1.6: The ontology alignment process for two ontologies O and O'.

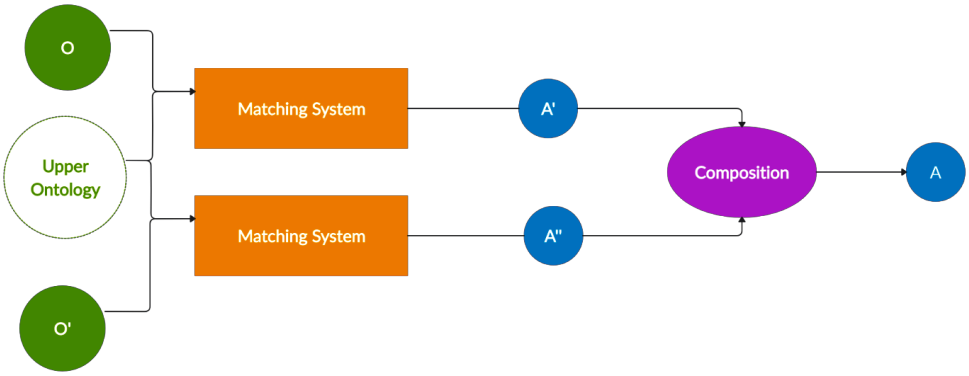


Figure 1.7: The indirect matching process of two ontologies O and O' via an upper ontology.

1.3. RESEARCH OBJECTIVE AND RESEARCH QUESTIONS

The primary motivation of this dissertation is to address interoperability in logistics with the aid of ontology alignment in order to enable IT system of logistics stakeholders to share data among each other. In principle, ontologies are used to develop IT systems, which represent how concepts in an application are linked to each other. Since ontologies are not similar for different IT systems, the data exchange is not a straightforward task and requires an intermediate step for the reconciliation of the differences. While ontology alignment is a promising strategy for such a reconciliation, it is essential to verify if the alignment systems can discover an alignment that drastically reduces the human effort to acquire the final alignment. Hence, the principal objective of this dissertation is as follows:

> *To address interoperability between heterogeneous IT systems in logistics by using ontology alignment.*

To accomplish this objective, it is required to look into two nearly independent topics, which are described in the following:

- We first need to delve into the ontology alignment literature to identify the state-of-the-art advances in line with enabling the interoperability in logistics. To that end, it is essential to identify the progress and challenges in developing ontology alignment systems in view of the objective of this thesis. In addition, it is required to have a proper evaluation and comparison means by which the alignment systems are evaluated and compared, and the most appropriate systems are selected for logistics.

- Second, it is essential to inspect the current methodologies for dealing with the heterogeneity in logistics and discuss their possible advantages/shortcomings. We then verify the applicability of ontology alignment to enabling interoperability in logistics by creating ontologies based on two logistics data models and applying the state-of-the-art alignment systems to the created ontologies.

For each of the above topics, a research question and several sub-questions are raised, whose answers can help address the objective of the dissertation. The first question involves the assessment of the state-of-the-art advances in ontology alignment with the purpose to identify potential research gaps that need to be addressed for accomplishing the research objective. The second question is the applicability of the findings from the responses to the first question to the logistics domain that includes designing logistics ontologies and checking the performance of ontology alignment systems in this domain. In the following, the research questions are discussed in more detail.

RQ1. WHAT IS THE EVOLUTION AND PROGRESS OF ONTOLOGY ALIGNMENT AND WHAT ENHANCEMENT NEEDS TO BE MADE IN VIEW OF THE RESEARCH OBJECTIVE?

First of all, we need to review and analyze the ontology alignment literature for identifying the state-of-the-art progress of the field and further finding the research gaps that can serve the objective of this dissertation. This is particularly essential since tremendous effort has been taken to improve and evolve the field of ontology alignment in the most recent two decades. As a result of such efforts, plenty of publications and materials exist for this domain that need to be analyzed in view of the objective of this dissertation.

RQ1.1. IS THE ONTOLOGY ALIGNMENT PROBLEM SOLVED IN AN EFFICIENT WAY?

Ontology alignment problem can be converted to a non-convex programming with zero-one constraints [17]. Since this problem is non-deterministic polynomial-time (NP) hard, finding its optimal solution within a reasonable time is not possible. This is particularly important in situations like matching the ontologies of the IT systems, where there are usually time and memory limitations. Hence, by answering this question, we are looking for a more efficient approach for the ontology alignment problem that can be used for logistics interoperability.

RQ1.2. CAN WE FAVOR ONE ONTOLOGY ALIGNMENT SYSTEM OVER ANOTHER?
Ontology alignment systems are currently compared based on the average of a performance score over multiple benchmarks. The decision of one alignment system being better is made simple: If the average of performance scores of an alignment system is higher than that of the other, regardless of the magnitude of the difference, then the system with a higher average is favored. However, there is no principled way to determine a difference as significant. On top of that, averages are not statistically safe and appropriate due to their sensitivity to outliers. It means that the fair (poor) performance of an alignment system over only one benchmark can compensate its poor (fair) performance over all the other benchmarks. Therefore, a methodology with more substantial evidence is required to favor one alignment system over another.

RQ1.3. CAN WE GIVE MORE MEANING TO AN INDIVIDUAL PERFORMANCE METRIC?
The performance metrics like precision are computed as figures, which are based on a ratio related to the true positives and true negatives of an alignment. For the multiple benchmarks case, the average of the performance scores is used as the overall performance indicator of an alignment system. Either computing a performance score or its average for an alignment, a figure represents the performance of a system, which is not informative enough and cannot substantiate the better performance of an alignment system. For instance, the precision of two alignment systems, one with two true positives and two true negatives and the other with 100 true positives and 100 true negatives, is identical, while the number of the correspondences is significantly different. Therefore, the current evaluation based on performance scores needs to be replaced by a methodology that is possibly more informative. Such an evaluation can lead to a more meaningful comparison between two alignment systems as well.

RQ1.4. HOW TO COMPARE ALIGNMENT SYSTEMS WITH RESPECT TO MULTIPLE PERFORMANCE METRICS?
The previous two questions solely considered the comparison and evaluation with respect to one performance metric only. However, it is typically essential to include multiple performance metrics, each indicates an aspect of accomplishment of an alignment system. For instance, a critical performance metric is execution time that needs to be included in the evaluation and comparison. Further, it is also the case that the importance of these performance metrics is distinct for different matching tasks and applications of ontology alignment. Often, experts or users would like to express their preferences over various performance metrics for one specific ontology alignment task and/or application. As a result, it is essential to incorporate these preferences into a comparison methodology based on multiple performance metrics, and subsequently, select the most appropriate alignment system.

RQ2. TO WHAT EXTENT DOES ONTOLOGY ALIGNMENT ADDRESS INTEROPERABILITY BETWEEN IT-SYSTEMS IN LOGISTICS IN PRACTICE?
Enabling interoperability in logistics using ontology alignment is the primary motivation of this dissertation. A simple example at the beginning of this chapter showed that there is a significant discrepancy between the perspectives of players or different

logistics stakeholders and that ontology alignment is a potential solution to address the present heterogeneity and to enable interoperability in logistics. To verify the applicability of ontology alignment in practice, it is essential to apply the state-of-the-art ontology alignment systems to more realistic logistics data models. The following two questions can help address the applicability of ontology alignment to logistics interoperability.

RQ2.1 WHAT ARE THE STATE-OF-THE-ART ADVANCES IN ENABLING INTEROPERABILITY IN LOGISTICS?
We first need to look into the state-of-the-art advances in enabling interoperability in logistics. In this regard, we need to check the data models and data structures currently being used in logistics, and verify if ontologies have been used for logistics interoperability. The identification of current data models, structures, and potentially logistics ontologies, is the first step of applying ontology alignment to logistics.

RQ2.2 WHAT IS THE RESULT OF APPLYING ONTOLOGY ALIGNMENT SYSTEMS TO LOGISTICS?
To check the applicability of ontology alignment to logistics, we need to apply the ontology alignment systems to different models in the domain and then analyze the outcome of alignment systems over such ontologies. We also need to verify if the outcome of an alignment system is reliable enough and can reduce the human effort to a minimum.

1.4. CONTRIBUTIONS AND GUIDE TO READERS

In particular, this dissertation consists of five different contributions to the field of ontology alignment. Chapter 2 contributes a quantitative approach to the current ontology alignment review by using bibliometric techniques. Chapter 3 contributes several matching strategies and a new ontology alignment system, which can solve ontology alignment problem more efficiently. Chapters 4-6 contribute new methodologies to comparison methods for ontology alignment systems, while Chapter 5 includes a methodology for evaluating ontology alignment systems as well. Finally, Chapter 7 contains a new alignment benchmark from logistics and a new application to which ontology alignment can be applied. The following lists give a short description of the contributions based on their relevance to ontology alignment, and how these contributions pertain to the research questions.

1.4.1. OUTLINE BY CONTRIBUTIONS

In the following, the contributions of this dissertation is listed as their relevance to the ontology alignment field.

- *A quantitative literature review.* We analyze the literature of ontology alignment by a quantitative approach. Almost all the existing ontology alignment reviews take a qualitative approach by analyzing several hundreds of ontology alignment research outputs. We instead employ a quantitative approach and analyze around 2,975 articles that pertain to ontology alignment. Different types of analyses are conducted that are described in Chapter 2.

- *Matching technique and matching system.* We develop a new matching system, called SANOM (simulated annealing-based ontology matching), by using several matching techniques, such as extended Soft TF-IDF (term frequency-inverse document frequency) and simulated annealing. Chapter 3 includes the detailed description of the proposed alignment system and the matching techniques it uses.

- *Evaluating alignment systems.* We propose a Bayesian model for evaluating ontology alignment systems. The current practice for evaluating ontology alignment systems is by computing a score. Instead, the proposed Bayesian model yields a distribution with respect to a performance metric, e.g., precision, that is significantly more informative than computing a score. Chapter 5 provides a detailed description of the rationale of using Bayesian statistics for evaluation, and gives the proposed Bayesian model.

- *Comparing alignment systems.* We propose several methodologies rooted in statistics and multi-criteria decision-making (MCDM) for comparing different alignment systems over single or multiple benchmarks. In principle, there are few studies on comparing alignment systems with no use of a proper means for comparison. We study the comparison of alignment systems by using statistical and MCDM methods in Chapters 4-6.

- *Practical application.* We also investigate the applicability of ontology alignment to logistics interoperability. It is a new problem to which ontology alignment is a potential solution. Chapter 7 is dedicated to the applicability of ontology alignment to logistics, where we study the importance of the problem, create two ontologies based on two well-known logistics data models, and apply the state-of-the-art alignment systems to verify the suitability of ontology alignment for logistics interoperability.

1.4.2. OUTLINE BY RESEARCH QUESTIONS

The following list presents the contributions of this dissertations by their pertinence to the research questions put forward in the previous section:

- *Literature review and analysis.* In response to RQ1, we review the literature of ontology alignment by inspecting 2,975 articles pertinent to this field. Based on the literature analysis, we identify several research gaps pertinent to the objective of this dissertation. Chapter 2 is dedicated to the literature review and analysis.

- *An efficient alignment system.* In response to RQ1.1, an efficient alignment system based on the simulated annealing is developed. The alignment system is significantly fast and efficient in terms of memory complexity, and has comparable results with respect to the state-of-the-art ontology alignment systems. Chapter 3 provides the description of the proposed alignment system.

- *Statistical inference for favoring one alignment system over another.* In response to RQ1.2, we study the statistical tests for comparing ontology alignment systems over single or multiple benchmarks. We also compare statistical tests with each

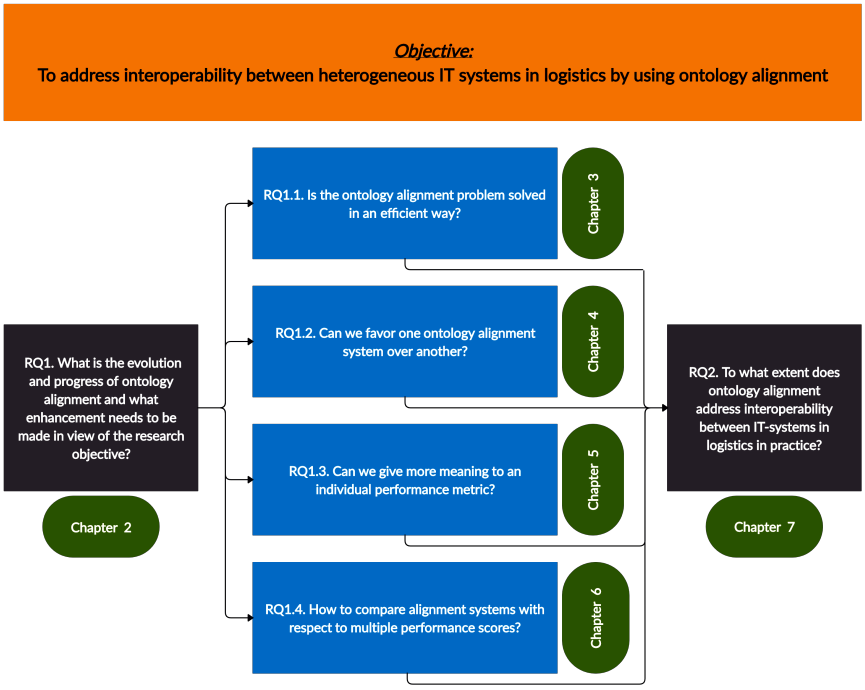


Figure 1.8: The objective of this dissertation as well as the research questions that are addressed in different chapters of this dissertation.

other and provide a list of appropriate tests for a given number of benchmarks. Chapter 4 is dedicated to comparing ontology alignment systems using statistical inference.

- *Bayesian interpretation of an individual performance metric.* In response to RQ1.3, we develop a Bayesian model that can produce a distribution with respect to each performance metric, instead of a score. A distribution with respect to a metric provides more information than a score and also help us have a more meaningful comparison among ontology alignment systems. Chapter 5 gives the details of the proposed Bayesian model, along with a Bayesian test for comparison.

- *MCDM-based comparison of alignment systems.* In response to RQ1.4, an MCDM-based methodology is proposed that can compare and rank ontology alignment systems with respect to multiple performance metrics. In addition, it accommodates the importance of each individual performance metric based on human preferences. Chapter 6 includes the proposed MCDM-based approach for comparing ontology alignment systems.

- *Ontology alignment for logistics interoperability.* In response to RQ2, we apply ontology alignment to enabling interoperability in logistics. We will first create two ontologies based on two logistics data models and subject them to the state-of-

the-art ontology alignment systems. We also discuss the proper experimental settings, e.g., indirect matching, for aligning ontologies in logistics. Chapter 7 is dedicated to addressing the logistics interoperability by using ontology alignment.

Figure 1.8 shows the research questions and the chapters that are dedicated to address them.

REFERENCES

[1] E. Styles, *Attention, perception and memory: an integrated introduction* (Psychology Press, 2004).

[2] R. Ross, *The ladder of inference,* The fifth discipline fieldbook: Strategies and tools for building a learning organization , 242 (1994).

[3] R. Descartes and L. J. Lafleur, *Meditations on first philosophy* (Bobbs-Merrill New York, 1960).

[4] G. Antoniou and F. Van Harmelen, *A semantic web primer* (MIT press, 2004).

[5] W. R. van Hage, *Evaluating ontology-alignment techniques,* (2009).

[6] V. Ermolayev and M. Davidovsky, *Agent-based ontology alignment: basics, applications, theoretical foundations, and demonstration,* in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics* (ACM, 2012) p. 3.

[7] L. Laera, V. Tamma, J. Euzenat, T. Bench-Capon, and T. Payne, *Reaching agreement over ontology alignments,* in *International Semantic Web Conference* (Springer, 2006) pp. 371–384.

[8] O. Bodenreider, *The unified medical language system (umls): integrating biomedical terminology,* Nucleic acids research **32**, D267 (2004).

[9] L. Daniele and L. F. Pires, *An ontological approach to logistics,* Enterprise Interoperability, Research and Applications in the Service-oriented Ecosystem, IWEI'13 Proceedings , 199 (2013).

[10] A. Tolk and J. A. Muguira, *The levels of conceptual interoperability model,* in *Proceedings of the 2003 fall simulation interoperability workshop*, Vol. 7 (Citeseer, 2003) pp. 1–11.

[11] J. Euzenat, P. Shvaiko, *et al.*, *Ontology matching*, Vol. 18 (Springer, 2007).

[12] G. Wang, *Schema mapping for data transformation and integration,* Ph.D. thesis, UC San Diego (2006).

[13] V. Mascardi, V. Cordì, and P. Rosso, *A comparison of upper ontologies.* in *Woa*, Vol. 2007 (2007) pp. 55–64.

[14] E. Jiménez-Ruiz and B. C. Grau, *Logmap: Logic-based and scalable ontology matching,* in *International Semantic Web Conference* (Springer, 2011) pp. 273–288.

[15] D. Faria, B. S. Balasubramani, V. R. Shivaprabhu, I. Mott, C. Pesquita, F. M. Couto, and I. F. Cruz, *Results of aml in oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 122.

[16] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, *et al.*, *Results of the ontology alignment evaluation initiative 2016,* in *OM: Ontology Matching* (No commercial editor., 2016) pp. 73–129.

[17] M. Bayati, D. F. Gleich, A. Saberi, and Y. Wang, *Message-passing algorithms for sparse network alignment,* ACM Transactions on Knowledge Discovery from Data (TKDD) **7**, 3 (2013).

[18] J. Wang, Z. Ding, and C. Jiang, *Gaom: Genetic algorithm based ontology matching,* in *Services Computing, 2006. IEEE Asia-Pacific Conference on* (IEEE, 2006) pp. 617–620.

[19] J. Bock and J. Hettenhausen, *Discrete particle swarm optimisation for ontology alignment,* Information Sciences **192**, 152 (2012).

2

LITERATURE ANALYSIS AND REVIEW

For the few scientists who earn a Nobel Prize, the impact and relevance of their research is unquestionable. Among the rest of us, how does one quantify the cumulative impact and relevance of an individual's scientific research output?

Jorge E. Hirsch

This chapter is dedicated to reviewing and analyzing ontology alignment publications. First, a recent framework for classifying ontology alignment publications is revisited and explained in detail. Then, we delineate the ontology alignment field by analyzing a core set of research outputs from the domain. In this regard, the related publication records are extracted for the period of 2001 to 2018 by using a proper inquiry on the well-known database Scopus. This chapter details the evolution and progress of ontology alignment since its genesis by conducting two classes of analyses, semantic and structural, on the retrieved publication records from Scopus. Semantic analysis entails the overall discovery of concepts, notions, and research lines flowing underneath ontology alignment, while the structural analysis provides a meta-level overview of the field by probing into the collaboration network and citation analysis in author and country levels. In addition to these analyses, the chapter ends with discussions regarding the limitations of the field and puts forward lines for its further progress.

2.1. INTRODUCTION

This chapter is dedicated to literature review and analysis. We first look into and revisit a recent framework for classifying ontology alignment publications. We then bring forth a bibliometric approach to analyze the growth and advancement of ontology alignment. In this regard, we searched Scopus to extract research outputs regarding ontology alignment. We based the bibliometric analysis on the Scopus data, since other databases such as Web of Science (WoS) do not index the ontology matching workshop, the primary venue in this field. We retrieve and analyze around 2,975 research outputs from Scopus, including articles, conference papers, book chapters, and reviews.

Bibliometrics is a quantitative approach to study scientific activities. At its most fundamental level, bibliometrics aims to unveil the latent dynamics of scientific research and analyze its key influential factors. Content, citation, and collaboration analyses are among the commonly-practiced types of analyses within bibliometrics. In this regard, many researchers use such analyses to delineate the importance of their fields, the impact of the lead researchers, or gauge the impact of a particular research output [1]. In recent years, bibliometric analysis has drawn a lot of attention that covers a broad spectrum of application domains [2]. Some of the studies in bibliometrics have more methodological orientations and try to study the existing bibliometric measures, e.g., citation and impact factor, or to come up with new ones. For instance, Chorus et al. [3] defined a metric for self-citation and studied trends in impact factor biased self-citations of scholarly journals. In the other research, Thelwall et al. [4] made a comparison for 11 altmetrics in WoS to understand the relationship between real citations and alternative metrics in social media. In another prominent study, Ke et al. [5] made a large-scale analysis of the sleeping beauty (SB) phenomenon in science and introduced a parameter-free measure that quantifies the extent to which a specific paper can be considered as an SB.

The other line of studies in bibliometrics calibrates research activities to provide insights regarding the dynamic and the vital influential factors behind scientific research. Citation analysis [6, 7], co-authorship analysis [7–9], and co-occurrence word analysis [10] are prevalent in this application domain of bibliometrics. For instance, in the study carried out by Bromham et al. [11] on the Australian Research Council's grant proposal data, they discuss the relationship between research interdisciplinarity and the chance of winning grants, and realized that a higher degree of interdisciplinarity leads to a lower probability of being funded. Also, in another research focusing on collaboration in the field of Genomics, Petersen et al. [12] discovered interdisciplinary research draw more attention, and consequently, get more citations.

On the other side of the spectrum, bibliometrics is utilized to address much broader goals. Some studies use bibliometric analysis for answering questions that are not for the purpose of scientific activities evaluation. This is a very recent approach toward bibliometrics, which can provide an opportunity for other disciplines to benefit from the tools and techniques developed in bibliometric analysis. For instance, Candia et al. [13] studied the problem of collective memory decay using multiple datasets including American Physical Society (APS) papers and the United States Patent and Trademark Office (USPTO) patents. In the other work, Guimera et al. [14] studied the self-assembly of creative teams in the collaboration network using empirical study over a bibliometric dataset that constitutes 50 years records of recognized journals in social psychology,

ecology, economics, and astronomy. In another research, Liu et al. [15] studied the phenomenon of a hot streak for individuals career by combining over 20,000 researcher profiles from Google Scholar and WoS. Ebrahimi Fard et al. [16] also used bibliometric analysis to study the readiness of academia amid a war with the diffusion of fake-news in social media.

For the bibliometric analysis in this chapter, we carry out two classes of analyses on the retrieved articles from Scopus. The first is *semantic analysis* concerning the overall discovery of concepts, notions, and research lines flowing underneath the scientific disciplines. For this analysis, two types of analysis are used, *topic* analysis and *thematic* analysis. We first use latent Dirichlet allocation (LDA) [17] to model the topics underlying the ontology alignment bibliometric data. To do so, the title, abstract, and keywords of each document were subjected to LDA, and six topics were extracted accordingly. Although the topics are extracted based merely on the words and their frequency in each document, the extracted topics are interestingly meaningful and delineate different applications to which ontology alignment can be applied or the problems it can address. Another analysis in this category is *thematic*, in which we show the shares of ontology alignment to top-cited articles and top percentile journals, as well as the fundamental disciplines contributing to ontology alignment. In addition to semantic analysis, we perform *structural analysis* to obtain a meta-level overview of the field. We break the structural analysis into two categories. First, we analyze the collaborations between different authors and countries in ontology alignment based on their co-authorship (collaboration analysis), and then gauge the impact of researchers and countries by analyzing their number of published articles and their number of citations, as well as visualizing their citation networks (impact analysis). The analyses of bibliometric data help us address some current issues in the field and also provide some solutions for its further improvement.

The remainder of this chapter is structured as follows. We first review a classification of ontology alignment papers in Section 2.2. Section 2.3 is dedicated to the methodology by which we retrieve ontology alignment research outputs for bibliometric analyses, as well as the tools that are used to analyze them. Semantic analysis is covered in Sections 2.4 and 2.5, where the topic analysis is presented in the former and the thematic analysis is discussed in the latter. Section 2.6 is devoted to the collaboration analysis, and the impact analysis at the author and country levels is explained in Section 2.7. We conclude the chapter and discuss the lessons learned from the analyses in Section 2.8.

2.2. STATE-OF-THE-ART PROGRESS IN ONTOLOGY ALIGNMENT

In this section, we provide the state-of-the-art progress of ontology alignment and classify the contributions to this domain. Figure 2.1 displays six different types of contributions to ontology alignment, that is a revisited version presented in [18]. In the following, each of these classes is described in more detail.

2.2.1. REVIEW ARTICLES

This class of articles are the ones devoted to reviewing the field of ontology alignment, as well as those that detail the state-of-the-art contributions and future challenges. The

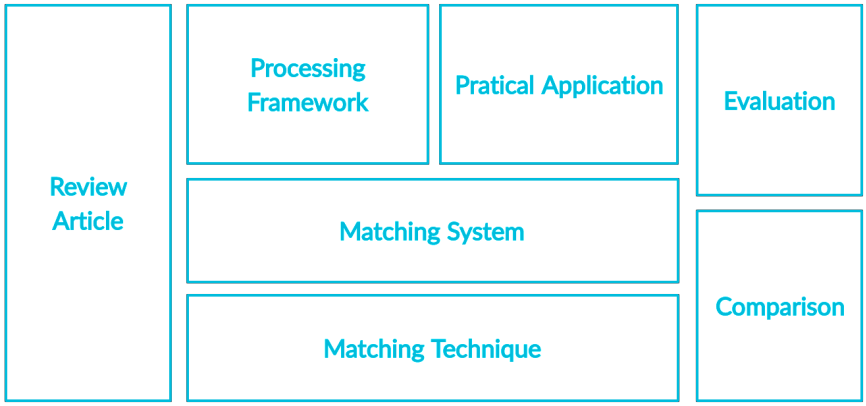


Figure 2.1: Ontology alignment article classification, a revisited version of that presented in [18].

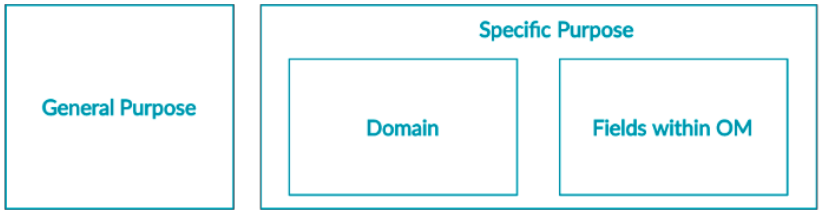


Figure 2.2: The types of review paper [18].

articles within this class can also be categorized as *general purpose* and *specific purpose*, where the former includes the publications that offer insights into the ontology alignment field in general without any concentration on a sub-problem or an application, while the latter is a more focused review on a specific problem in ontology alignment or an application that ontology alignment can be applied to. Figure 2.2 shows the types of ontology alignment review articles.

There are several articles related to the general purpose review articles, among them are surveys [19–21], state-of-the-art articles [22, 23], and publications addressing the future challenges of the field [24, 25].

The articles related to the specific purpose category contains the research that concentrates on a specific problems in ontology alignment. Examples of such articles are large-scale ontology alignment [26], complex ontology alignment [27], instance matching [28], and matching across different languages [29]. These articles fall in the *fields within ontology matching* subcategory of the *specific purpose* category. Besides, there are several other articles that survey applying ontology alignment to particular domains such as medicine [30], geography [31], and agriculture [32]. These articles are classified as *domain* specific purpose.

Aside from the type, the review papers can be classified as *quantitative* and *qualitative* based on the analysis approach they employ. Based on the papers retrieved from Scopus, some of which are also mentioned above, all the ontology alignment review pa-

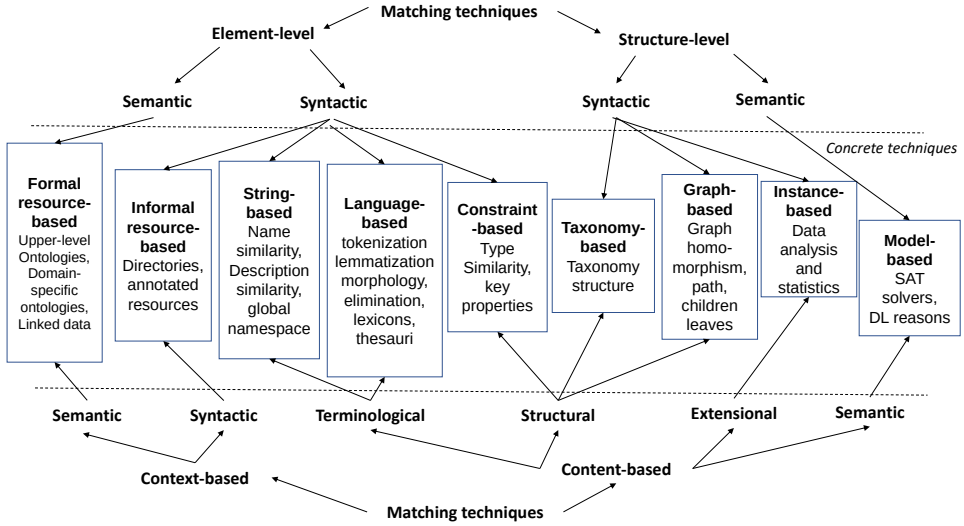


Figure 2.3: The classification of simple matching techniques [33].

pers use the qualitative approach and are based on the opinions of the researchers conducting the literature review. To the best of our knowledge, however, no paper considers a quantitative approach like bibliometric analysis to delineate the ontology alignment field.

2.2.2. MATCHING TECHNIQUE

Matching techniques are the building blocks of matching systems to align the ontologies in question. In principle, there are two types of techniques, simple and complex, the former of which refers to the simple similarity metrics of two concepts of ontologies in question, while the latter considers the concepts of ontologies from a higher point of view.

Simple matching techniques have been well-studied in the literature and a considerable amount of techniques are currently available. A thorough classifications of these techniques, encompassing almost all other classifications, are presented in Figure 2.3, adopted from [33]. This classification from the top shows the interpretation of different techniques with respect to the input information, while the classification from the bottom takes into account the type of input being used by matching techniques.

The first tier from the top-down classifications includes two subcategories of element-level and structure-level. The element-level techniques involve the metrics that consider the concepts in isolation, and ignore their positions in their corresponding ontology, while the structure-level techniques analyze the entities based on their positions in their ontologies to obtain correspondences. Following the top-down interpretation, the next level classifies the matching techniques into syntactic and semantic measures, the former of which limits their calculation to the ontologies and entities in questions, while the latter uses a formal semantic resource such as upper ontologies.

From the bottom of this classification, the matching techniques are divided into content-based and context-based classes. The content-based techniques base their computations on the internal information from the ontologies in question, while the context-based techniques use external resources for similarity computation. Context-based techniques can be classified into semantic and syntactic groups that are discussed above, and content-based techniques are further divided into the four categories listed in the following:

- **Terminological:** The methods that consider their inputs as strings.
- **Structural:** These methods base their computation on the positions of entities in the associated ontologies.
- **Extensional:** The methods that discover correspondences based on the available instances.
- **Semantic:** The methods that need some semantic interpretations of input that is often based on a reasoner.

The next tier in both top-down and bottom-up classifications involves the matching techniques that are elaborated in the following:

- **Formal resource-based** methods involves the techniques that use formal resources to support and possibly enhance the alignment of two ontologies in question. Examples of resources are upper level ontologies, domain-specific ontologies, and stored alignments of ontologies being matched previously.
- **Informal resource-based** techniques are similar to the formal resource-based methods, but they use informal resources like directories and annotated resources to support the matching.
- **String-based** methods are arguably the most important similarity metrics that base their similarity computation on the names of the concepts in question. There are several string similarity metrics, including but not limited to, Jaccard, Levenstein, Jaro, Jaro-Winkler, and TF-IDF [34].
- **Language-based** techniques take the semantics of the strings into account by using some external resources such as WordNet [35]. The pre-processing strategies such as tokenization, lemmatization, or stop-word removal lie within this category.
- **Constraint-based** techniques consider the internal structure of entities, such as the domain and range of data properties, to calculate the similarity.
- **Graph-based** methods consider the ontology alignment problem as graph homomorphism problem, where ontologies are regarded as labelled graphs.
- **Taxonomy-based** methods are not commonly applied to ontology alignment. This class can be seen as a particular case of graph-based methods, where only the specialization relation is considered.

- **Instance-based** methods use the instances of concepts in the ontology to compute the similarity of the concepts. The underlying idea is that if two concepts share similar instances, then the concepts are similar as well.
- **Model-based** methods take advantage of semantic interpretation related to the input ontologies. The matching techniques based on description logic reasoning techniques are an example of this class.

Another category of matching techniques is complex matching techniques that use simple matching techniques and devise new complex methods. Examples of such techniques are soft Jaccard and soft TF-IDF methods, where the simple techniques are used and extended [36].

2.2.3. MATCHING SYSTEMS

Matching systems take advantages of (possibly) several matching techniques and strategies in order to align the ontologies in question. Basically, there are four types of matching systems:

- **Schema-based systems** match the given ontologies based on the schema-level input information.
- **Instance-based systems** consider the instances of concepts for alignment. The underlying idea is that two concepts are similar given they share similar instances.
- **Mixed approach** uses both schema and instance information for alignment.
- **Metamatching systems** concern with combining different alignment systems and tune several hyperparameters such as threshold and weights for alignment systems.

The focus of this dissertation is on schema-based ontology alignment, which can be modeled using a zero-one non-convex programming [37]. Solving such optimization problems are very difficult in polynomial time, and obtaining an approximation solution often requires massive time and memory. One viable way for solving such problems is evolutionary algorithms (EAs) that has been already practiced in the ontology alignment literature.

There are two different ways to apply EAs to the ontology alignment problem. The first approach is the so-called *meta-matching*, whose goal is to find heuristically the hyper-parameters of an alignment system. Generally, a set of similarity measures for each pair of entities are selected, and the goal is to achieve the optimal weights for the chosen similarity metrics. Another critical parameter usually computed by meta-matching techniques is the threshold according to which the final alignment will be obtained. The major shortcoming of the meta-matching is that they often need a reference alignment, or a part of it, in order to identify the hyper-parameters. In reality, however, the reference alignment of ontologies in question is often unavailable, and the applicability of such systems is thus restricted. Such a drawback is present in most of the meta-matching systems using EAs [38–40]. To our knowledge, there is only one meta-matching

system which is able to discover alignments of two given ontologies needless of having the reference alignment [41]. In their proposed system, Xue et al. have used two heuristic measures which are not reliant on a reference alignment. The measures are *MatchFmeasure* and *Unanimous Improvement Ratio (UIR)* based on which a memetic algorithm is applied to identify the alignment.

The second way of using EAs is to solve the ontology alignment problem directly. Similar to the meta-matching techniques, there are multiple systems which require a reference alignment. These systems optimize various objective functions such as F-measure [42, 43] and a weighted sum of similarity metrics [44]. Such systems also have narrow applicability in the real world situations since no gold standard is available in reality. In addition, there are several EA-based systems suitable for real-world situations. Wang et al. are arguably the first ones who used an evolutionary algorithm, i.e., genetic, to find an alignment between two given ontologies [45]. Their proposed system, GAOM[1], models a possible alignment as a population member (chromosome). They further define the intension of a concept as a set containing its name, properties, and instances, and the extension of a concept as its relations (i.e., object property) to some other entities at the same ontology. Based on the intensional and extensional features, the fitness of a chromosome is computed, and the optimal alignment is discovered using the genetic algorithm. GAOM suffers from several drawbacks. First and foremost, it solely matches the classes, not the object or data properties, although they are used to measure the similarity of classes. On top of that, it is not clear how the structural similarity of concepts is considered.

A well-developed system, called MapPSO [46], identifies the alignment based on the discrete particle swarm optimization (PSO). MapPSO is able to align classes and properties of two given ontologies without a reference alignment. This system utilizes lexical, linguistic, and structural similarity metrics to determine the fitness of a particle. Aside from its salient characteristics, MapPSO has several severe drawbacks as well. First, there is no pre-processing, e.g., tokenization and stemming, over the names of various entities. In their alignment algorithm, the Levenstein string similarity metric [47] is directly applied to the names of two entities to gauge their similarity. This approach has low applicability to real-world ontologies, since the concepts are likely to be named as the combination of various tokens. As a result, the similarity computation of names merely based on the Levenstein metric would lead to overall poor performance, since recent studies have accented the role of string similarity metrics for ontology matching [34]. Such names cannot be discovered in WordNet [35] neither so that the linguistic similarity used in MapPSO would not lead to a significant mapping discovery when linguistic heterogeneity is present. Yet another subtle but essential drawback of MapPSO is that the same string similarity metric has been used for matching properties. Nonetheless, the sole consideration of the names of properties would increase simultaneously the false negative and false positive [48].

There are also several pitfalls inherited from PSO. First of all, it is a population-based evolutionary algorithm, and MapPSO used it in a way that it needs to generate a significant number of particles in order to transition to the next generation and to find the optimum of the given problem. Such populations need to be stored in the main mem-

[1] Stands for **G**enetic **A**lgorithm-based **O**ntology **M**atching

ory so that it requires a considerable amount of space. The computation of populations fitness would also be time-consuming. Further, PSO is suffering from the so-called premature convergence, which makes it converge to the local optima.

2.2.4. PROCESSING FRAMEWORK

A class of articles in this category involves the research that *processes and exploits the ontology alignment* for some purposes such as ontology merging [49, 50], ontology evolution [51], reasoning based on the discovered correspondences as rules [52], and ontology argumentation [53]. The other class of articles in this category concerns with *alignment framework and format* [54], where the process does not finish by identifying the alignment, and further actions such as alignment validation are available for the user to perform.

2.2.5. PRACTICAL APPLICATIONS

This category includes the articles that use ontology alignment to address a specific problem. Among these applications are semantic web service discovery [55, 56], P2P systems [57], and multi-agent systems [58–60].

2.2.6. EVALUATION

This category includes the articles that study the evaluation of ontology alignment systems. They are divided in three categories: (i) Performance metrics; (ii) Benchmarks; (iii) Evaluation method;.

PERFORMANCE METRICS

Performance metric articles involve the papers that put forward new metrics. In general, there are several common performance metrics that are frequently used. The three widely-used performance metrics for ontology alignments are precision, recall, and F-measure. Given an alignment A and the reference A^*, precision is the ratio of true positives to the total correspondences in the alignment generated by a system, and is defined as:

$$Pr(A, A^*) = \frac{|A \cap A^*|}{|A|}, \tag{2.1}$$

where Pr is the precision and $|.|$ is the cardinality operator.

Recall is another popular measure, which is computed as the ratio of the true positives to the total number of correspondences in the reference. Thus, it can be computed as:

$$Re(A, A^*) = \frac{|A \cap A^*|}{|A^*|}, \tag{2.2}$$

where Re is the recall.

Each of precision and recall represents only one aspect of the alignment systems; the former only considers the correctness of the alignment, while the latter accentuates the completeness of the alignment with respect to a reference. As a combination of both,

F-measure, as the harmonic mean of precision and recall, is often used, i.e.,

$$\text{F-measure}(A, A^*) = 2\frac{Pr(A, A^*) \times Re(A, A^*)}{Pr(A, A^*) + Re(A, A^*)}.$$

Aside from these popular performance metrics, there are two important principles for a given alignment. The first is *conservativity* [61, 62], which states that, with regard to the alignment being generated, the system must not impose any new semantic relationship between the concepts of the ontologies involved. The second is *consistency*, which states that the discovered correspondences should not lead to unsatisfiable classes in the merged ontology [62].

There is also a metric called *Recall+*, which indicates the portion of correspondences that a system cannot readily detect. When this performance metric has a higher score, that indicates that the associated system is able to identify the most non-trivial, i.e., non-syntactically identical, correspondences between two given ontologies. In addition, execution time is another important indicator of the performance of the alignment systems, that also has to be taken into account.

Aside from these metrics, contributions in performance metrics involve developing new metrics such as relaxed precision and recall [63], semantic precision and recall [64], scores related to credibility and stability [65], as well as metrics for evaluating interactive ontology alignment [66].

The performance metrics for evaluation can be distinguished based on their need for a reference alignment. While most of the performance metrics are computed based on a given reference alignment, several metrics such as inconsistency [62] evaluate an alignment without having the associated reference. The mapping repair methods [67–69], which amend the alignment problems like inconsistency, fall within this category as well.

BENCHMARK

The second category is *benchmark* that concerns with the dataset or the benchmark that is used to evaluate an alignment system. The benchmarks used for evaluation are typically the benchmarks provided by the OAEI organizers. In the following, the benchmarks that will be used in the next chapters are itemized:

- **Benchmark:** As one the few synthesized tracks, the benchmark track includes generating two ontologies based on a *seed* in such a way that the correct alignment between the two generated ontologies was known. It used to be one of the most essential tracks at the OAEI, but it is no longer a part of it since 2016.
- **Anatomy:** As one of the eldest OAEI tracks, anatomy track consists of matching the adult mouse anatomy to a part of NCI thesaurus describing the human anatomy.
- **Conference:** This track involves matching and aligning seven ontologies from different conferences. For this track, there are two different reference alignments, i.e., certain and uncertain.
- **Disease and Phenotype:** The OAEI disease and phenotype track comprises matching different disease and phenotype ontologies that consists of two tasks. The first

one is to align the human phenotype (HP) ontology to the mammalian phenotype (MP), and the second is to align the human disease ontology (DOID) and the orphanet and rare diseases ontology (ORDO).

- **Large Biomedical Track:** The aim of this track is to find alignments between the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI) ontologies. The ontologies are large and contain tens of thousands of classes.

- **SPIMBENCH:** This track aims to determine when two OWL instances describe the same Creative Work. There are two datasets, *Sandbox* and *Mainbox*, each of which has a *Tbox* as the source ontology and *Abox* as the target. Tbox contains the ontology and instances, and it has to be aligned to Abox, which only contains instances. The difference between Sandbox and Mainbox is that the reference of the former is available to the participants in advance, while the latter is a blind matching task and participants do not know the real alignment in advance.

EVALUATION METHODS

The techniques and strategies such as statistical inference [70] that are used for evaluation lie within this category. Similar to performance metrics, evaluation methods can also be divided into the methods that need the reference alignment and those that do not. The example of the former includes several performance metrics such as semantic precision and recall [64], while the example of the latter is the sample evaluation of alignment systems [70].

2.2.7. COMPARISON

The articles for comparison can be divided into two groups. The first group includes the articles that compare matching systems or matching techniques for a particular domain/problem. For instance, there is a solid comparison of string similarity metrics for different OAEI tracks [34], as well as a comparison for property matching [48]. The second group includes the methodologies for comparison. Although different methods have been developed for evaluating ontology alignment systems, there is a lack of a solid methodology for comparing alignment systems or selecting the most appropriate system for a particular matching task.

2.3. RESEARCH METHODOLOGY FOR BIBLIOMETRIC ANALYSIS

In this section, we first discuss the research strategy that is employed to extract ontology alignment research outputs from two well-known databases, WoS and Scopus. We then explain the tools and methods that are used for the analysis of extracted bibliometric data in further sections. Figure 2.4 displays the research methodology being used. The top panel of Figure 2.4 plots the search methodology in Scopus, the bottom-left panel shows the refinement of the retrieved outputs from Scopus, and bottom-right panel illustrates the types of analyses being conducted and their corresponding sections within this chapter. In the following, each of these steps are discussed in more detail.

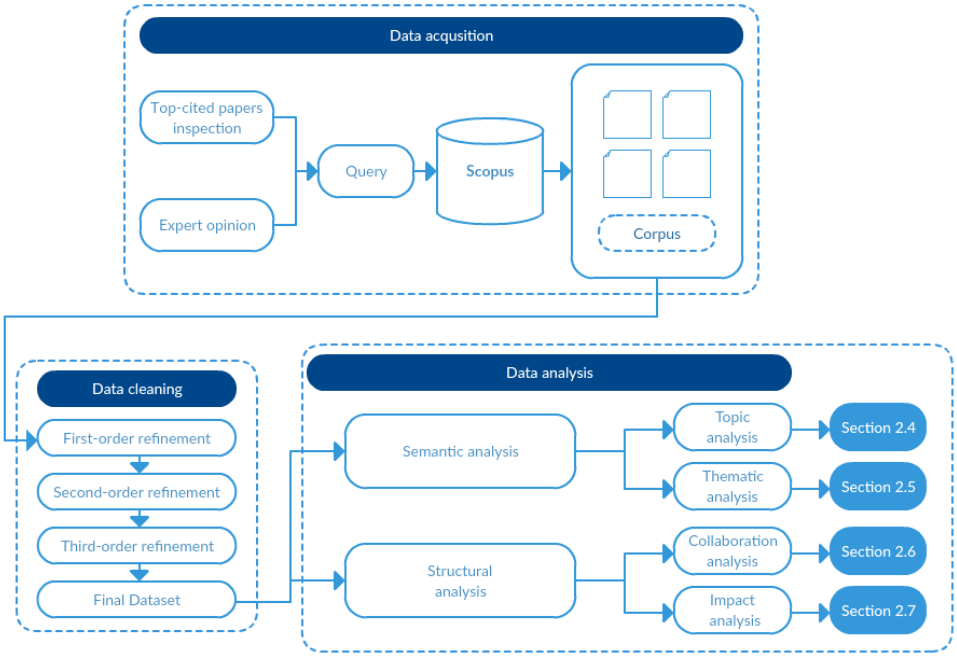


Figure 2.4: The flowchart of the research methodology being used as well as the analyses conducted in this chapter.

2.3.1. ONTOLOGY ALIGNMENT BIBLIOMETRIC SEARCH APPROACH

Bibliometric approaches aim at the quantitative analysis of research outputs, such as publications and patents, in order to comprehend and track the scale, direction, and the innovation of a field. The major prerequisite for such an analysis is to find the relative research outputs according to which the analysis could be performed. In this regard, there are several well-known databases such as Scopus and Thomson Reuters Web of Science, from which research outputs can be retrieved by proper queries.

For the bibliometric analysis, there are several standard ways to extract the pertinent research outputs to a problem/domain. Index-based methods [71] use the categories already defined by the publication database and retrieve research outputs accordingly. The approach is simple, but the search is restricted to the indices created by journals. In addition, we observed that there is no particular index for ontology matching in several publishers such as IEEE and ACM. Another approach is based on citation and co-citation [72], wherein one first needs to find a core corpus of research outputs that everyone agrees upon. The basic corpus of publications then evolves by using its citations and co-citations. The major drawback of this technique is that it is difficult to replicate, and there is no consensus on the interpretation of citations and co-citations. For ontology alignment, in particular, finding a core amount of publications which everyone agrees upon is not easy to acquire. One potential way would be to use the papers published in the ontology matching workshop, but the number of articles in the workshop is quite restricted so that the final corpus would not include the exhaustive set of

all publications for this problem. Another way to get the bibliometric data is to detect a set of journals dedicated to a domain and analyze their published articles [73]. For ontology alignment, unfortunately, there is no particular journal to conduct the analysis. On top of that, ontology alignment is interdisciplinary by nature, since it is used as a pre-processing strategy in many circumstances and has thus diverse applications. As a result, research outputs are not restricted to a specific journal or domain.

One of the most popular yet straightforward methods is to use several expert-defined keywords based on which research outputs are retrieved [74, 75]. This method is semi-automatic, since the results of the search are then reviewed by an expert to exclude the irrelevant items for the subsequent analyses. After inspecting the keywords of top 50 cited research outputs in this domain and further discussion with the experts in this domain, we arrived at three main keywords: “ontology alignment", “ontology matching", and “ontology mapping". The keyword "ontology" alone refers to a more general concept in the Semantic Web and adds research outputs that are irrelevant to ontology alignment. Thus, we need to conduct the search based on the keyword “ontology alignment", which is interchangeably referred to as “ontology matching" or “ontology mapping" as well. Thus, these terms should be considered for searching the databases. We further realize that the ontology alignment evaluation initiative (OAEI) is also essential, since it might also add some research outputs. Since the papers that contain “ontology alignment evaluation initiative" are completely covered by articles retrieved solely by the keyword “ontology alignment", this term is redundant. However, “OAEI" must be used as another keyword. The search based on these keywords is reasonable, since it delivers a considerable number of articles in ontology alignment. Thus, the keywords being used to retrieve ontology alignment research outputs are:

- *Ontology alignment*;
- *Ontology matching*;
- *Ontology mapping*;
- *OAEI*.

We use the keywords listed above to retrieve research outputs related to ontology alignment. Then, the identified research outputs need to be processed by an expert to verify if they are relevant to the ontologoy alignment domain. To this end, we conducted an inquiry in WoS by searching the identified keywords in the title, abstract, and keywords of research outputs. The result of the search included 1,536 articles spanning from 1999 to 2018. The 1,536 articles were processed in three different phases to omit the articles that are not pertinent to the domain. In each of these phases, each paper is flagged as either one of the following options:

- **Relevant** articles that are related to the ontology alignment domain;
- **Irrelevant** articles that do not contribute to ontology alignment;
- **Uncertain** articles that could not be flagged in the associated phase, so that they are passed to the next phase.

Table 2.1: Four steps for filtering the ontology alignment research outputs. *Total* is the number of items at the beginning of each step, *Relevant* and *Irrelevant* denote the number of items that are flagged as related and unrelated to ontology alignment, respectively, and *Uncertain* is the number of items that could not be flagged as either relevant or irrelevant in the corresponding phase so that they are passed to the next phase. The search query for retrieving data from Scopus is: *TITLE-ABS-KEY ("ontology Alignment" OR "ontology matching" OR "ontology mapping" OR "OAEI").*

	Total	Relevant	Irrelevant	Uncertain	Description
1	-	-	-	3289	Retrieving bibliometric data from Scopus
2	3289	1820	225	1244	Inspecting the publication items by revising the title only
3	1244	1094	53	97	Inspecting the publication items by revising the abstract only
4	97	61	36	0	Inspecting the whole paper
sum	-	2975	314	-	

In the first phase, the title of papers was considered, since most of the related works to ontology alignment could be easily detected by merely their titles. In this phase, 1,166 items were identified as relevant or irrelevant, and 370 items were passed to the second phase. In the second phase, the abstract of the remaining papers was considered, according to which 316 articles were recognized as relevant, and the remaining 54 papers were passed to the third phase. In the final stage, the 54 articles were thoroughly inspected and the papers were classified as relevant and irrelevant. In total, 1,420 research outputs were labeled as relevant to ontology alignment and the remaining of articles were eliminated.

After rigorous examinations of the remaining articles, we realized that research outputs regarding the ontology matching workshop are not indexed by WoS. Since this workshop is the essential venue of this domain, we refused to continue the analysis based on WoS data. Therefore, we conducted the same search strategy in Scopus and realized that the items recovered by this database include the articles from the ontology matching workshop as well. The inquiry in Scopus retrieved 3,289 articles from 2001 up until 2018. Although it does not index several papers from the late 90s and early 2000 e.g., [76, 77], it includes all the papers from the ontology matching workshop. Since the number of articles that are not indexed by Scopus is not significant, especially compared to WoS, we use Scopus data for further analyses. The retrieved papers from Scopus underwent the same procedure as that of WoS articles in order to discard the irrelevant papers. After conducting the three phases of processing research outputs, 2,975 articles are labeled as relevant to ontology alignment. Table 2.1 tabulates the steps for obtaining and cleaning the bibliometric data from Scopus. The analyses discussed in next sections are performed on the remaining items.

2.3.2. TOOLS AND METHODS FOR BIBLIOMETRIC ANALYSIS

In this section, we explain the methods and tools that are used to analyze the 2,975 related research outputs. For topic modeling, we use latent Dirichlet allocation (LDA) [17], a statistical method that aims to find the underlying topics in ontology alignment based on the articles published in this domain in the period 2001-2018. We use the LDA imple-

Table 2.2: The tools used for the analyses conducted in this chapter.

Type of analysis		Tools			
		LDA	SciVal	VOSViewer	Gephi
Semantic	Topic Analysis	✓			
	Thematic Analysis		✓	✓	
Structural	Collaboration Analysis		✓	✓	✓
	Impact Analysis		✓	✓	✓

mentation in MATLAB to analyze the articles. For the thematic analysis, we discuss the number of all publications as well as publications in top journals along with the disciplines that contribute to ontology alignment.

In addition, the collaborations between authors and countries worldwide are analyzed and the level of international and academic-corporate collaborations over the last few years are discussed. We then probe into the contributions and impacts of authors and countries in ontology alignment. For these analyses, we use VOSviewer [78], SciVal[2], as well as Gephi for network visualization [79]. Some of the analyses presented in this section are limited to the six most recent years due to the fact that more bibliometric metadata are only available in recent years. Table 2.2 shows the tools used in different anlyases in this chapter.

2.4. TOPIC ANALYSIS OF ONTOLOGY ALIGNMENT

Topic modeling is a statistical approach to discover the underlying topics of a set of documents based on the frequency of words that appeared in the documents. Topic modeling is generally used to find the underlying hidden semantic structure of a text body. One of the most well-known algorithms for modeling the topic is latent Dirichlet allocation (LDA) [17], which can find the hidden topics of a set of texts with a given number of topics.

In this analysis, we aim at analyzing different topics in order to provide a broader picture of the domain, as well as the problems to which ontology alignment is an essential contribution. In this regard, the title, abstract, and keywords of the retrieved research outputs were subjected to LDA, and the identified topics were visualized using word clouds. It is typically necessary to conduct several pre-processing procedures on the given data in order for the final topics to be more meaningful. The following preprocessing strategies are used before applying LDA:

- **Tokenization:** Tokenization means that sentences are broken into their constituent words. Tokenization transforms a document into a bag of words, which are useful for further topic analysis.

[2] https://www.scival.com

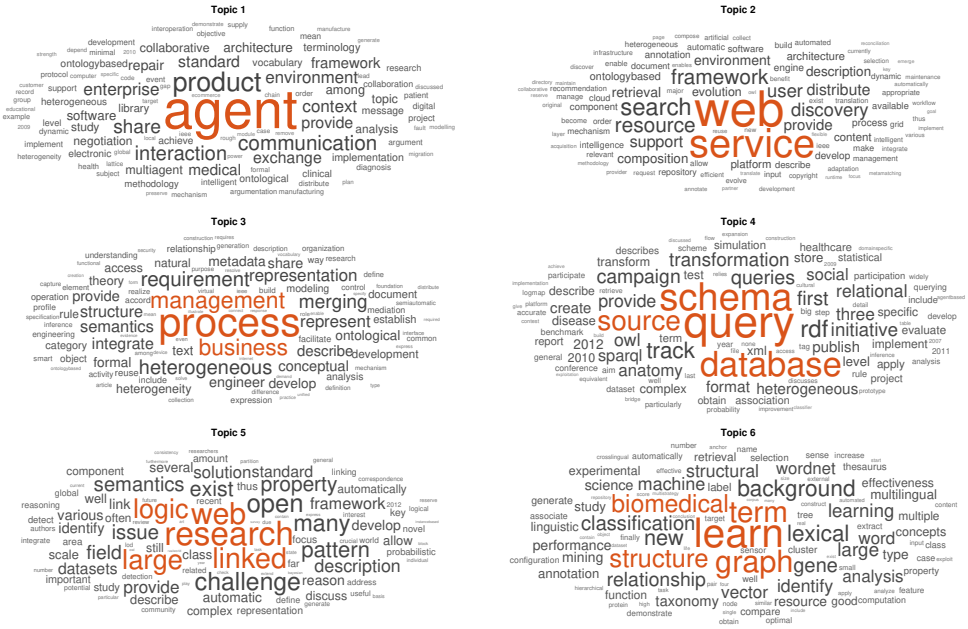


Figure 2.5: Six topics of ontology alignment based on the bibliometric data 2001-2018. The topics are identified by LDA [17] and visualized in MATLAB.

- **Meaningless Word Removal:** For meaningful analysis, stop words and the words whose length is less than three were removed, since they do not have any further impact on the analysis. At the same time, we removed the terms used for the query, since they exist in all the retrieved documents, while they do not convey any useful information for ontology alignment topics.

- **Lemmatization:** Words were also lemmatized, which means that the verbs in the past or future tense are changed into their present tense, and other third person words are replaced by their corresponding first person. We further used the Porter stemming algorithm [80], which replaces the words with their associated roots.

The processed documents were then subjected to LDA for topic modeling. LDA also needs to have the number of topics identified in advance. We tried a variety of numbers between four up until 14, and we found out that more meaningful topics are detected by six topics. Figure 2.5 displays the topics identified by LDA. In this figure, the size of each term indicates its importance, and the most vital terms are also identified by the orange color. Note that the order of topics is arbitrary, hence, the topics that appeared at the beginning do not have any privilege over those presented later on. In the following, we analyze the topics discovered by LDA.

- **Agent-based Modeling:** Agent-based modeling is used to simulate the actions and interactions of independent agents in order to estimate their impacts on the overall system. Agents communicate together by exchanging messages structured in

different syntaxes such as Agent Communication Language developed in Foundation for Intelligent Physical Agents (FIPA-ACL) and Knowledge Query and Manipulation Language (KQML). These syntaxes determine only the overall structure of the messages and not their contents. The actual content of a message expressed by an agent is typically modeled by an ontology. As a result, when two independent agents communicate with each other, it is unlikely that they can understand each other if they do not use the same ontology for communication. In this regard, ontology alignment has been extensively used by the agents to understand various messages in different formats. One of the first problems to which ontology alignment was applied is agent communication, where the paper was published in 2002 [58]. There are also several systems that employ agent-based modeling for automatic ontology alignment, where they call such systems as *agent-based ontology alignment* [59, 60, 81]. In this view, the mappings between two ontologies are deemed as a product of communications between two intelligent agents [82]. Topic 1 in Figure 2.5 illustrates the topic related to articles of ontology alignment that used the notions of agent-based modeling. In this topic, as expected, the term “agent" is identified as the most central word. There are also several other terms related to agent-based modeling. For instance, “communication" and “interaction" as are usually paired with “agent", and the terms “collaborative", “negotiation", and “exchange" that refer to collaborations, negotiation, and data exchange between agents, are the features that agents can be equipped with by using ontology alignment. There are also some general terms of agent-based modeling such as “environment", “software" (as in software agent), “multiagent", that are visible in Topic 1 of Figure 2.5.

- **Web Service Discovery:** Web Services are the services provided by some providers exposed their particular services to a broad audience by using Web technologies. Semantic Web Services (SWS) are conceptual specification of web services to describe the services more richly so that their discovery by requesters become even easier. Web Service discovery is the process of finding a service that meets a goal. Sometimes a request cannot be responded merely by a single service, but by a composition of services. In this situation, it is required to have a composition process, which integrates several services in order to meet a particular need of a requester. SWS can be modeled by different standards such as Web Services Description Language (WSDL) [83] and OWL-S [84], and different terminologies are used by different providers/requesters. As a result, the discovery of services requires the use of ontology alignment techniques so that the heterogeneity between different services is reconciled and the discovery success increases significantly. SWS discovery was also one of the first applications that ontology alignment could address. The first paper employing ontology alignment for SWS discovery dates back to 2003 [55], and it has been since used in other studies [56, 85, 86].

  Topic 2 in Figure 2.5 is devoted to this important application of the ontology alignment. As it is readily observable, the terms “web" and “services" are detected as the main terms, and there are also other terms such as “discovery" and “composition" that are the general terms in the Semantic Web Service domain.

- **Process Model Matching:** Process models comprise a set of related activities or tasks which need to be done in a specific structure (sequential and/or parallel) to produce a service or product. The matching of these processes is of the essence for several tasks such as system validation and process harmonization [87–89]. In this regard, ontology alignment systems or techniques can be used. A more specific application is matching the business processes [90–93], where the process models are typically related to e-commerce [94].

  In the OAEI 2016 and 2017, there was a track that included matching different process models of the university admission systems. As a result, the problem is completely well-known by the ontology alignment community as well. As Topic 3 in Figure 2.5 shows, LDA has been able to detect the importance of this problem for ontology alignment. The terms “process" and “business" are at the heart of this topic, which accentuate the importance of process model matching in ontology alignment. The term “management" also has a significant weight. Interestingly, “business process management" refers to a domain where matching of processes has become a major research area [95].

- **Query Answering:** Information provided by different sources is not described by a unified schema on the Web. At the same time, users do not utilize the same terminology in their search queries. Thus, a semantic query answering is required to rewrite the query in order to provide sensible results. Since both the information on the web and the queries are the reasons for the discrepancy, ontology alignment can be helpful to address this challenge and improve the relevance of the retrieved information. Thus, ontology alignment has been used extensively in this regard. As Topic 4 in Figure 2.5 illustrates, this problem is quite important in ontology alignment. The term “query" is the most accented term, and there are some other related terms such as “relational", “database", “schema". In the OAEI 2014 and 2015, there was a track for answering queries by the aid of ontology alignment systems, which indicates that this problem is also well-known to the community.

- **Linked Data and Logic:** One of the primary objectives of the Semantic Web is to link different data sources on the Web to other available resources so that useful information can be provided from the available data. Since the published data on the Web is designed by many, interlinking of these data is not straightforward due to their heterogeneous nature. As a result, ontology alignment is a potential solution to fulfill this essential objective of the Semantic Web [96, 97]. This is the reason that “linked" in Topic 5 of Figure 2.5 has been centralized. Another vital term in this topic is “logic". Logic has been widely used to align two different ontologies [98–101]. One of the well-known systems is LogMap [101], which is based on logic and is one of the top systems at the OAEI in the recent decade. Also, logic has been used to repair the alignment automatically obtained from alignment systems [67–69].

- **Machine Learning and Biomedical Ontology:** Topic 6 is a mixture of a well-known approach for ontology alignment and one essential domain to which ontology alignment has been applied. There are several ontology alignment systems which

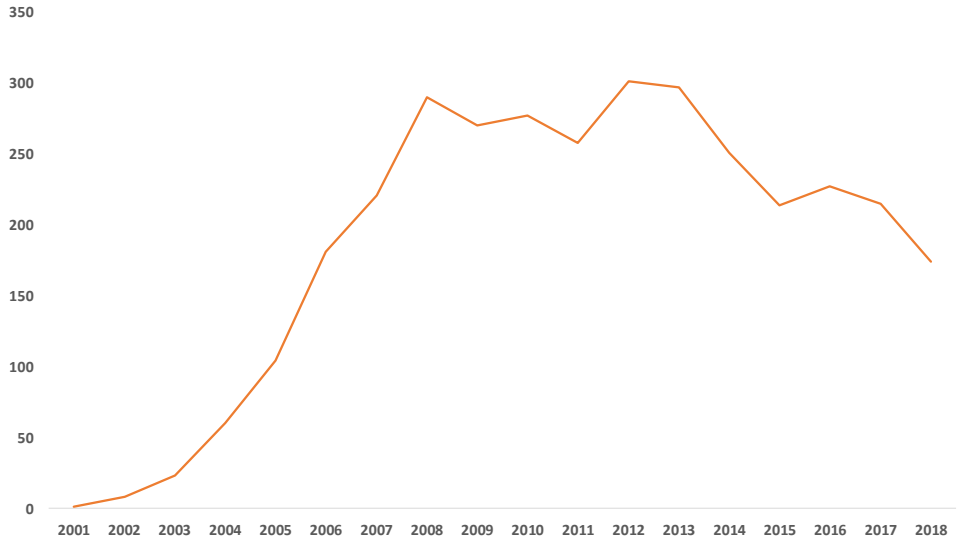


Figure 2.6: The number of documents published about ontology alignment on Scopus between 2001 to 2018.

use machine learning techniques for alignment. In fact, machine learning techniques are one of the first approaches that are used for aligning ontologies [102, 103]. There are also many machine learning-based systems that require to have a gold standard for training [104–106]. These systems are sometimes called pre-trained systems [33] and need to have [a part of] the reference alignment for training. The system is then ready to map the remaining of the same ontologies or other ontologies in the same domain. The terms “learn", “machine", “learning", and “classification" in Topic 6 are the indicators of these alignment systems.

Another term in this topic is “biomedical", which is one of the most important domains to which ontology alignment has been applied. The anatomy track, which involves matching the adult mouse anatomy and a part of NCI thesaurus comprising the human anatomy, is one of the first tracks in the OAEI [107]. There are several other tracks such as disease and phenotype [108] and large biomedical [109–111] tracks which have been recently added to the OAEI. Therefore, there is no surprise to see this term as an important topic of ontology alignment. The terms “large" and “background" are also related to this theme, since there is one large biomedical track in the OAEI and it is the common practice to use background knowledge such as UMLS [112] for matching ontologies in the biomedical domain.

2.5. THEMATIC ANALYSIS

In this section, the thematic analysis of ontology alignment publications is presented based on the collected bibliometric data between 2001 and 2018. We first study the number and types of research outputs, followed by the contributions of ontology alignment

publications to the top-cited and top journal percentiles. Afterward, the disciplines contributed to ontology alignment are discussed.

2.5.1. NUMBER AND TYPES OF PUBLISHED DOCUMENTS

In this subsection, we discuss the number and the types of research outputs in the ontology matching data from 2001 to 2018. The essence of having the automatic mapping between two ontologies was discussed in the late 90s and early 2000 for different problems such as ontology merging [77, 113] and further in business-to-business (B2B) electronic commerce [57], where mappings between ontologies, taxonomies, and the classification system were required. In the preceding years, the existence and importance of an automatic mapping were discussed in several other problems such as agent communication in multi-agent systems [58]. Ever since, ontology alignment has been the topics of numerous research studies, by which various problems have been addressed. Figure 2.6 shows the number of research outputs from 2001 to 2018. According to this figure, the number of outputs has been steadily increased until 2008, when around 290 research articles are published. From 2008 up until 2013, the number of publications has been approximately the same, where the maximum number of outputs is in 2013 with 300 publications. From 2013, the number of documents has experienced a steady decrease, where its minimum number reached in 2018 with 175 research outputs. Interestingly, in 2013, Shvaiko and Euzenat [25] showed the improvement of the field based on their analysis on the state-of-the-art ontology matching systems and the results of evaluations, while they observed that the speed of the ontology alignment progress was slowing down. The slow progress in the field has shown itself in the number of publications in the field as one important criterion.

We also analyze the types of research outputs in the ontology alignment field. Figure 2.7 displays the percentage of different types of papers published in the ontology matching domain between 2001 up until and including 2018. According to this figure, the vast majority of research outputs, i.e., around 65%, are published in conference venues. It is no surprise since the main venue for this field is the ontology matching workshop held in International Semantic Web Conference (ISWC), where there has been several papers and posters along with the alignment contest. Aside from conference papers, journal articles comprise 25% of the publications and are ranked as second type of publications in ontology matching. *Conference reviews* and *book chapters* are the other major types of articles in this domain.

2.5.2. OUTPUTS IN TOP PERCENTILES WORLDWIDE

In this subsection, the appearance of ontology alignment research outputs in top-cited and journal percentiles in the six most recent years is explored.

We first look into the ratio of ontology alignment publications in the top-cited percentiles. The distribution of these top-cited articles in the six most recent years is shown in Figure 2.8, where the lighter color shows the percentile in top 10% most cited and the darker denotes the percentile in top 1% most cited articles worldwide. According to this figure, ontology alignment research outputs constitute one percent, 0.9%, and 0.5% of the top 1% most cited articles worldwide in years 2013, 2015, and 2017, respectively. At the same time, there is no ontology alignment output in the top 1% most cited for 2014

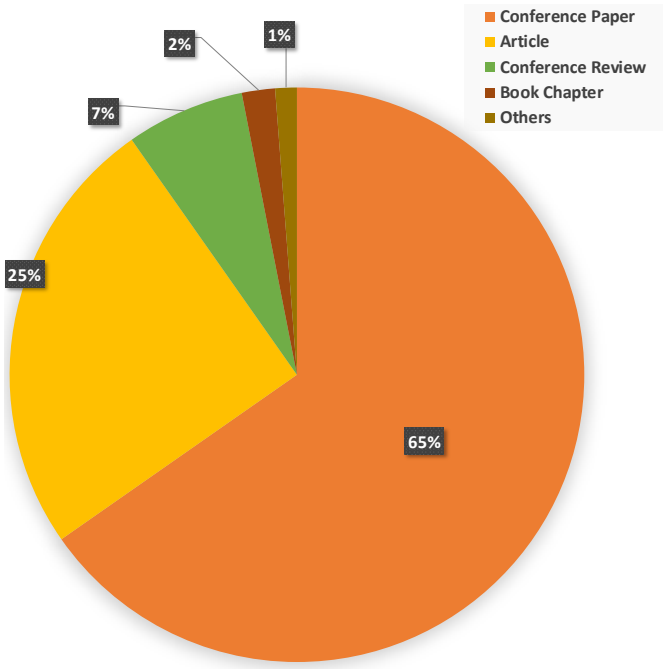


Figure 2.7: The types of documents published about ontology alignment on Scopus between 2001 to 2018.

and 2016. It is also readily seen that ontology alignment outputs form 6.1%, 5.3%, and 5.8% of the top 10% most cited article worldwide in years 2013, 2016, and 2018. Interestingly, although the research outputs from 2016 and 2018, which have a considerable amount of papers in the top 10% most cited articles, they do not have any in the top 1% most cited research outputs worldwide. The top-cited articles in the six most recent years are tabulated in Table 2.3. As expected, four of these articles are published in 2013 so that this year is considered as the best year in the six most recent years in terms of the number of research outputs in the top 1% and the top 10% most cited articles.

We also analyze the share of ontology alignment in the top journals identified by CiteScore. Figure 2.9 illustrates the ratio of ontology alignment outputs in top journals from 2013 to 2018. According to this figure, the ontology alignment share in the top 1% journals is 2.1%, 1.7%, 1.9%, and 1.5% for 2013, 2015, 2017, and 2018, respectively. Similarly, the share in the top 10% journals is 13.8% as the highest, followed by 11.8% in 2014, 9.9% in 2015, and 9.2% in 2018. There are a steady decrease and increase in publishing in the top 10% journals from 2013 to 2016 and from 2016 to 2018, respectively.

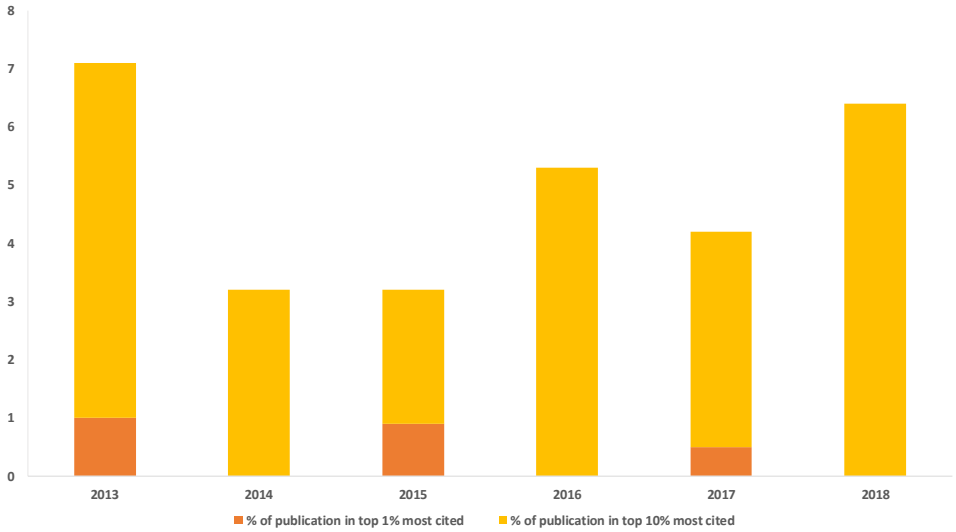


Figure 2.8: The share of ontology alignment research outputs to the top 1% and the top 10% most cited articles published in all disciplines.

Table 2.3: Five top-cited publications in ontology alignment in the six most recent years.

	Title	Main Authors	Year
1	Ontology matching: State of the art and future challenges	Shvaiko, P., Euzenat, J.	2013
2	Ontology matching: Second edition	Shvaiko, P., Euzenat, J.	2013
3	Ontology matching: A literature review	Otero-Cerdeira, L., Rodríguez-Martínez, F.J.	2015
4	Scaling semantic parsers with on-the-fly ontology matching	Kwiatkowski, T., Choi, E., Artzi, Y.	2013
5	The AgreementMakerLight ontology matching system	Faria, D., Pesquita, C., Santos, E.	2013

2.5.3. DISCIPLINES CONTRIBUTING TO ONTOLOGY ALIGNMENT

In this section, we consider the disciplines that contribute to ontology alignment. In this regard, we take advantage of all science journal classification (ASJC) used in Scopus and visualize the main areas along with their subcategories that contribute to the growth and evolution of the ontology alignment field.

Figure 2.10-(a) displays the main subject areas contributed to ontology alignment based on publication data between 2001 and 2018. As expected, computer science is the area with the maximum number of publications and constitutes 55% of the overall research articles. More in detail, Figure 2.10-(b) displays the subcategories of computer science contributing to ontology alignment. According to this figure, *general computer science* has the most of published articles, followed by *software* (12.2%), *computer net-*

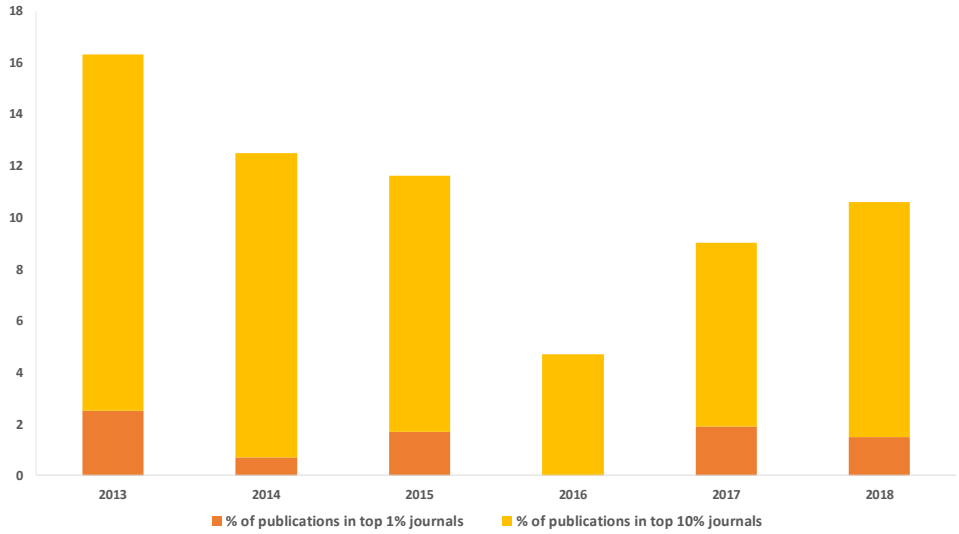


Figure 2.9: The share of ontology alignment research outputs to the top 1% and the top 10% journals of all disciplines.

works and communication (11.8%), and *information systems* (9.7%).

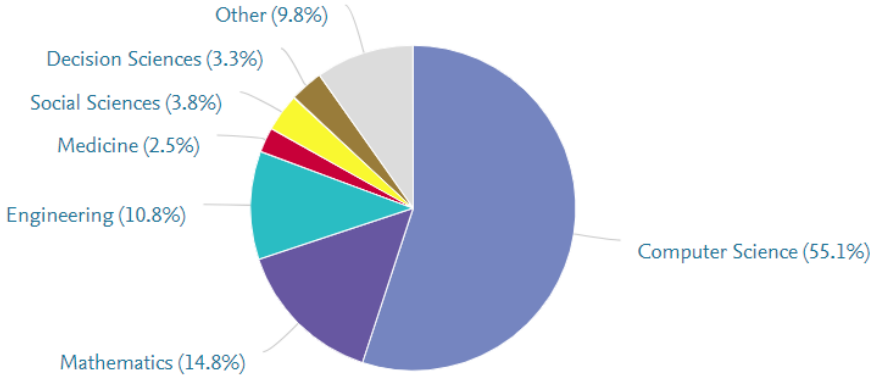
The second major discipline in the ontology alignment development is *mathematics*, which forms 14.8% of the overall research outputs. In particular, Figure 2.10-(c) illustrates the subcategories of mathematics, which indicates that *theoretical computer science* makes up 60.9% of the overall articles related to this category, and it is followed by *general mathematics* (9.3%) and *modeling and simulation* (9.0%).

The third main area is *engineering*, which forms 10.8% of all publications in this field. More in detail, Figure 2.10-(d) displays the subcategories of engineering contributing to ontology alignment. According to this figure, *control and system engineering* has 29.7% of publications and is followed by *general engineering* (26.4%) and *electrical and electronic engineering* (20.4%).

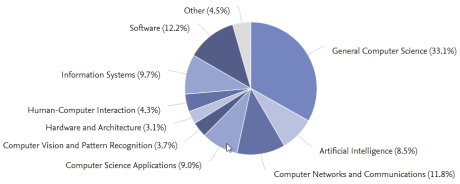
Social science is another important area contributing to ontology alignment, which constitutes 3.8% of all publications. Figure 2.10-(e) shows the subcategories of social science which have the highest share in ontology alignment research outputs. According to this figure, *library and information system* has the largest part of publications, i.e., 39.2%, and *education* (19.6%) and *linguistic and language* (16.5%) follow it.

Decision science has 3.3% of overall research outputs in ontology alignment. Figure 2.10-(f) shows that *information system and management* (83.3%), *general decision science* (11.1%), and *management science and operations research* (5.6%) are the subcategories in this category with contributions to the development of ontology alignment.

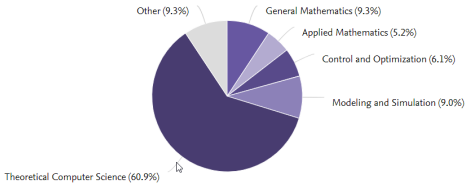
As the last subject area contributing to ontology alignment, *medicine* makes up 2.5% of publications. More in detail, Figure 2.10-(g) displays the subcategories of this area with the shares in research outputs. According to this figure, *health informatics* (70.5%), *general medicine* (8.2%), and *medicine (miscellaneous)* (8.2%) are the subcategories with



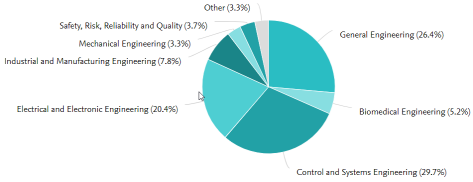
(a) The main subject areas of ontology alignment.



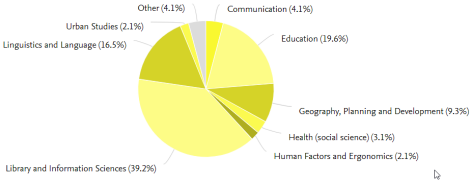
(b) Subcategories of computer science



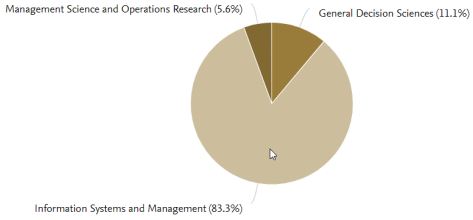
(c) Subcategories of mathematics



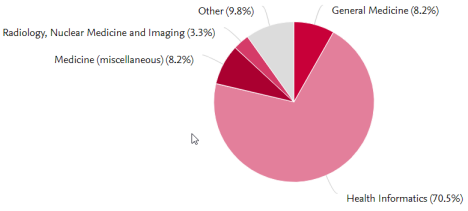
(d) Subcategories of engineering



(e) Subcategories of social science



(f) Subcategories of decision science



(g) Subcategories of medicine

Figure 2.10: The Disciplines and their associated subcategories contributed to ontology alignment.

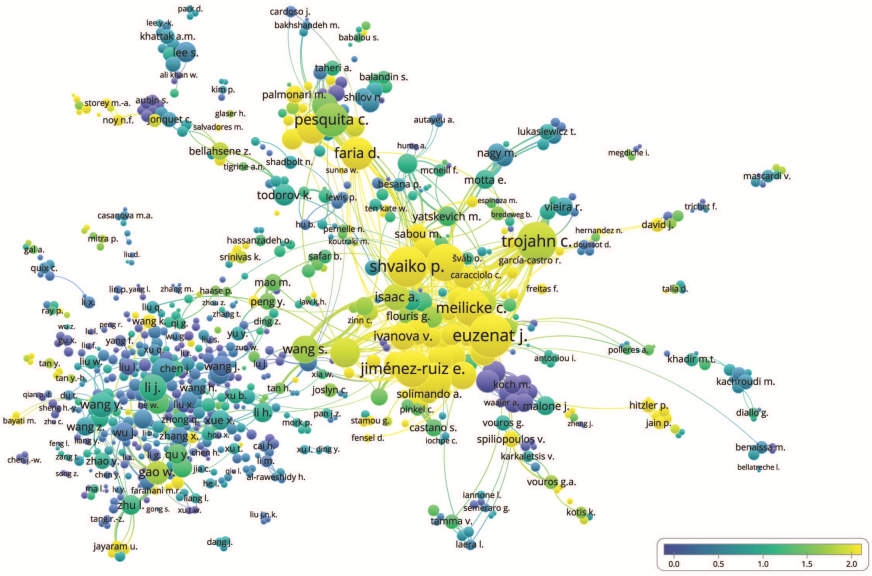


Figure 2.11: Collaborations of authors in ontology alignment based on the bibliometric data 2001-2018. The size of nodes is proportionate to the number of collaborative publications by the associated author, the color of nodes represent the average citations of authors per paper, and the thickness of each edge is commensurate with the number of collaborative articles of the authors at the two ends.

contributions to ontology alignment.

2.6. RESEARCH COLLABORATION IN ONTOLOGY ALIGNMENT

The collaboration of researchers within an area widens the impact and scope of the corresponding scientific field. As a result, it is of the essence to monitor, and even encourage, collaboration between different researchers and institutes all over the world. In this section, research collaboration in ontology alignment is investigated. In this regard, we consider the collaboration between authors as well as countries, and identify their most collaborative elements. We further analyze the trend of collaboration in recent years and academic-corporate collaboration in ontology alignment.

2.6.1. AUTHOR COLLABORATION

Collaboration among authors of ontology alignment is visualized by Figure 2.11 based on the bibliometric data 2001-2018. The authors in this graph are represented by nodes, where their size is proportionate to the number of collaborative articles, and their color denotes the average number of citations per publication according to the collected data from 2001-2018. Also, the thickness of each edge between two authors is commensurate with the number of collaborative publications of the authors at the two ends.

According to Figure 2.11, the organizers of the OAEI located at the center of the figure have great collaboration, and the most collaborative authors are also coming from this

Table 2.4: The ontology alignment researchers with the maximum number of collaborative publications. The first column tabulates the number of collaborative articles, the second column denotes the number of all collaborative researchers, and the third column is the average citation per publication.

Author	# Co-authored Articles	# All Co-authors	Average Citations
Euzenat J.	82	240	70.92
Shvaiko P.	63	205	116.58
Jiminez-Ruiz E.	63	168	16.86
Trojahn C.	59	204	13.98
StuckenSchmidt H.	59	177	18.95
Ferrara A.	56	163	31.76
Meilicke C.	54	160	22.72

community. In particular, J. Euzenat, P. Shvaiko, and E. Jimenez-Ruiz have 82, 63, and 63 co-publications, respectively, and lead the list of the author collaborations. Table 2.4 tabulates the top collaborative authors in ontology alignment. In this table, the number of articles co-authored with others, the total number of all co-authors, and the average citation per paper are displayed in the first to third columns, respectively. According to this table, J. Euzanat, P. Shvaiko, and C. Trojahn have the maximum number of collaborations in terms of the total number of co-authors.

Aside from the OAEI organizers, the developers of AgreementMaker and AgreementMakerLight (AML) [114], i.e., D. Faria, C. Pesquita, and I. Cruz, that are positioned at the top of Figure 2.11, have significant collaborations with each other. Also, D. Faria is one of the OAEI organizers, and along with C. Pesquita, have several collaborations with the OAEI community as well. Other members in this cluster have collaborations solely with each other.

Another dense cluster is placed at the bottom left of Figure 2.11. A closer look at the authors indicates that they are mainly from China, and have few collaborations with researchers outside of their country. The exceptions are S. Wang with 46, Juanzi Li with 48, and Yingjie Li with 40 co-authorship, including collaboration with other groups around the world. The main reason of such collaborations is that they mainly studied outside of China, and had a better opportunity for international collaborations. Other authors from this community who had mainly collaborated intra-nationally are Y. Wang with 44, J. Wang with 37, and S. Zhang with 32 co-authorships.

The color of nodes in Figure 2.11 is proportionate to the average number of citations, where nodes closer to the yellow color have a higher number of citations per paper. It is interesting to see that the OAEI organizers, who vastly collaborate with other researchers, have the maximum citation average compared to others. Based on Table 2.4, Shvaiko with 116.58 citations per publication leads the author list in terms of average citations, followed by J. Euzenat with 70.92, and A. Ferrara with 31.76 citations per paper. Another important point is about the Chinese community. Among these researchers, the authors who have collaborations with international communities have higher average citations. This confirms previous studies that multinational research collaboration are associated with increased citations [115, 116]. In general, international research collaboration is recognized as a means of cultivating research quality, enhanced resource uti-

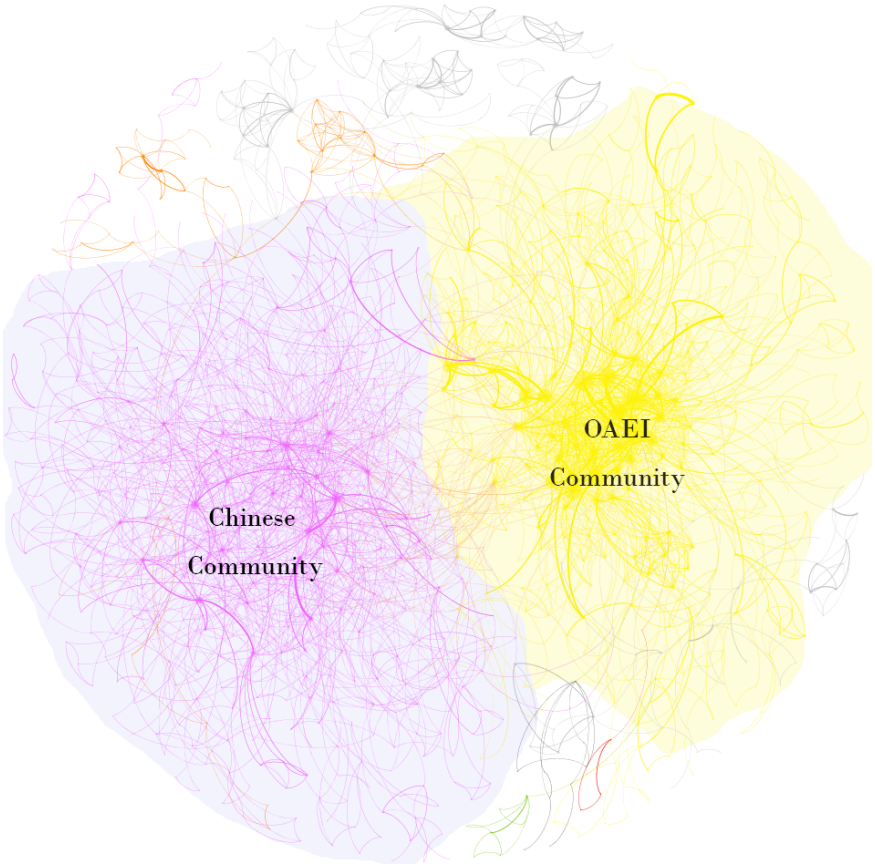


Figure 2.12: Communities of collaborations in ontology alignment based on the bibliometric data 2001-2018.

lization, and high impact [115, 117]. It also has indirect strategic, economic, or political benefits [118]. In fact enlarging team sizes, increasing interdisciplinarity, and intensifying ties across institutional and geographic borders is a signal of the field evolution from a solitary enterprise to an expanding social movement [115].

To further detect the communities of collaborations in ontology alignment, the co-authorship network in Figure 2.11 was subjected to a community detection algorithm, and the major collaborative communities were detected accordingly. There are quite a few community detection algorithms [119], and we choose Louvain algorithm [120], due to its speed, scalability, and simplicity [121, 122]. It is also one of the most popular community detection algorithms and has been implemented in many software and programming packages such as Gephi. Figure 2.12 plots the identified communities, where each community is depicted in a particular color. Six communities were identified by the algorithm, two of which are quite significant and include 75% of all researchers in this domain, and are shown in yellow and purple colors. The former is the OAEI organizers who have created a very large community. The thickness of edges of this network also

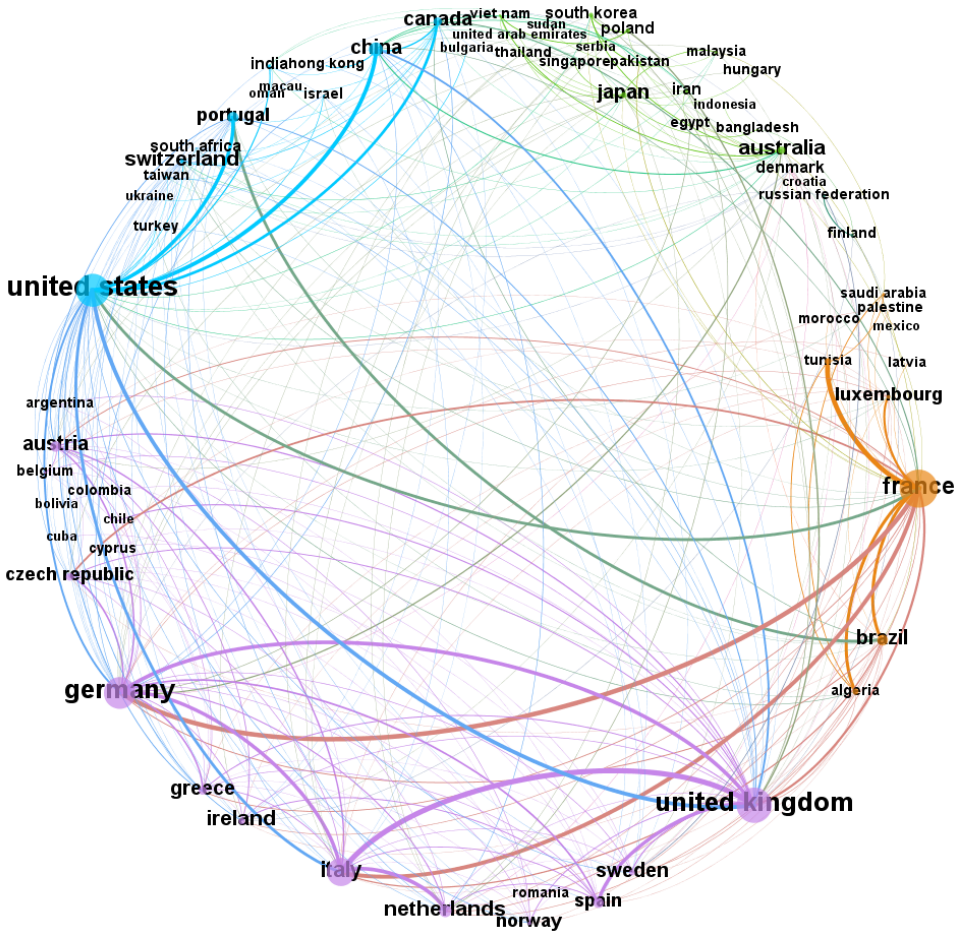


Figure 2.13: Collaborations of countries in ontology alignment based on the bibliometric data 2001-2018. The size of nodes represents the number of all collaborative papers of researchers from the associated country.

indicates that researchers in the OAEI community greatly collaborate with each other. The other dominant community is the Chinese, in which the collaborations, though not as significant as the OAEI community, are remarkable. Also, this figure indicates that the collaborations between Chinese and OAEI communities are not of significance, and researchers primarily cooperate with other researchers from the same community.

2.6.2. COUNTRY COLLABORATION

In this section, collaborations between countries are discussed and visualized. Figure 2.13 displays the co-authorship between different countries using a graph. The nodes of this graph are the countries with the size of node being proportionate to the number of publications collaborating with other countries, and the strength of the edge between each pair of countries is commensurate to the number of publications written jointly by

authors of the corresponding countries. According to this figure, France, UK, Germany, the US, and Italy are respectively the top five countries having the most collaborative papers worldwide.

The top five countries have the most collaborative research outputs together. For French authors of ontology alignment publications, the international collaboration is mostly with Germany (with 13% share of all collaborative research outputs), Italy (12%) and the US (8%). For the UK, the co-authorship is most frequently with Italy (14%), Germany 13%, and the US (11%). German international co-authorships are mostly dominated by collaborations with France (14%), the UK (13%), and Italy (12%). US international ontology alignment collaborations are also most commonly with the UK (12%), then with researchers from France (10%) and Italy (9%). Italian ontology researchers most frequently collaborate with researchers from the UK (16%), then with colleagues from France (14.3%) and Germany (14%). In sum, these five countries are the most collaborative countries worldwide that mainly cooperate with each other.

We further analyze the collaboration between countries along with the number of their publications and citations they have received. In this regard, Figure 2.14 plots the collaboration between countries with the size of nodes being proportional to the number of publications of the country, while Figure 2.15 shows the same graph with the size of nodes being commensurate to the average number of citations per publication. According to Figure 2.14, China has the maximum number of publications among other countries with 578 research outputs forming around 20% of all ontology alignment research outputs. At the same time, the number of citations of this country is not commensurate to the number of publications. That is to say, research outputs with at least one Chinese author have not gained enough attention. One of the primary reasons is the lack of international collaborations, as Figure 2.13 simply explains. The top five collaborative countries are readily seen to have more publications and average citations among other countries. This corroborates the importance of collaborations, since collaborations increase the visibility of research outputs to a wider audience, and the research studies thus get the attention they deserve. Another important point is the positive correlation between co-authorships and the size of publication outputs. It also makes sense since cooperation between researchers helps the use of the common wisdom, which is then emerged as more publications for the whole group.

2.6.3. INTERNATIONAL COLLABORATION

In this subsection, we take a closer look at the levels of collaborations in the six most recent years. In this regard, we count the number of research outputs published by authors from different countries (international collaboration), people from different institutes within a country (only national collaboration), different researchers of an institute (only institutional collaboration), and single-authored ones (no collaboration). Figure 2.16 plots the share of ontology alignment published articles in each of these categories. According to this figure, as expected, the maximum collaboration is among the researchers from the same institutes, and the only national and only institutional collaborations have approximately similar shares of ontology alignment research outputs. Single-authored articles are at the bottom of this list and from around 10% of overall publications.

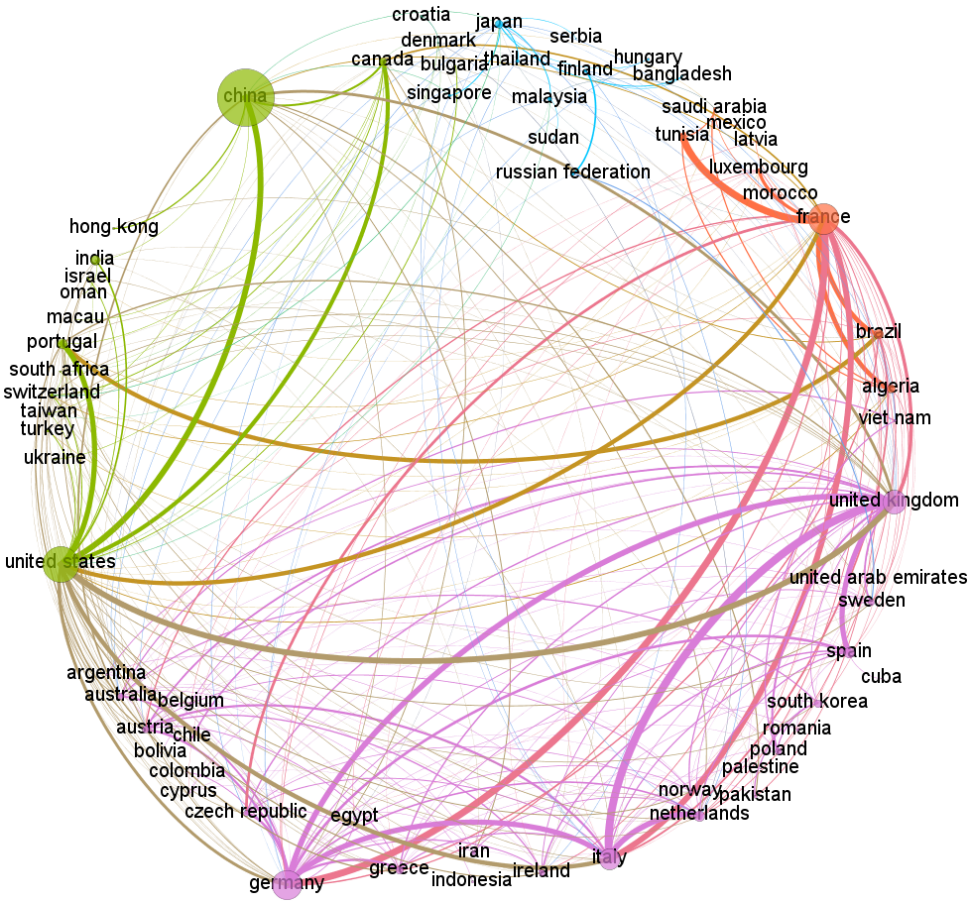


Figure 2.14: Collaborations of countries in ontology alignment based on the bibliometric data 2001-2018. The size of nodes is proportionate to the number of published documents with at least one author from the associated country.

It is also seen from Figure 2.16 that international collaboration has increased in the six most recent years, from 18.9% in 2013 to 27% in 2018. The maximum international collaborations is also in the year 2017, which constitute 28.9% of all ontology alignment research outputs. The collaborations inside the institutions have experienced a significant decrease, from 54.6% in 2013 to 45.5% in 2018. Similarly, the collaborations between institutions within a country have declined, while the single-authored articles have been growing. Ironically, the share of international collaborations (the most desired) and the share of no collaboration have been increased together over the last few years.

2.6.4. ACADEMIC-CORPORATE COLLABORATION

In this section, we consider the relationship between academic institutions and corporations based on ontology alignment published articles in the six most recent years.

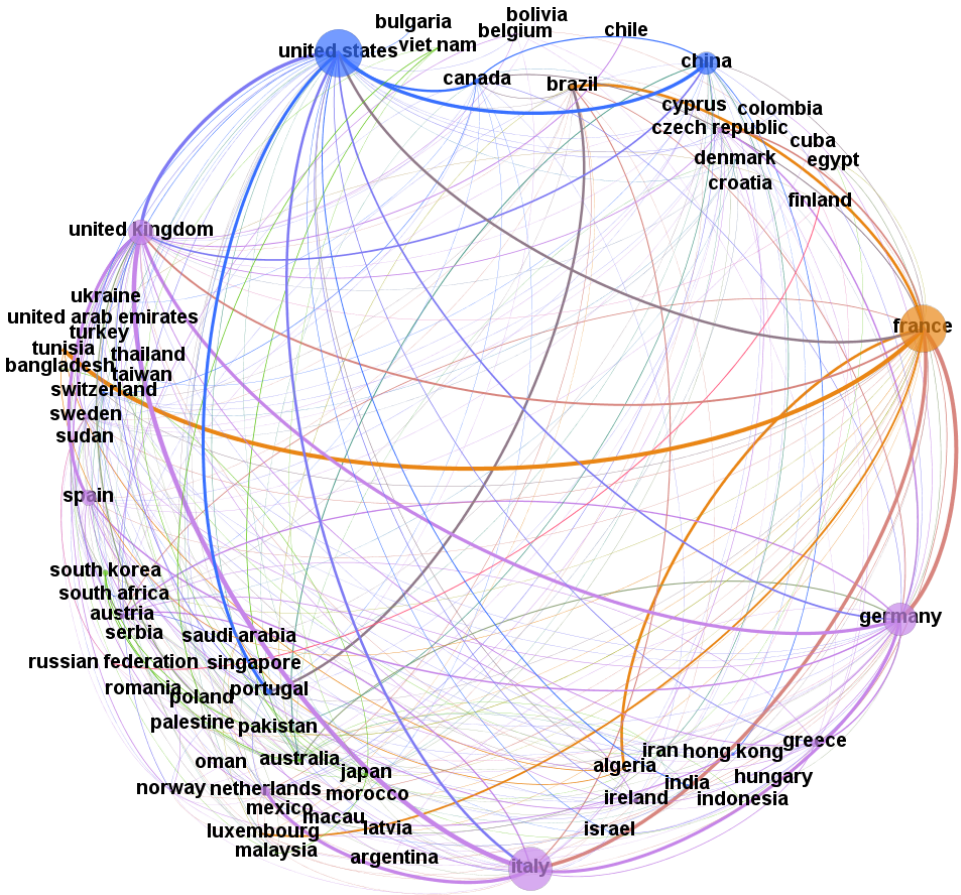


Figure 2.15: Collaborations of countries in ontology alignment based on the bibliometric data 2001-2018. The size of nodes is proportionate to the number of citations of the corresponding country.

In an academic-corporate relationship, collaboration is of the essence for both sides. It helps the academia to ensure industrial relevance in its research [123], and on the other hand, it provides the opportunity of knowledge complementary and risk sharing with the corporations [124]. Figure 2.17 displays the share of published articles by the academic-corporate collaborations. According to this figure, a tiny portion of papers is published jointly by academia and corporations, with a maximum of 3% in 2014. The minimum portion is also from years 2013 and 2018 with 1.7% of all ontology alignment published articles in the corresponding years. This figure elaborates that the relation between industry and academic institutions is not strong enough, since vast majority of collaborations is inside either academia or industry.

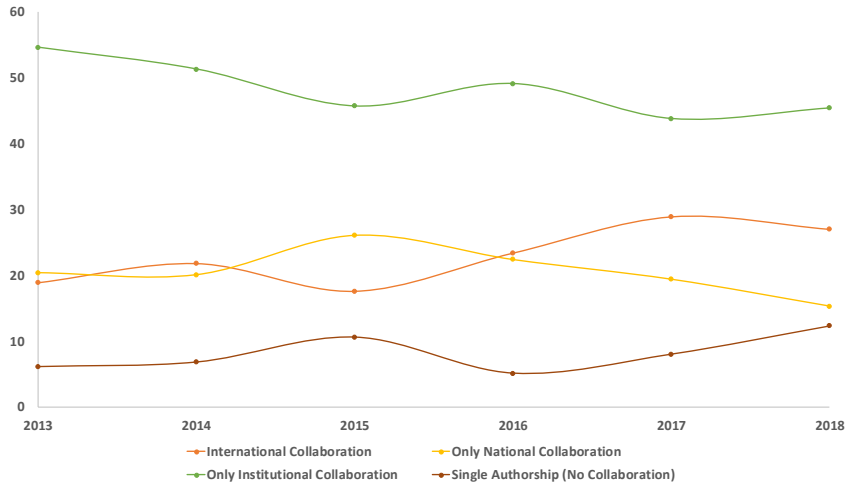


Figure 2.16: The share of different types of Collaborations in ontology alignment in the six most recent years.

2.7. CONTRIBUTION AND IMPACT IN ONTOLOGY ALIGNMENT

In this section, the impact of authors and countries that are active in ontology alignment is investigated. In this regard, we count the number of documents published per authors or countries and the number of citations of their documents to analyze their influence on the field of ontology alignment.

2.7.1. CONTRIBUTION AND IMPACT OF AUTHORS IN ONTOLOGY ALIGNMENT

In this subsection, the influence of authors on ontology alignment is investigated by counting the publications of top authors and their number of citations. Figure 2.18 plots the publication count of the top 10 authors in ontology alignment. The top 10 authors have published around 308 articles spanning from 2001 to 2018, which constitute 10.3% of all publications in ontology alignment. According to Figure 2.18, J. Euzenat tops the list of authors with 70 publications in ontology alignment, and C. Trojahn with 50, E. Jimenez-Ruiz and H. Stuckenschmidt with 45 and P. Shvaiko with 38 research outputs follow.

We further analyze the ontology alignment authors in terms of their number of citations. In this regard, Figure 2.19 plots the top 10 authors with the maximum number of citations. According to this figure, J. Euzenat leads the list with 4,610 citations, followed by P. Shvaiko with 3,847, M. Scholemmer with 1,040, and Stuckenschmidt with 834 citations.

To provide a broader view of the author impacts on ontology alignment, we visualize the citation map of the author using VOSviewer. Figure 2.20 visualizes the citation map of ontology alignment authors, where the size of nodes and their colors are proportionate to the citations and average citations per paper, respectively. Also, the thickness of edges is proportional to the number of times the authors at the two ends have cited

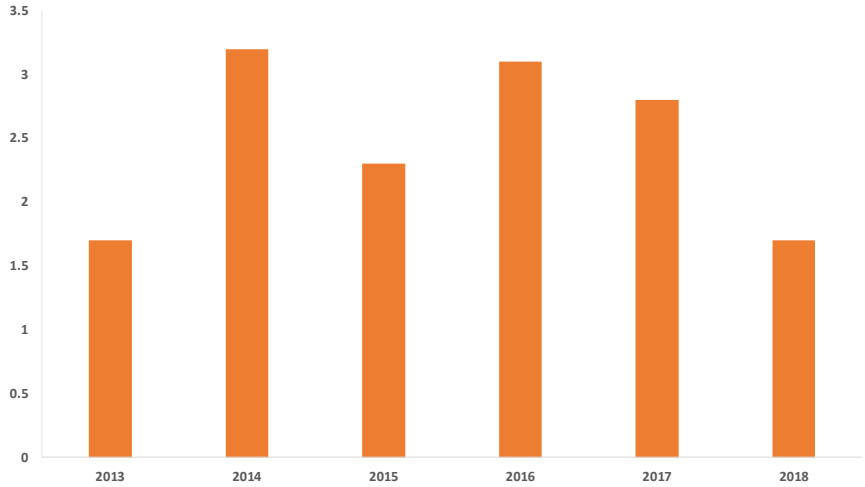


Figure 2.17: The share of academic-corporate collaborations in ontology alignment.

each other. Aside from the top authors shown in Figure 2.19, another critical observation from Figure 2.20 is that the researchers who are active in the OAEI have received more attention in comparison to others. The bottom-right region of this figure shows that the OAEI organizers and participants have higher average citations than other researchers.

Aside from the OAEI organizers and participants, several other researchers have been cited well. Y. Kalfoglou and M. Schorlemmer have a well-cited review paper published in 2003, and two methods for ontology alignment published in 2002 and 2003 [125–127]. N. Noy has also conducted several fundamental research in ontology alignment in the first years of this century [76, 77, 113, 128]. Y. Li, J. Tang, and J. Li have developed (along with colleagues) the ontology alignment system, called RiMOM, where their seminal work was published in 2008 [129, 130]. M. Ehrig has published the book entitled *ontology alignment: bridging the semantic gap* in 2006, which has got more than 600 citations to date [131]. As one can readily realize, these researchers have received a great amount of attention because of their research in the first decade of this century. However, there are several well-cited researchers among the OAEI organizers, whose research studies are more recent. The examples are E. Jiminez-Ruiz who developed LogMap [101], D. Faria and C. Pesquita, who developed AML [114]. As a result, the attention of ontology alignment is more focused on the OAEI in recent years, and other researchers in this field have not drawn significant attention.

2.7.2. CONTRIBUTION AND IMPACT OF COUNTRIES IN ONTOLOGY ALIGNMENT

In this section, we analyze the impact and contributions of different countries on the evolution of ontology alignment. First, we visualize the number of ontology alignment publications for each country. Figure 2.21 displays the top 10 countries in terms of their publication count. According to this figure, China leads the list with 578 publications,

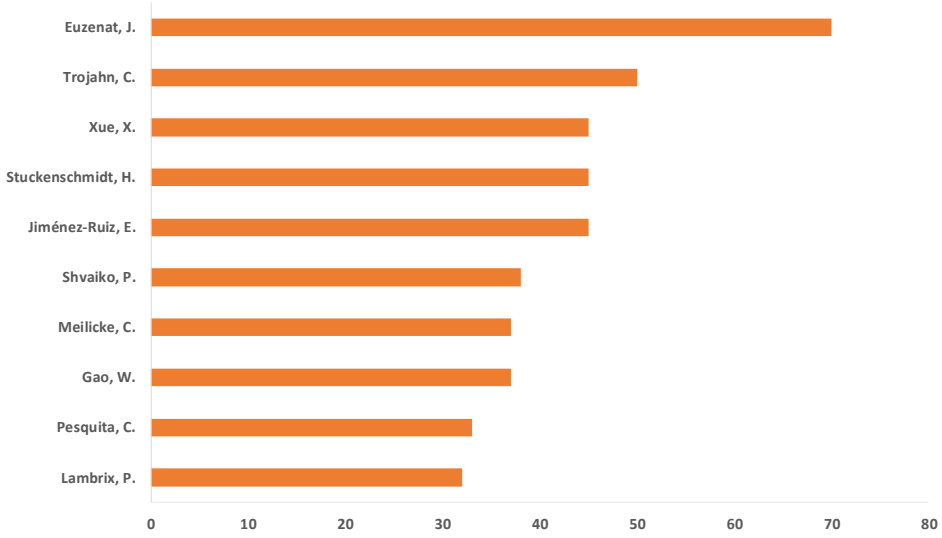


Figure 2.18: The top 10 authors of ontology alignment in terms of their number of published documents based on the bibliometric data 2001-2018.

which has a quite well difference with the second country, the US, with 360 publications in ontology alignment. France with 306, Germany with 295, and the UK with 234 publications in ontology alignment are the next countries in this list.

We further analyze the countries concerning their number of citations. Figure 2.22 shows the top 10 countries regarding the number of citations. Based on this figure, the US tops the list with 6,662 citations, followed by France with 6,563 and Italy with 6,027 citations. China, while has the maximum number of ontology alignment citations, is the sixth country in this ranking with 3,080 citations overall.

Besides, we visualize the citation network of countries using VOSviewer. Figure 2.23 illustrates the citation network, where the nodes are the countries whose size are proportionate to the number of citations based on the bibliometric data 2001-2018. According to this figure, the number of citations of Chinese researchers is not commensurate with their number of publications: While they publish more than any country, they are ranked sixth concerning the number of citations. Italy has the highest average citations per publication with 29.54 for 204 documents, followed by France with 21.45, Spain 20.04, the US with 18.51, Germany with 15.33, and the UK with 14.65 citations per publication. Moreover, based on Figure 2.23, the US has the maximum number of citations, followed by France, Italy, Germany, and the UK.

In addition, the community detection algorithm has been applied to the network in Figure 2.23. Every community in this figure represents a set of countries whose researchers refer to each others' studies more often. Basically, two major communities were detected that are shown in purple and green colors. One community comprises the US, UK, China, and Spain along with some other countries with less significant contribution to ontology alignment. Another community consists of France, Italy, and Germany

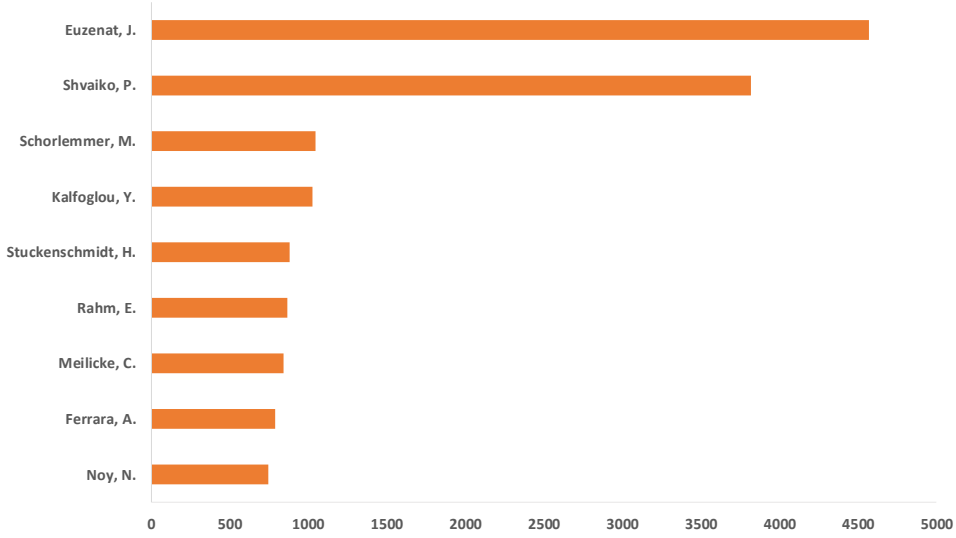


Figure 2.19: The top 10 authors of ontology alignment in terms of their number of citations based on the bibliometric data 2001-2018.

as the major countries along with some less considerable ones such as the Netherlands and Portugal.

2.8. CONCLUSION AND DISCUSSION

In this chapter, we revisited ontology alignment by revising and analyzing the ontology alignment publications. First, a recent framework for classifying ontology alignment publications was revisited and discussed in detail. The classifications of ontology alignment research outputs highlighted the fact that several areas have not received enough attentions. For instance, no methodology for comparing alignment systems has been developed, while one of the primary motivations of the OAEI is to compare the alignment systems together. In the following chapters, we address comparing ontology alignment comparison by looking into several methods from statistics and multi-criteria decision-making (MCDM) and put forward proper tools and methodologies to compare the alignment systems.

In addition, the ontology alignment problem can be solved more efficiently, especially in the light of the aim of this dissertation for enabling interoperability in logistics that requires a fast alignment system. In this regard, we adopt an evolutionary algorithm that can solve the non-deterministic polynomial-time (NP) ontology alignment problem more efficiently in terms of time and memory complexity, as well as precision and recall.

Another important point, which also accented by the bibliometric analysis, was a scarcity of practical applications for ontology alignment. This issue would restrict the applicability of ontology alignment and hinders its progress. In this regard, we study the applicability of ontology alignment to logistics, where interoperability is of the utmost

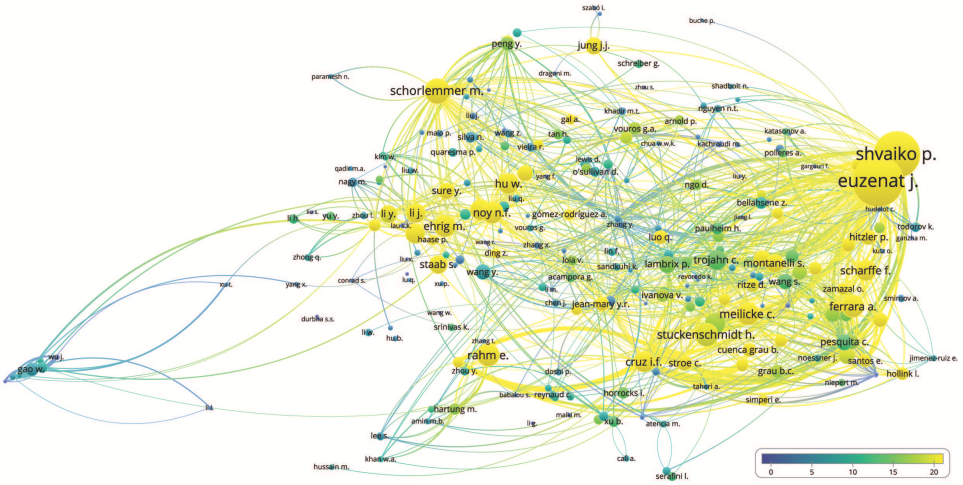


Figure 2.20: The citation map of ontology alignment researchers, where the size of nodes is proportionate to the number of citations and its color represents the average citations per paper for the associated researcher. The thickness of edges in this network is proportional to the number of times the authors at the ends have cited each other.

importance and heterogeneity is inevitable.

We also made an inquiry in Scopus and extracted articles pertinent to ontology alignment. After inspecting the articles and excluding irrelevant items to ontology alignment, 2,975 articles remained based on which bibliometric analyses were carried out. We conducted two classes of bibliometric analyses, semantic and structural. Semantic analysis entails the overall discovery of concepts, notions, and research lines flowing underneath ontology alignment, while the structural analysis provides a meta-level overview of the field by probing into the collaboration network and citation analysis in author and country levels. Each of these analyses was divided into two subcategories. In the semantic analysis, we first conducted a topic modeling on the extracted bibliometric data by subjecting title, keywords, and abstracts of articles to the latent Dirichlet allocation (LDA), a well-established statistical method for modeling topics. Although the topics were detected based merely on the frequency of the words in the articles, the identified topics readily referred to a problem that ontology alignment can address or a domain to which it has been applied. The other semantic analysis was thematic, wherein the number of annual publications, the types of research outputs, the share of ontology alignment in top-cited articles and top journals, and contributing disciplines to ontology alignment were explained and discussed in detail.

The second class of analyses was structural, wherein we carried out two classes of probings. First, we studied the collaborations of ontology alignment researchers. We observed that international collaboration has been improved over the last few years. Also, the collaboration between academia and corporations were gauged, which was not significant enough. We also observed that ontology alignment researchers fall into two major communities. The first community comprises the OAEI organizers, and the second

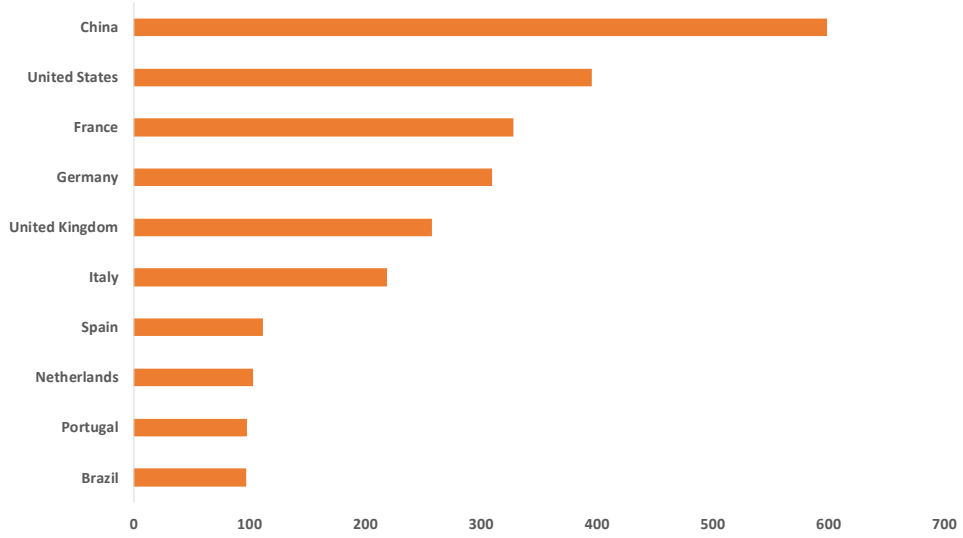


Figure 2.21: The top 10 countries of ontology alignment in terms of their number of published documents based on the bibliometric data 2001-2018.

consists of the Chinese scholars. Although the researchers in these communities cooperate closely with each other, the communities, especially the Chinese, seem like isolated islands that do not interact with the other researchers from other communities. We also observed that, although the number of publications with at least one Chinese researcher is considerable, they do not get enough attention, possibly the attention that they deserve. Aside from authors, we also analyzed the collaborations between countries and realized that the top five most collaborative countries mostly work together more than any other country in the world. Another structural analysis was regarding the impact and contribution of researchers and countries for the field of ontology alignment. In these analyses, we identified the authors and countries with maximal influence on the field by counting their number of publications and citations. We also visualized their citation network, where we identified two hosts of countries that mostly cite each other.

We observed that the articles from the *ontology matching workshop* are not indexed by Web of Science (WoS). This workshop is particularly important for the ontology alignment community, since highly-cited researchers from the field actively participate in the venue. Also, the articles from the workshop represent the state-of-the-art challenges and novelties in this domain. In addition, the OAEI contest is also a part of this workshop, wherein new challenges are introduced and new systems are developed to overcome those challenges.

Based on the contributions of ontology alignment to the top-cited articles and top journal percentiles, we perceived that ontology alignment is indeed a very essential field of study, research outputs of which receive significant attention, and fundamental research studies are conducted and published in top journals. However, this field has only one dedicated venue, the ontology matching workshop, that is not indexed by WoS. This

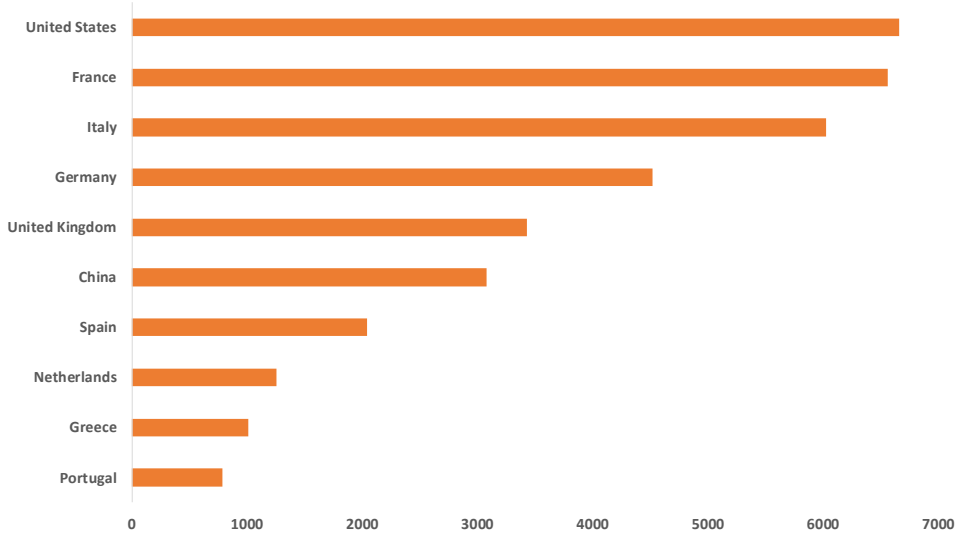


Figure 2.22: The top 10 authors of ontology alignment in terms of their number of citations based on the bibliometric data 2001-2018.

calls for having more serious venues such as a dedicated journal for ontology alignment. This is essential also because the pertinent articles to ontology alignment are dispersed in various journals, while the conference papers are mainly published in the ontology matching workshop. Having a journal devoted to ontology alignment helps the researcher in this field to focus on one venue. Also, there have been several special issues on ontology alignment over the last few years that are also hosted by different journals. A dedicated journal can be the host of such special issues as well.

The topics identified in the semantic analysis were directly related to the problems and applications related to ontology alignment. Some of these topics are well-established in ontology alignment, and the researchers of the field are aware of it. However, there are several other topics that are totally neglected by ontology alignment researchers, and the OAEI organizers in particular. By viewing the OAEI benchmarks, one readily gets the impression that ontology alignment is particularly useful for the biomedical domain, since most of the tracks are from this domain, i.e., anatomy, disease and phenotype, and large biomedical. While it is indisputable that ontology alignment has been successfully applied to various biomedical ontologies, its use and applicability are not restricted to this domain. We encourage the OAEI organizers and other researchers in the field to prepare several other standard benchmarks for evaluating ontology alignment systems. The benchmarks can be from Semantic Web Services, agent-based modeling, knowledge graphs, and business processes.

We also identified two primary communities, OAEI and Chinese, for ontology alignment, in both of which around 75% of ontology alignment researchers fall. The collaborations between these two communities are not significant, and researchers of these communities collaborate mainly with other researchers in the same community. Based

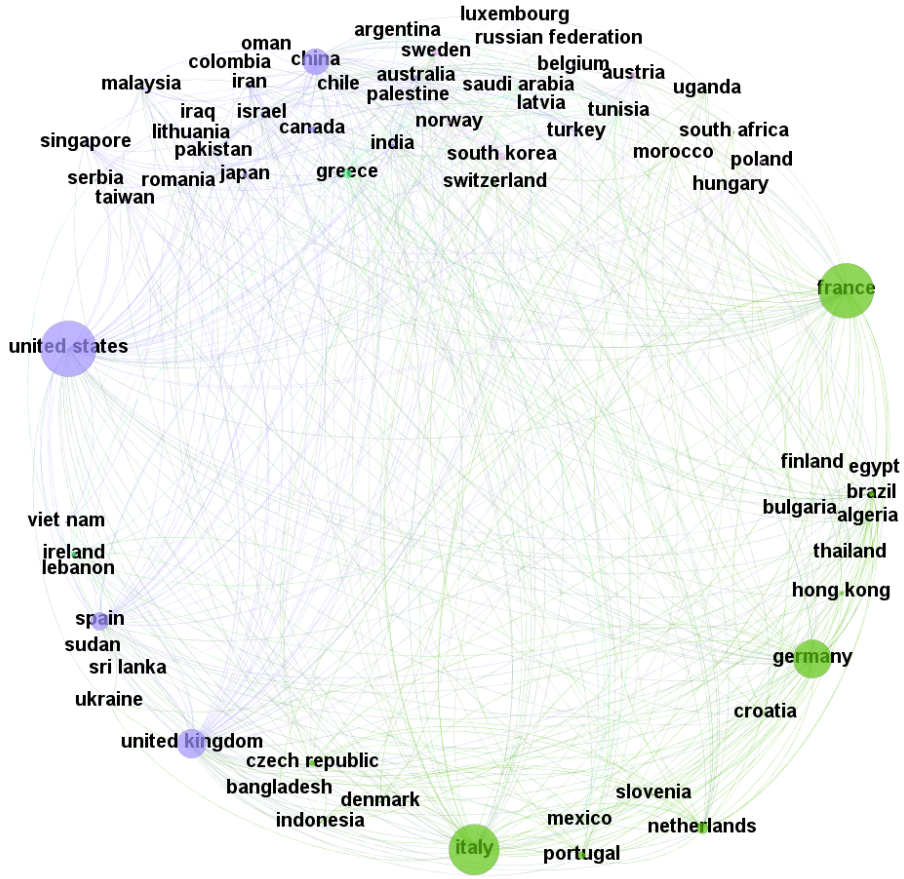


Figure 2.23: The citation map of countries contributing to ontology alignment, where the size of nodes is proportionate to the number of citations.

on the citation analysis, we realized that the research of the Chinese community does not get considerable attention, while the number of research outputs of this community is significant. One reason for not getting enough attention is probably the lack of international collaboration in the Chinese community. Another reason might be that the contributions of this community do not address the current state-of-the-art challenges so that they do not draw enough attention. In any case, ontology alignment should take advantage of the potentials in the Chinese community. Thus, the collaborations between these two communities are strongly recommended, making the Chinese researchers familiar with the current challenges in the domain. In addition, the Chinese community will have the opportunity to disseminate their research to a wider audience, and can play a more vital role for progressing the field of ontology alignment. A potential direction for future research is to inspect the research of these two communities to identify precisely what they are working on and on which potential topics they can cooperate.

We discussed and visualized the impact of researchers and countries on ontology alignment as well. We observed that the OAEI organizers and participants get a considerable amount of attention. There are several other researchers with significant attention, but their research studies were mainly carried out in the first decade of this century. Thus, if researchers want to get attention, they must be involved in the OAEI community and participate actively in the contest.

Another critical observation from the analyses in this chapter is the insignificant collaborations between academia and corporations. We observed that the academic-corporate relations constitute on average, around 2% of all publications in ontology alignment. One possible reason is that the teams inside the enterprises have the ability to resolve the problems. However, that does not explain the inconsiderable relation between academia and industry. The more realistic reason is that the companies have not realized that ontology alignment can enhance their business functions. This is also coming from the fact that most of the standard benchmarks are restricted to particular problems and domain, while ontology alignment can be widely used to address disparate issues. One way to increase the impact and use of ontology alignment is to conduct several qualitative case studies to show how ontology alignment can automate the manual procedures, and consequently, increase the profits of the companies. Another avenue for the progress of ontology matching is to dedicate more funding to the applicability of ontology alignment and find untapped domains and problems to which ontology alignment is a potential solution. Such projects create new benchmarks for the OAEI and extend the ontology alignment applicability so that a broader audience can understand the importance of the field. This can also help overcome the slow progress we have observed from 2013. The bibliometric analysis presented in this chapter also covered the lack of a literature review based on a quantitative approach that, to the best of our knowledge, has not been considered before.

REFERENCES

[1] A. Mallik and N. Mandal, *Bibliometric analysis of global publication output and collaboration structure study in microrna research,* Scientometrics **98**, 2011 (2014).

[2] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, *et al.*, *Science of science,* Science **359**, eaao0185 (2018).

[3] C. Chorus and L. Waltman, *A large-scale analysis of impact factor biased journal self-citations,* PLoS One **11**, e0161021 (2016).

[4] M. Thelwall, S. Haustein, V. Larivière, and C. R. Sugimoto, *Do altmetrics work? twitter and ten other social web services,* PloS one **8**, e64841 (2013).

[5] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, *Defining and identifying sleeping beauties in science,* Proceedings of the National Academy of Sciences **112**, 7426 (2015).

[6] Y. He and S. C. Hui, *Mining a web citation database for author co-citation analysis,* Information processing & management **38**, 491 (2002).

[7] P. O. Seglen and D. W. Aksnes, *Scientific productivity and group size: A bibliometric analysis of norwegian microbiological research,* Scientometrics **49**, 125 (2000).

[8] W. Glänzel, *Science in scandinavia: A bibliometric approach,* Scientometrics **48**, 121 (2000).

[9] U. Schmoch and T. Schubert, *Are international co-publications an indicator for quality of scientific research?* Scientometrics **74**, 361 (2008).

[10] Y. Ding, G. G. Chowdhury, and S. Foo, *Bibliometric cartography of information retrieval research by using co-word analysis,* Information processing & management **37**, 817 (2001).

[11] L. Bromham, R. Dinnage, and X. Hua, *Interdisciplinary research has consistently lower funding success,* Nature **534**, 684 (2016).

[12] A. M. Petersen, D. Majeti, K. Kwon, M. E. Ahmed, and I. Pavlidis, *Cross-disciplinary evolution of the genomics revolution,* Science advances **4**, eaat4211 (2018).

[13] C. Candia, C. Jara-Figueroa, C. Rodriguez-Sickert, A.-L. Barabási, and C. A. Hidalgo, *The universal decay of collective memory and attention,* Nature Human Behaviour **3**, 82 (2019).

[14] R. Guimera, B. Uzzi, J. Spiro, and L. A. N. Amaral, *Team assembly mechanisms determine collaboration network structure and team performance,* Science **308**, 697 (2005).

[15] L. Liu, Y. Wang, R. Sinatra, C. L. Giles, C. Song, and D. Wang, *Hot streaks in artistic, cultural, and scientific careers,* Nature **559**, 396 (2018).

[16] A. Ebrahimi Fard and C. Scott, *Assessing the readiness of the academia in the topic of false and unverified information,* To be appeared in ACM Journal of Data and Information Quality (JDIQ) .

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation,* Journal of machine Learning research **3**, 993 (2003).

[18] L. Otero-Cerdeira, F. J. Rodríguez-Martínez, and A. Gómez-Rodríguez, *Ontology matching: A literature review,* Expert Systems with Applications **42**, 949 (2015).

[19] S. M. Falconer, N. F. Noy, and M.-A. D. Storey, *Ontology mapping-a user survey.* in *OM* (Citeseer, 2007).

[20] P. Shvaiko and J. Euzenat, *A survey of schema-based matching approaches,* in *Journal on data semantics IV* (Springer, 2005) pp. 146–171.

[21] Y. K. Hooi, M. F. Hassan, and A. M. Shariff, *A survey on ontology mapping techniques,* in *Advances in Computer Science and its Applications* (Springer, 2014) pp. 829–836.

[22] J. Zhu, *Survey on ontology mapping,* Physics Procedia **24**, 1857 (2012).

[23] E. Droge, *Guidelines on ontology matching,* Information-Wissenschaft und Praxis **61**, 143 (2010).

[24] K. Kotis and M. Lanzenberger, *Ontology matching: current status, dilemmas and future challenges,* in *2008 International Conference on Complex, Intelligent and Software Intensive Systems* (IEEE, 2008) pp. 924–927.

[25] P. Shvaiko and J. Euzenat, *Ontology matching: state of the art and future challenges,* IEEE Transactions on knowledge and data engineering **25**, 158 (2013).

[26] P. Ochieng and S. Kyanda, *Large-scale ontology matching: State-of-the-art analysis,* ACM Computing Surveys (CSUR) **51**, 75 (2018).

[27] É. Thiéblin, O. Haemmerlé, N. Hernandez, and C. Trojahn, *Survey on complex ontology matching,* .

[28] S. Castano, A. Ferrara, D. Lorusso, and S. Montanelli, *On the ontology instance matching problem,* in *2008 19th International Workshop on Database and Expert Systems Applications* (IEEE, 2008) pp. 180–184.

[29] B. Fu, R. Brennan, and D. O'Sullivan, *Multilingual ontology mapping: Challenges and a proposed framework,* in *Workshop on Matching and Meaning-Automated Development, Evolution and Interpretation of Ontologies* (2009) pp. 1–31.

[30] P. Wennerberg, *Aligning medical domain ontologies for clinical query extraction,* in *Proceedings of the 12th conference of the european chapter of the association for computational linguistics: student research workshop* (Association for Computational Linguistics, 2009) pp. 79–87.

[31] B. Tomaszewski and E. Holden, *The geographic information science and technology and information technology bodies of knowledge: an ontological alignment,* in *Proceedings of the 13th annual conference on Information technology education* (ACM, 2012) pp. 195–200.

[32] B. Lauser, G. Johannsen, C. Caracciolo, W. R. van Hage, J. Keizer, and P. Mayr, *Comparing human and automatic thesaurus mapping approaches in the agricultural domain,* in *International Conference on Dublin Core and Metadata Applications* (2008) pp. 43–53.

[33] J. Euzenat, P. Shvaiko, *et al.*, *Ontology matching*, Vol. 18 (Springer, 2007).

[34] M. Cheatham and P. Hitzler, *String similarity metrics for ontology alignment,* in *International Semantic Web Conference* (Springer, 2013) pp. 294–309.

[35] G. A. Miller, *Wordnet: a lexical database for english,* Communications of the ACM **38**, 39 (1995).

[36] W. Cohen, P. Ravikumar, and S. Fienberg, *A comparison of string metrics for matching names and records,* in *Kdd workshop on data cleaning and object consolidation*, Vol. 3 (2003) pp. 73–78.

[37] M. Bayati, D. F. Gleich, A. Saberi, and Y. Wang, *Message-passing algorithms for sparse network alignment,* ACM Transactions on Knowledge Discovery from Data (TKDD) **7**, 3 (2013).

[38] G. Acampora, H. Ishibuchi, and A. Vitiello, *A comparison of multi-objective evolutionary algorithms for the ontology meta-matching problem,* in *Evolutionary Computation (CEC), 2014 IEEE Congress on* (IEEE, 2014) pp. 413–420.

[39] M. Martínez-Romero, J. M. Vázquez-Naya, F. J. Nóvoa, G. Vázquez, and J. Pereira, *A genetic algorithms-based approach for optimizing similarity aggregation in ontology matching,* in *International Work-Conference on Artificial Neural Networks* (Springer, 2013) pp. 435–444.

[40] X. Xue and S. Liu, *Compact evolutionary algorithm based ontology meta-matching,* in *International Conference on Smart Vehicular Technology, Transportation, Communication and Applications* (Springer, 2017) pp. 213–221.

[41] X. Xue and Y. Wang, *Optimizing ontology alignments through a memetic algorithm using both matchfmeasure and unanimous improvement ratio,* Artificial Intelligence **223**, 65 (2015).

[42] J. M. Gil, J. F. A. Montes, E. Alba, and J. Aldana-Montes, *Optimizing ontology alignments by using genetic algorithms,* (2008).

[43] X. Xue, J. Liu, P.-W. Tsai, X. Zhan, and A. Ren, *Optimizing ontology alignment by using compact genetic algorithm,* in *Computational Intelligence and Security (CIS), 2015 11th International Conference on* (IEEE, 2015) pp. 231–234.

[44] X. Xue, Y. Wang, and W. Hao, *Optimizing ontology alignments by using nsga-ii.* International Arab Journal of Information Technology (IAJIT) **12** (2015).

[45] J. Wang, Z. Ding, and C. Jiang, *Gaom: Genetic algorithm based ontology matching,* in *Services Computing, 2006. IEEE Asia-Pacific Conference on* (IEEE, 2006) pp. 617–620.

[46] J. Bock and J. Hettenhausen, *Discrete particle swarm optimisation for ontology alignment,* Information Sciences **192**, 152 (2012).

[47] L. Yujian and L. Bo, *A normalized levenshtein distance metric,* IEEE transactions on pattern analysis and machine intelligence **29**, 1091 (2007).

[48] M. Cheatham and P. Hitzler, *The properties of property alignment.* in *OM* (2014) pp. 13–24.

[49] G. Stumme and A. Maedche, *Fca-merge: Bottom-up merging of ontologies,* in *IJCAI*, Vol. 1 (2001) pp. 225–230.

[50] N. F. Noy, M. A. Musen, *et al.*, *Algorithm and tool for automated ontology merging and alignment,* in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831*, Vol. 115 (sn, 2000).

[51] Y. Ding and S. Foo, *Ontology research and development. part 2-a review of ontology mapping and evolving,* Journal of information science **28**, 375 (2002).

[52] J. Zhang, P. Lin, P. Huang, and G. Wu, *Research on semantic web service composition based on ontology reasoning and matching,* in *International Conference on Information Computing and Applications* (Springer, 2011) pp. 450–457.

[53] C. Trojahn, P. Quaresma, and R. Vieira, *Exploiting majority acceptable arguments for ontology matching,* International Journal of Artificial Intelligence **8**, 1 (2012).

[54] N. F. Noy and M. A. Musen, *The prompt suite: interactive tools for ontology merging and mapping,* International journal of human-computer studies **59**, 983 (2003).

[55] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh, *Ontology alignment for linked open data,* in *International semantic web conference* (Springer, 2010) pp. 402–417.

[56] Z. Duo, L. Juan-Zi, and X. Bin, *Web service annotation using ontology mapping,* in *Service-Oriented System Engineering, 2005. SOSE 2005. IEEE International Workshop* (IEEE, 2005) pp. 235–242.

[57] L. Obrst, R. E. Wray, and H. Liu, *Ontological engineering for b2b e-commerce,* in *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001* (ACM, 2001) pp. 117–126.

[58] F. Wiesman, N. Roos, and P. Vogt, *Automatic ontology mapping for agent communication,* in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2* (ACM, 2002) pp. 563–564.

[59] V. Ermolayev and M. Davidovsky, *Agent-based ontology alignment: basics, applications, theoretical foundations, and demonstration,* in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics* (ACM, 2012) p. 3.

[60] L. Laera, V. Tamma, J. Euzenat, T. Bench-Capon, and T. Payne, *Reaching agreement over ontology alignments,* in *International Semantic Web Conference* (Springer, 2006) pp. 371–384.

[61] A. Solimando, E. Jiménez-Ruiz, and G. Guerrini, *Detecting and correcting conservativity principle violations in ontology-to-ontology mappings,* in *International Semantic Web Conference* (Springer, 2014) pp. 1–16.

[62] A. Solimando, E. Jiménez-Ruiz, and G. Guerrini, *A multi-strategy approach for detecting and correcting conservativity principle violations in ontology alignments.* in *OWLED* (2014) pp. 13–24.

[63] M. Ehrig and Y. Sure, *Ontology mapping–an integrated approach,* in *European Semantic Web Symposium* (Springer, 2004) pp. 76–91.

[64] J. Euzenat, *Semantic precision and recall for ontology alignment evaluation.* in *IJCAI* (2007) pp. 348–353.

[65] X. Niu, H. Wang, G. Wu, G. Qi, and Y. Yu, *Evaluating the stability and credibility of ontology matching methods,* in *Extended Semantic Web Conference* (Springer, 2011) pp. 275–289.

[66] H. Paulheim, S. Hertling, and D. Ritze, *Towards evaluating interactive ontology matching tools,* in *Extended Semantic Web Conference* (Springer, 2013) pp. 31–45.

[67] C. Meilicke, H. Stuckenschmidt, and A. Tamilin, *Repairing ontology mappings,* in *AAAI*, Vol. 3 (2007) p. 6.

[68] C. Meilicke, *Alignment incoherence in ontology matching*, Ph.D. thesis, Universität Mannheim (2011).

[69] C. Pesquita, D. Faria, E. Santos, and F. M. Couto, *To repair or not to repair: reconciling correctness and coherence in ontology reference alignments,* in *Proc. 8th ISWC ontology matching workshop (OM), Sydney (AU), page this volume* (2013).

[70] W. R. van Hage, *Evaluating ontology-alignment techniques,* (2009).

[71] S. Cozzens, S. Gatchair, J. Kang, K.-S. Kim, H. J. Lee, G. Ordóñez, and A. Porter, *Emerging technologies: quantitative identification and measurement,* Technology Analysis & Strategic Management **22**, 361 (2010).

[72] M. Zitt and E. Bassecoulard, *Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences,* Information processing & management **42**, 1513 (2006).

[73] L. Leydesdorff and P. Zhou, *Nanotechnology as a field of science: Its delineation in terms of journals and patents,* Scientometrics **70**, 693 (2007).

[74] A. Kuzhabekova and J. Kuzma, *Mapping the emerging field of genome editing,* Technology Analysis & Strategic Management **26**, 321 (2014).

[75] A. L. Porter, J. Youtie, P. Shapira, and D. J. Schoeneck, *Refining search terms for nanotechnology,* Journal of nanoparticle research **10**, 715 (2008).

[76] N. F. Noy and M. A. Musen, *Smart: Automated support for ontology merging and alignment,* in *Proc. of the 12th Workshop on Knowledge Acquisition, Modelling, and Management (KAW'99), Banf, Canada* (Citeseer, 1999).

[77] N. F. Noy, M. A. Musen, *et al.*, *Algorithm and tool for automated ontology merging and alignment,* in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831*, Vol. 115 (sn, 2000).

[78] N. van Eck and L. Waltman, *Software survey: Vosviewer, a computer program for bibliometric mapping,* Scientometrics **84**, 523 (2009).

[79] M. Bastian, S. Heymann, and M. Jacomy, *Gephi: an open source software for exploring and manipulating networks,* in *Third international AAAI conference on weblogs and social media* (2009).

[80] M. F. Porter, *An algorithm for suffix stripping,* Program **14**, 130 (1980).

[81] Y. Kalfoglou and M. Schorlemmer, *If-map: An ontology-mapping method based on information-flow theory,* in *Journal on data semantics I* (Springer, 2003) pp. 98–127.

[82] M. Schorlemmer, Y. Kalfoglou, and M. Atencia, *A formal foundation for ontology-alignment interaction models,* International Journal on Semantic Web and Information Systems (IJSWIS) **3**, 50 (2007).

[83] R. Akkiraju, J. Farrell, J. A. Miller, M. Nagarajan, A. P. Sheth, and K. Verma, *Web service semantics-wsdl-s,* (2005).

[84] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne, *et al., Owl-s: Semantic markup for web services,* W3C member submission **22** (2004).

[85] E. Sirin, B. Parsia, and J. Hendler, *Filtering and selecting semantic web services with interactive composition techniques,* IEEE Intelligent Systems **19**, 42 (2004).

[86] A. Fellah, M. Malki, and A. Elçi, *Web services matchmaking based on a partial ontology alignment,* Int. Journal of Information Technology and Computer Science (IJITCS) **8** (2016).

[87] M. Weidlich, T. Sagi, H. Leopold, A. Gal, and J. Mendling, *Predicting the quality of process model matching,* in *Business Process Management* (Springer, 2013) pp. 203–210.

[88] A. Gater, D. Grigori, and M. Bouzeghoub, *Complex mapping discovery for semantic process model alignment,* in *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services* (ACM, 2010) pp. 317–324.

[89] Y. Belhoul, M. Haddad, A. Gater, D. Grigori, H. Kheddouci, and M. Bouzeghoub, *Spectral graph approach for process model matchmaking,* in *2013 IEEE International Conference on Services Computing* (IEEE, 2013) pp. 408–415.

[90] R. Dijkman, M. Dumas, L. Garcia-Banuelos, and R. Kaarik, *Aligning business process models,* in *2009 IEEE International Enterprise Distributed Object Computing Conference* (IEEE, 2009) pp. 45–53.

[91] J. J. Jung, *Semantic business process integration based on ontology alignment,* Expert Systems with Applications **36**, 11013 (2009).

[92] J. Fengel, *Semantic technologies for aligning heterogeneous business process models,* Business Process Management Journal **20**, 549 (2014).

[93] B. G. Humm and J. Fengel, *Semantics-based business process model similarity,* in *International Conference on Business Information Systems* (Springer, 2012) pp. 36–47.

[94] N. Silva, J. Rocha, and J. Cardoso, *E-business interoperability through ontology semantic mapping,* in *Working Conference on Virtual Enterprises* (Springer, 2003) pp. 315–322.

[95] C. Klinkmüller, I. Weber, J. Mendling, H. Leopold, and A. Ludwig, *Increasing recall of process model matching by improved activity label matching,* in *Business process management* (Springer, 2013) pp. 211–218.

[96] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh, *Ontology alignment for linked open data,* in *International semantic web conference* (Springer, 2010) pp. 402–417.

[97] H. L. Kim, A. Passant, J. G. Breslin, S. Scerri, and S. Decker, *Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces,* in *2008 IEEE International Conference on Semantic Computing* (IEEE, 2008) pp. 315–322.

[98] M. Niepert, C. Meilicke, and H. Stuckenschmidt, *A probabilistic-logical framework for ontology matching,* in *Twenty-Fourth AAAI Conference on Artificial Intelligence* (2010).

[99] M. Kolli and Z. Boufaida, *A description logics formalization for the ontology matching,* Procedia Computer Science **3**, 29 (2011).

[100] H. Nottelmann and U. Straccia, *A probabilistic, logic-based framework for automated web directory alignment,* in *Soft Computing in Ontologies and Semantic Web* (Springer, 2006) pp. 47–77.

[101] E. Jiménez-Ruiz and B. C. Grau, *Logmap: Logic-based and scalable ontology matching,* in *International Semantic Web Conference* (Springer, 2011) pp. 273–288.

[102] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, *Learning to map between ontologies on the semantic web,* in *Proceedings of the 11th international conference on World Wide Web* (AcM, 2002) pp. 662–673.

[103] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, *Ontology matching: A machine learning approach,* in *Handbook on ontologies* (Springer, 2004) pp. 385–403.

[104] M. Rubiolo, M. L. Caliusco, G. Stegmayer, M. Coronel, and M. G. Fabrizi, *Knowledge discovery through ontology matching: An approach based on an artificial neural network model,* Information Sciences **194**, 107 (2012).

[105] M. Mao, Y. Peng, and M. Spring, *An adaptive ontology mapping approach with neural network based constraint satisfaction,* Web Semantics: Science, Services and Agents on the World Wide Web **8**, 14 (2010).

[106] J. Huang, J. Dang, J. M. Vidal, and M. N. Huhns, *Ontology matching using an artificial neural network to learn weights,* in *IJCAI workshop on semantic Web for collaborative knowledge acquisition*, Vol. 106 (2007).

[107] Z. Dragisic, V. Ivanova, H. Li, and P. Lambrix, *Experiences from the anatomy track in the ontology alignment evaluation initiative,* Journal of biomedical semantics **8**, 56 (2017).

[108] I. Harrow, E. Jiménez-Ruiz, A. Splendiani, M. Romacker, P. Woollard, S. Markel, Y. Alam-Faruque, M. Koch, J. Malone, and A. Waaler, *Matching disease and phenotype ontologies in the ontology alignment evaluation initiative,* Journal of biomedical semantics **8**, 55 (2017).

[109] E. Jiménez-Ruiz, A. Agibetov, M. Samwald, and V. Cross, *Breaking-down the ontology alignment task with a lexical index and neural embeddings,* arXiv preprint arXiv:1805.12402 (2018).

[110] E. Jiménez-Ruiz, C. Meilicke, B. C. Grau, and I. Horrocks, *Evaluating mapping repair systems with large biomedical ontologies.* Description Logics **13**, 246 (2013).

[111] E. Jiménez-Ruiz, A. Agibetov, M. Samwald, and V. Cross, *Breaking-down the ontology alignment task with a lexical index and neural embeddings,* arXiv preprint arXiv:1805.12402 (2018).

[112] O. Bodenreider, *The unified medical language system (umls): integrating biomedical terminology,* Nucleic acids research **32**, D267 (2004).

[113] N. F. Noy and M. A. Musen, *The prompt suite: interactive tools for ontology merging and mapping,* International journal of human-computer studies **59**, 983 (2003).

[114] D. Faria, B. S. Balasubramani, V. R. Shivaprabhu, I. Mott, C. Pesquita, F. M. Couto, and I. F. Cruz, *Results of aml in oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 122.

[115] D. Hsiehchen, M. Espinoza, and A. Hsieh, *Multinational teams and diseconomies of scale in collaborative research,* Science advances **1**, e1500211 (2015).

[116] A. Van Raan, *The influence of international collaboration on the impact of research results: Some simple mathematical considerations concerning the role of self-citations,* Scientometrics **42**, 423 (1998).

[117] G. Radda, *Biomedical research and international collaboration,* Science **295**, 445 (2002).

[118] L. Georghiou, *Global cooperation in research,* Research policy **27**, 611 (1998).

[119] S. Fortunato, *Community detection in graphs,* Physics Reports **486**, 75 (2010).

[120] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks,* Journal of Statistical Mechanics: Theory and Experiment **2008**, P10008 (2008).

[121] S. Emmons, S. Kobourov, M. Gallant, and K. Börner, *Analysis of network clustering algorithms and cluster quality metrics at scale,* PloS one **11**, e0159161 (2016).

[122] P. Chopade and J. Zhan, *A framework for community detection in large networks using game-theoretic modeling,* IEEE Transactions on Big Data **3**, 276 (2016).

[123] C. Wohlin, A. Aurum, L. Angelis, L. Phillips, Y. Dittrich, T. Gorschek, H. Grahn, K. Henningsson, S. Kagstrom, G. Low, *et al.*, *The success factors powering industry-academia collaboration,* IEEE software **29**, 67 (2011).

[124] O. Al-Tabbaa and S. Ankrah, *'engineered'university-industry collaboration: A social capital perspective,* European Management Review (2018).

[125] Y. Kalfoglou and M. Schorlemmer, *Ontology mapping: the state of the art,* The knowledge engineering review **18**, 1 (2003).

[126] Y. Kalfoglou and M. Schorlemmer, *If-map: An ontology-mapping method based on information-flow theory,* in *Journal on data semantics I* (Springer, 2003) pp. 98–127.

[127] Y. Kalfoglou and M. Schorlemmer, *If-map: An ontology-mapping method based on information-flow theory,* in *Journal on data semantics I* (Springer, 2003) pp. 98–127.

[128] P. Mitra, N. F. Noy, and A. R. Jaiswal, *Omen: A probabilistic ontology mapping tool,* in *International Semantic Web Conference* (Springer, 2005) pp. 537–547.

[129] J. Tang, J. Li, B. Liang, X. Huang, Y. Li, and K. Wang, *Using bayesian decision for ontology mapping,* Web Semantics: Science, Services and Agents on the World Wide Web **4**, 243 (2006).

[130] J. Li, J. Tang, Y. Li, and Q. Luo, *Rimom: A dynamic multistrategy ontology alignment framework,* IEEE Transactions on Knowledge and data Engineering **21**, 1218 (2008).

[131] M. Ehrig, *Ontology alignment: bridging the semantic gap*, Vol. 4 (Springer Science & Business Media, 2006).

3

SIMULATED ANNEALING-BASED ONTOLOGY MATCHING

Scientific knowledge is in perpetual evolution; it finds itself changed from one day to the next.

Jean Piaget

Ontology alignment is a fundamental task to reconcile the heterogeneity among various information systems using distinct information sources. The evolutionary algorithms (EAs) have been already considered in the literature as the primary strategy to develop an ontology alignment system. However, such systems have two significant drawbacks: They either need a reference alignment which is often unavailable, or they utilize the population-based EAs in a way that they require massive computations and memory. This chapter presents a new ontology alignment system, called SANOM, which uses the well-known simulated annealing as the principal technique to find the mappings between two given ontologies when no reference alignment is available. In contrast to population-based EAs, the simulated annealing need not generate populations, which makes it significantly swift and memory-efficient for the ontology alignment problem. In this chapter, we model the ontology alignment problem as optimizing the fitness of a state whose optimum is computed via the simulated annealing. An alignment fitness function is developed which takes advantages of various similarity metrics including string, linguistic, and structural. A randomized warm initialization is especially-tailored for the simulated annealing in order to expedite its convergence. The experiments illustrate that SANOM is competitive with the state-of-the-art systems, and is significantly superior to other EA-based systems.

3.1. INTRODUCTION

This chapter is dedicated to developing a new alignment system based on simulated annealing. In general, alignment systems based on the evolutionary algorithms (EAs) have several advantages. First, they can be used as a general framework whose objective function can be simply modified for different alignment problems, that enable these algorithms to be applicable to ontologies from different domains. For logistics, which is the domain of study of this dissertation, we can experiment different similarity metrics and select the most appropriate ones for usage. In addition, EAs are often amenable to parallel computing that can be applied to ontology alignment as well. Therefore, they are able to handle large-scale ontology alignment problems in a timely manner and can keep up with the progress of computer architecture. Furthermore, these algorithms can be terminated anytime and yet they provide a near-optimal solution. The ability to be parallel and providing near-optimal solution any time the algorithm is halted are particularly essential for aligning ontologies in logistics IT systems, where there are often time and resource restrictions.

In this chapter, the simulated annealing (SA) is used as the primary strategy to find the alignment between two ontologies in question. SA has several salient features which makes it practically more efficient than other evolutionary techniques. SA mimics the slow cooling in metallurgy in a way that it decreases slowly a temperature value which is high at the beginning of the process. When the temperature is high, the probability to transition to a worse state (based on the fitness function) is higher. As the temperature decreases, the odds of moving to a worse solution diminishes as well. Accepting a worse solution at the beginning would help explore the whole solution space so that the chance of the premature convergence significantly falls. Along with its convergence, SA is more time- and memory-efficient than the population-based EAs since it only operates on one single state based on which it will produce a *successor*. Therefore, it requires less memory to store the populations as well as less time for computing the fitness of multiple chromosomes in a population.

Aside from the inherent characteristics of SA, there are several other advantages of the system proposed in this chapter, SANOM[1]. In contrast to MapPSO, SANOM performs a complete pre-processing step which is proved to enhance significantly the performance of matching [1]. Further, it benefits from the so-called Soft TF-IDF (term-frequency and inverse document frequency) string metric [2] and generalizes it with two base similarity metrics. One of the string similarity metrics is Jaro-Winkler to compare the names solely, and the second is a WordNet-based metric to gauge the linguistic proximity of tokens. The proposed Soft TF-IDF is able to detect the correspondences whose parts of names have been stated by different but synonymous tokens. For matching properties, SANOM will use the notion of the *core* concept, defined in [3], as an extra name for the given properties. This would increase the likelihood of mapping while the false positive decreases as well.

Among EA-based ontology alignment systems, MapPSO is the only one participated in the OAEI and its implementation is also freely available[2]. Thus, we particularly compare SANOM with MapPSO in terms of the execution time as well as efficiency gauged

[1] Stands for **S**imulated **AN**nealing-based **O**ntology **M**atching.
[2] https://sourceforge.net/projects/mappso/

by various performance metrics.

The contributions of this chapter can be summarized as follows:

- An alignment is modeled as a state whose optimum based on a fitness function will solve the ontology matching problem;
- An intrinsic fitness function is developed by using various similarity measures. In this regard, the Soft TF-IDF metric is extended by using two base similarity metrics; one for the strings similarity of tokens and one for their linguistic relations;
- The simulated annealing is adjusted to find the alignment between two given ontologies. In this regard, a randomized greedy algorithm is developed for the initialization which expedites the convergence of the algorithm;
- The proposed system is evaluated with the OAEI anatomy, conference, and disease and phenotype tracks.

The preliminary implementation of SANOM participated in the OAEI 2017 [4], and the current, enhanced implementation participated in the OAEI 2017.5 [5] and 2018 [6].

The chapter is structured as follows. Section 3.2 dedicates to the basic concepts of simulated annealing. The computation of the alignment fitness using string and structural similarity metrics are discussed in Section 3.3. Section 3.4 contains the details of the proposed system, SANOM, and the experimental results are presented in Section 3.5. Finally, the chapter is concluded in Section 3.6.

3.2. SIMULATED ANNEALING

Simulated annealing is a probabilistic approach to estimate the global optimum of problems which cannot be solved by the standard optimization techniques [7]. As the name suggests, this technique simulates the annealing in metallurgy which slowly cools the materials to decrease their defects. Such a controlled cooling is implemented in the simulated annealing as the probability to transition to a worse solution. The probability of a move to a worse solution is proportionate to the temperature: The higher the temperature, the more chance to move to a worse solution. Such a feature would enable SA to explore the whole search space when the temperature is high, making it not converge prematurely unlike the genetic and swarm intelligence algorithms.

In contrast to population-based EAs, SA only operates on one possible solution, called *state*, and tries to improve it to get a better solution. Such an enhancement is performed by creating a new successor in the neighborhood of the current state, and then probabilistically transition to it. Let S be the current state and S' be the *successor* (or the neighbor) created based on the current state. The proposed move from S to S' happens based on a fitness function: If the fitness of S' is superior to S, then the transition *certainly* happens, and it *probably* occurs otherwise.

The probability of a move when the fitness of the successor is less than the current state is commensurate with the value of their fitness and the temperature. In more detail, let $f(S)$ and $f(S')$ be the fitness of the current and successor states, respectively. If $f(S') > f(S)$, then the transition to *successor* certainly happens. Otherwise, the likelihood of its

occurrence is $e^{\frac{\Delta E}{T}}$, where $\Delta E = f(S') - f(S)$. Therefore, the probability of moving to S', shown by P_{move}, can be rewritten as:

$$P_{move} = min\left(e^{\frac{\Delta E}{T}}, 1\right), \tag{3.1}$$

where T is the temperature. It is evident from equation (3.1) that if $f(S') > f(S)$, then $e^{\Delta E/T} > 1$ and $P_{move} = 1$. Thus, the proposed move to S' will certainly happen. Otherwise, the transition is reliant on ΔE and T: The greater ΔE or smaller T, the smaller chance to accept the move to a state with lower fitness.

Having generated the successor and having then computed the probability of accepting the move using equation (3.1), the move is accepted or rejected in practice by sampling from a uniform distribution in the interval $[0,1]$. If the sampled value is less than P_{move}, then the move to S' is accepted. Otherwise, the transition is rejected, and a new successor is produced based on the current state.

The temperature also plays a critical role in the transition to a worse state. The simulated annealing algorithm starts with a higher temperature so that the transition to worse states are more likely at the beginning. However, the temperature is getting lower and lower as the time goes by, making it less likely to move to a worse solution. This enables SA to explore the whole solution space at the beginning to find the global optima and to prevent premature convergence. The overall SA algorithm is summarized in Algorithm 1.

Algorithm 1 Simulated annealing

Input: $S = S_0$, maxIter
for iter=1:maxIter **do**
  $T = updateTemperature(iter, maxIter)$
  $S' = generateSuccessor(S)$
  Compute $P_{move} = Pr(S, S', T')$ by equation (3.1)
  Sample r from the uniform distribution in the interval $[0,1]$
  if $P_{move} > r$ **then**
    $S = S'$
  end if
end for
Output Final state S

3.3. ALIGNMENT FITNESS

At the heart of any evolutionary algorithm, there must be a way to measure the fitness of different solutions based on which the evolution happens. First, a precise definition is presented for the alignment fitness.

Definition 4 (Alignment fitness) *Given an alignment A between two ontologies O_1 and O_2, the fitness of A is computed by the function F: $A^* \rightarrow R$ (where A^* is the set of all possible*

alignments), and is defined as

$$F(A) = \sum_{c \in A} f(c),$$

where $f : A \rightarrow R$ computes the fitness of each correspondence in the given alignment A.

The alignment fitness definition reveals the need of computing the fitness of each correspondence for having the overall fitness of a given alignment. The function f calculates the similarity of two entities in correspondence c.

The first way of calculating the similarity is to consider the names (e.g., URI, label, comments, etc.) of two given entities and determine their sameness using either string similarity metrics or their linguistic relations using WordNet [8]. The ways of finding the similarity between two classes and two properties are different in the proposed system. It is also possible to consider the positions of two entities in their ontologies as a meter of similarity. For instance, if two classes match from two ontologies, the likelihood of mapping their subclasses increases. Such metrics are referred to as structural similarity metrics. In a nutshell, $f(c) = f_{string}(c) + f_{strucutural}(c)$, where $f_{string}(c)$ and $f_{strucutural}(c)$ are the string and structural similarity measures, respectively. In the remainder of this section, the appropriate similarity metrics which are utilized in SANOM are reviewed.

3.3.1. STRING SIMILARITY METRIC

In this section, the techniques for computing the similarity between the strings of two entities are revised. We take advantage of the Soft TF-IDF (term frequency-inverse document frequency) with two base similarity metrics. The reason of using this metric is that it can be generalized to accommodate multiple base similarity metrics. There are some other metrics such as Soft Jaccard which have the same capability, but Soft TF-IDF has shown better performance in terms of both precision and recall in recent studies [1]. The base metrics for Soft TF-IDF are Jaro-Winkler, to deal with names as a sequence of characters, as well as Wu and Palmer [9], to compute the linguistic relatedness of two names using WordNet.

It is common that ontologies have several annotations which facilitate finding their peers in the other ontologies. Further, the sole comparison of properties names would lead to poor results. Therefore, we consider a set of names for each entity as follows:

- For classes, an essential name is their uniform resource identifier (URI). Besides, there are some annotations which might help the matching process. Among them are *label* and *comment* annotations which provide more information about the corresponding class. There are sometimes related synonyms for classes in an ontology, which should also be considered. For instance, the OAEI anatomy track has several related synonyms which would enhance the alignment outcome.

- For the property alignment, considering solely the names would mislead the matching process. As recommended by M. Cheatham and P. Hitzler [3], we consider the *core* concept as another name for each property. The core is the first verb, if exists, in the property name whose length is higher than three, otherwise the first noun along with its corresponding adjective. The Stanford part of speech tagger [10] is utilized to extract the core of each property.

Evidently, it is likely that each entity has more than one name, we therefore take the maximum similarity among various names as the string similarity among two corresponding entities. Let G and H be sets of names pertaining to two entities $e_1 \in O$ and $e'_1 \in O'$, then

$$f_{string}(c) = Sim(G,H) = \max_{g \in G, h \in H} \text{Soft TF-IDF(g,h)}, \tag{3.2}$$

where c represents a correspondence containing the mapping e_1 to e'_1, $Sim(G,H)$ is the similarity between the names of two entities, and Soft TF-IDF denotes the string similarity measure. In the further subsections, the Soft TF-IDF is explained. Prior to that, we need to use several pre-processing strategies to increase the chance of matching.

PRE-PROCESSING STRATEGIES

The modification of strings before the similarity computation is essential to increase the chance of mapping entities. The primary pre-processing strategies utilized in SANOM are:

- **Tokenization.** The terminology of concepts is usually constructed by a sequence of words. The words are often concatenated by white space, the capital letter of first letters, and several punctuations such as " – " or "_". Therefore, the initial strings can be broken into a bag of words, which is called *tokenization*.
- **Stop word removal.** Stop words refer to the common words which do not convey any particular meaning. The stop words can be distinguished by looking up the tokens (identified after tokenization) in a table storing the potential stop words. The Glasgow stop word list, which contains English stop words, is utilized in the current implementation[3].
- **Stemming.** Entities may refer to the same notion, but they may appear differently due to various verb tenses, plural/singular, and so forth. Therefore, we need to revert them to a standard form to be able to detect similar concepts which have been changed for the grammatical reasons. The Porter stemming method is used for this matter [11].

SOFT TF-IDF WITH MULTIPLE BASE SIMILARITY METRICS

TF-IDF, or cosine similarity, is one of the most popular strategies used in information retrieval [2]. To calculate TF-IDF, we need to compute the frequency of word w in bag of tokens S, e.g., $TF_{w,S}$, and the inverse fraction of strings which contain w, e.g., IDF_w. Then, TF-IDF of two given sets G and H is computed as:

$$\text{TF-IDF}(G,H) = \sum_{w \in G \cap H} V(w,s)V(w,T), \tag{3.3}$$

where V is defined as:

$$V(w,H) = \frac{log(TF_{w,H}+1).log(IDF_w)}{\sum_{w'} log(TF_{w',H}+1).log(IDF'_w)}. \tag{3.4}$$

[3] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

Equation (3.3) only considers the words which are exactly *identical* in both bags of words. However, *identity* can be interpreted differently, especially for ontology alignment. Thus, TF-IDF can be extended by defining the identity using a base string similarity measure. Given such a similarity metric and a threshold, set C is defined as the set of triples (g, h, sim), where $g \in G$ and $h \in H$ are tokens whose similarity sim is computed by the base similarity metric (and is greater than a given threshold). Having the set C, the Soft TF-IDF is defined as

$$\text{Soft TF-IDF}(G, H) = \sum_{w \in C} V(w, g) V(w, H) D(w, H), \tag{3.5}$$

where $D(w, H) = max_{v \in C} sim(w, v)$, and $sim(w, v)$ is the similarity of w and v in C. The base similarity metric gauges the similarity of tokens obtained from each name. In this study, we take advantages of two similarity metrics and take their maximum as the final similarity of two given tokens. The reason of considering the maximum similarity is that two tokens are assumed to be similar if their names are nearly identical or they are linguistically related. For instance, *ConferenceDinner* and *ConferenceBanquet* are deemed the same since the first token of two names is identical, and the second token is linguistically similar. The similarity metrics for measuring the strings similarity and lingual relatedness are:

- **Jaro-Winkler metric.** The combination of TF-IDF and Jaro-Winkler has shown promising performance in name entity matching [2] as well as in ontology alignment [1]. By the same token, SANOM exploits Jaro-Winkler with the threshold 0.9 as one of the base similarity metrics. The value of the threshold is in line with recent studies [1].

- **WordNet-based metric.** The linguistic heterogeneity is also prevalent in various domains. Therefore, the existence of a similarity metric to measure the lingual closeness of two entities is absolutely essential. In this study, the relatedness of two given tokens are computed by the Wu and Palmer measure [9] and is used as another base similarity metric. The threshold should be high enough since two distinct tokens are assigned a large similarity value according to Wu and Palmer, which solely checks the semantic relatedness of given tokens. Our investigation showed that any value less than 0.8 would practically mean that most of tokens are semantically related according to Wu and Palmer. Thus, one needs to specify a much higher threshold to avoid it. In the current implementation of SANOM, the threshold for this similarity metric is set to 0.95.

Using these two base similarity metrics, we can discover the concepts with synonymous changes in one or multiple tokens.

3.3.2. STRUCTURAL SIMILARITY

The preceding string similarity metric gives a high score to the entities which have lexical or linguistic proximity. Another similarity of two entities could be derived from their positions in the given ontologies.

We consider the following two structural similarity measures for the current implementation of SANOM:

- The first structural similarity is gauged by the subsumption relation of classes. If there are two classes C and C' whose superclasses are SC and SC' from two given ontologies O and O', then the matching of classes SC and SC' would increase the similarity of C and C'. Let c be a correspondence mapping SC to SC', then the increased similarity of C and C' is gauged by:

$$f_{structural}(C, C') = f(c). \tag{3.6}$$

- Another structural similarity is derived from the properties of the given ontologies. The alignment of two properties would tell us the fact that their corresponding domain and/or ranges are also identical. Similarly, if two properties have the analogous domain and/or range, then it is likely that they are similar as well.

  The names of properties and even their corresponding core concepts are not a reliable meter based on which they are declared a correspondence. A recent study has shown that the mapping of properties solely based on their names would result in high false positive and false negative rates [3], e.g., there are properties with identical names which are not semantically related, while there are semantically relevant properties with totally distinct names.

  The current implementation treats the object and data properties differently. For the object properties op_1 and op_2, their corresponding domains and ranges are computed as the concatenation of their set of ranges and domains, respectively. Then, the fitness of the names, domains, and ranges are computed by Soft TF-IDF. The final mapping of two properties is the average of top two fitness scores obtained by Soft TF-IDF. For the data properties, the fitness is computed as the similarity average of names and their corresponding domain.

  On the other flow of alignment, it is possible to derive if two classes are identical based on the properties. Let e_1 and e_2 be classes, op_1 and op_2 be the object properties, and R_1 and R_2 are the corresponding ranges, then correspondence $c = (e_1, e_2)$ is evaluated as

$$f_{structural}(c) = \frac{f_{string}(R_1, R_2) + f_{string}(op_1, op_2)}{2}. \tag{3.7}$$

3.4. ONTOLOGY ALIGNMENT USING SIMULATED ANNEALING

Having computed the alignment fitness (see Definition 4), the simulated annealing can be exploited to find the best possible alignment of two given ontologies. In this section, the necessary steps of the simulated annealing are described to solve the ontology alignment problem. Figure 3.1 displays the details of the SANOM implementation. In the following, we explain the modules of SANOM depicted in Figure 3.1.

3.4.1. ONTOLOGY PARSING AND SIMILARITY COMPUTATION

According to Figure 3.1, ontologies are first parsed using OWL API [12], and they are stored in a list of a data structure called *lexicon*. Lexicon is the main data structure of SANOM, which contains a list of hash sets for each concept in the ontology. This list

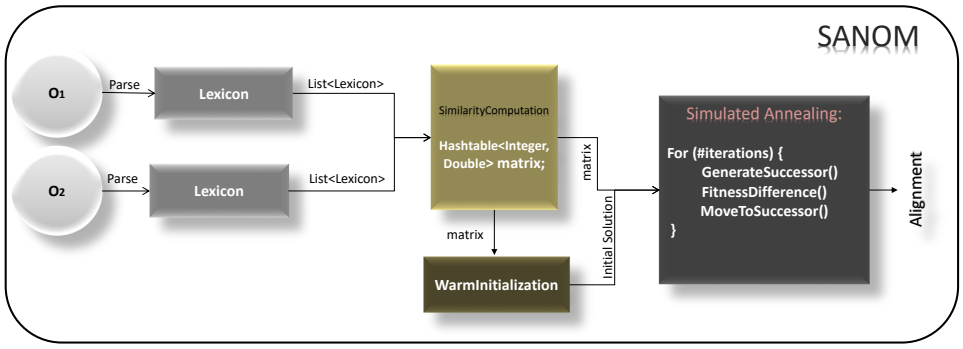


Figure 3.1: The architecture of SANOM.

contains all the names of a concept, e.g., URI and labels, which are tokenized and undergone the pre-processing strategies before being stored. Thus, a hash set in the list contains the tokens of a name of a concept from an ontology. The overall concepts of each ontology are stored in a list of lexicon. The list index is considered as the index of the corresponding concept in the ontology, which will be used to store the similarity of concepts.

After parsing the ontologies, we first need to compute the similarity of each concept from the first ontology to all concepts from the target. As is recently recommended [13], the similarity computations should be efficiently stored in a hash table. Further, we only store the similarities whose magnitude is bigger than a value. In the current implementation, this value is set to 0.5 [13]. The keys of the hash table can be simply generated based on the index of the concept of the first ontology to that in the target. Using hash tables decrease significantly the amount of the memory required for saving the similarities for large-scale ontologies.

Computing the similarity of entities is time-consuming due to the use of a nested loop, making us use the Java Fork/Join framework [14] to expedite the procedure. The Java Fork/Join framework uses the divide and conquer strategy so that it divides the initial big tasks into several small ones (the fork step) and then solves the smaller tasks. At the end, the small tasks are aggregated together (the join step). For computing the similarity of a concept from the first ontology, we divide the concepts of the second ontology into different disjoint partitions, and compute the similarities of concepts of the first ontology with those in smaller tasks. Having computed the similarities, we now look into other elements in SANOM. Prior to that, we need to model an alignment in order to be able to use the SA for ontology alignment.

3.4.2. REPRESENTATION OF AN ALIGNMENT

The state in the simulated annealing would represent an alignment. Therefore, the optimal mappings between ontologies in question are obtained by optimizing the state fitness. Let n and m be the number of concepts of the ontologies in question, the state $S \in R^n$ is an integer vector whose values are between 1 and m, i.e., $1 \leq S_i \leq m$. Hence, if the i^{th} cell contains number j, it indicates that (e_i, e'_j) is the related correspondence.

Definition 5 (Alignment State) *Alignment state between ontologies in question is defined as a set that contains pairs of concepts from two ontologies:*

$$S = \{(e_1, e'_{j_1}), (e_2, e'_{j_2}), ..., (e_n, e'_{j_n})\}, \tag{3.8}$$

where (e_i, e'_{j_i}) *is a correspondence mapping* $e_i \in O$ *to* $e'_{j_i} \in O'$*, and* $j_i \in \{1, 2, ..., m\}$.

Definition 5 formally describes the state in the simulated annealing as an alignment. From the implementation view, the set S can be defined as a vector whose entries would indicate a correspondence. More in detail, the element at the k^{th} position of this vector is the mapping (e_k, e_{j_k}).

In contrary to the population-based evolutionary strategies, the simulated annealing only operates on one single state and tries to evolve it in order to obtain a better solution. Therefore, it is more time- and memory-efficient. The length of S could be more optimal if we choose it as the minimum of m and n, especially if their difference is considerable. However, such an improvement is not significant and is ignored in the current implementation.

3.4.3. WARM INITIALIZATION WITH A RANDOMIZED GREEDY TECHNIQUE

To expedite the convergence of the simulated annealing, SANOM implements a randomized greedy technique for the initialization. An element from the alignment state is arbitrarily chosen by finding a random number r between 1 and n. Then, the entity $e_r \in O$ is mapped to entity $e'_{j_r} \in O'$ which has the maximum similarity, e.g., $arg\max_{e' \in O'} f(e_r, e') = e'_{j_r}$. The similarity of the correspondences $f(e, e_{j_r})$ is stored in a hash table, which can be retrieved immediately so that finding an initial solution is not time-consuming. It is evident that this way of mapping does not result in optimal alignment, but it is significantly better than using an arbitrary initial state.

The mapping is considered to be one-to-one, hence it must be fulfilled in the initialized state as well. Therefore, some auxiliary variables are required to check these constraints. We also need to compute the fitness of the initial solution, since it is required in the SA. The fitness can be simply calculated by adding the fitness of each correspondence we added to the alignment. Algorithm 2 summarizes the whole procedure of finding an efficient initial state.

3.4.4. GENERATING A SUCCESSOR

The simulated annealing finds the optimal solution by the transition to a new state which usually has a higher fitness value. The prerequisite to such a move is to first generate the next state in the neighborhood of the current.

We swap the elements of the current state in order to produce a successor. The number of elements to be swapped can be a fraction of the state length. In the current implementation, we alter q elements of the current state where $q = \lceil 5\% * |S| \rceil$ and $|S|$ is the length of the current state. The alteration happens by finding q distinct numbers between 1 and n, stored in vector k, and then exchanging the elements $s(k(i))$ and $s(k(i+1))$, where $k(i)$ is element i of vector k.

The fitness of the successor can also be computed based on that of the current state. More in detail, when the value at position $k(i)$ is replaced with the value of $k(i+1)$, it

Algorithm 2 Randomized greedy technique for initialization

Input: Set of entities of the source and target ontologies O and O'
$n = |O|$, $m = |O'|$, $counter = 0$, $fit = 0$, S
while counter < n **do**
  r = $generate - random - number(1, n)$
  If(Chosen-Before(r)) **continue;**
  $e'_{j_r} = arg\max_{e' \in O'} f(e_r, e')$
  fit += $f(e_r, e'_{j_r})$
  S(r) = j_r
  Remove(e_{j_r}, O')
  ++counter;
end while
Output State S and its fitness *fit*

means the correspondence $(e_{k(i)}, e'_{S(k(i))})$ is replaced with $(e_{k(i)}, e'_{S(k(i+1))})$, and the correspondence $(e_{k(i+1)}, e'_{S(k(i+1))})$ is substituted with $(e_{k(i+1)}, e'_{S(k(i))})$, for S being a state. As a result, we merely need to subtract the fitness values of the previous correspondences and add those of the new ones. The fitness of these correspondences have been already stored in a hash table. Algorithm 4 summarizes the overall procedure for creating a successor and its fitness. Since the fitness of correspondences are computed in a hash table before running the SA, the creation of a successor and its fitness is quite swift. It only requires to swap $k/2$ elements and conduct $4k/2$ additions/subtractions.

Algorithm 3 Generating a successor and its fitness calculation

Input: State S and its fitness f(S)
$n = |O|$, $m = |O'|$, $S' = S$, $f(S') = f(S)$
$q = \lceil 5\%n \rceil$
k=generate-distinct-number(q,1,n); // generating q distinct number in the interval [1,n]
for i < length(k); k+2 **do**
  swap(S',k(i),k(i+1)); // replacing the elements of S in the positions $k(i)$ and $k(i+1)$
  $f(S')$ -= $f(e_{k(i)}, e_{S(k(i))})$.
  $f(S')$ -= $f(e_{k(i+1)}, e_{S(k(i+1))})$
  $f(S')$ += $f(e_{k(i)}, e_{S(k(i+1))})$
  $f(S')$ += $f(e_{k(i+1)}, e_{S(k(i))})$
end for
Output State S' and its fitness $f(S')$

3.4.5. SANOM IN A NUTSHELL

SANOM first computes the similarity of each entity from the first ontology to the entities from the target. Then, the warm initialization would find a possibly good initial alignment state for the given ontologies. The initial alignment is then enhanced by the

simulated annealing by generating a successor, computing its fitness, and then moving to it. Such an enhancement is recurrently repeated for some number of iterations.

The number of iterations is a parameter which can be tuned by the user. According to the number of iterations, the temperature in each iteration can be tuned. Given the number of iterations $iter_{max}$ and the initial temperature T_{init} ($T_{init} = 1$ by default), the temperature at iteration $iter$, shown as T_iter, can be computed as

$$T_{iter} = (1 - \frac{iter}{iter_{max}}) T_{init}.$$

Having computed the temperature, the overall ontology alignment algorithm is summarized in Algorithm 4. In terms of the time complexity, the randomized greedy initialization is of order $O(nm)$ and it is only executed once, where m and n are the number of entities in the two ontologies to be aligned. The successor generation is quite swift, and it only requires $k/2$ swap operations and $4k/2$ additions/subtractions, where k is the number of elements in an alignment to be swapped. Thus, SA is very swift. However, the most time-consuming module is the similarity computations that are efficiently implemented using Java Fork/Join framework. As a result, the time complexity varies significantly with respect to n, since the string similarity is partly reliant on the length of the concept names (or the number of tokens), and structural similarity is dependent on the number of superclasses/subclasses. However, it is certain that both structural and string similarity metrics are required to be conducted for each pairs of entities from the two input ontologies.

The number of iterations is also proportionate to the number of entities in the input ontologies, since for larger ontologies we need to have a higher number of iterations so that the swap operator can be applied to as many potential correspondences as possible to optimize the overall search for an optimal alignment. Note that the fitness of a successor, which is the most time-consuming module in simulated annealing, can be computed very fast in SANOM, making the overall performance of the system optimal even for a large number of iterations.

3.5. EXPERIMENTAL RESULTS

To evaluate the efficiency and efficacy of the proposed ontology alignment system, several standard benchmarks with known reference alignment are required. In this section, the proposed system is evaluated and compared with the state-of-the-art alignment systems. We take advantages of the benchmarks from three tracks of the ontology alignment evaluation initiative (OAEI), i.e., anatomy, conference, and disease and phenotype tracks, to evaluate the performance of SANOM and compare it with several other alignment systems.

3.5.1. ANATOMY TRACK

SANOM is first applied to the anatomy track and compared with AML [15], XMap [16], LogMap and LogMapLite [17], KEPLER [18], Wiki3 [19], and ALIN [20]. The number of iteration was set to 1,000 for this track. We applied MapPSO, the only EA-based alignment system with available implementation, to the anatomy track, but its outcome was not

Algorithm 4 SANOM

Input: Source and target ontologies O_1 and O_2, number of iteration $iter_{max}$, initial temperature T=1.
Finding the initial state S and its fitness $f(S)$ by Algorithm 2
while $iter < iter_{max}$ **do**
$T_{iter} = (1 - \frac{iter}{iter_{max}}) T_{init}$.
S' and its fitness $f(S')$ are generated by Algorithm 3.
$\Delta E = f(S') - f(S)$.
$P_{move} = min\left(e^{\frac{\Delta E}{t}}, 1\right)$.
if P_{move} > random(0,1) **then**
$S = S'$
$f(S) = f(S')$
end if
end while
Output State S

Table 3.1: The precision, recall, and F-measure of participatory systems on the OAEI anatomy track.

System	Precision	F-measure	Recall
AML	0.95	0.943	0.936
XMap	0.926	0.893	0.863
KEPLER	0.958	0.836	0.741
LogMap	0.918	0.88	0.846
LogMapLite	0.962	0.829	0.728
SANOM	0.888	0.870	0.853
WikiV2	0.883	0.802	0.734
ALIN	0.996	0.506	0.339

acceptable with both precision and recall less than 0.05. Thus, we left it out for comparison on the anatomy track. Among the participating systems, LogMap and LogMapLite, SANOM, and ALIN are the systems that do not take advantages of any background knowledge such as UMLS Metathesaurus [21]. Hence, it is evident that these systems have lower performance with respect to those with biomedical background knowledge.

The systems are first compared based on precision, recall, and F-measure, which are tabulated in Table 3.1. According to this table, AML is the system with the highest discovery, followed by XMap. Both of these systems have utilized biomedical background knowledge which led to the better performance.

Among the system without the background knowledge, SANOM has the best performance with respect to recall which means that the proposed system could discover more correspondences. LogMap, on the other hand, has better performance in terms of precision. The difference between SANOM and LogMap are approximately 1% regarding recall and 3% with respect to precision. The overall difference between SANOM and LogMap is 1% regarding F-measure which is the trade-off between precision and recall. Further, SANOM is interestingly quite competitive with XMap as well; their difference

is nearly 1% in terms of precision and 2% regarding recall in spite of the fact that XMap takes advantages of UMLS, dedicated background knowledge in the biomedical domain.

3.5.2. CONFERENCE TRACK

The experiments in this section include the alignment of seven ontologies from the conference track of the OAEI: *cmt*, *conference*, *confOf*, *iasted*, *edas*, *ekaw*, *sigkdd*. These ontologies describe the conference organization from different proceedings; therefore, they are heterogeneous by nature.

Pairing every two ontologies together, there are overall 21 mapping tasks. The tasks entail, not only the alignment of classes, but the alignment of properties as well. Therefore, it is a suitable challenge to gauge the goodness of alignment systems for matching properties. Our experience over this track showed that SANOM converges to the optimal solution with less than 100 iteration in all tracks. Thus, the number of iterations was set to 100.

Along with the systems used for comparison in the anatomy track, MapPSO is also included for comparison, since it shows better performance in this track. The overall performance of alignment systems is also gauged by micro- and macro-averaging. Macro-averaging is the mean of performance scores over all tasks. For the micro-averaging, on the other hand, the false positive (FP), false negative (FN), and true positive (TP) in each task are first computed, and then the micro-averages of the precision and recall are defined as:

$$\hat{Precision} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i}, \qquad \hat{Recall} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FN_i}, \tag{3.9}$$

where TP_i, FN_i, and FP_i are respectively the true positive, false negative, and false positive for the i^{th} task and $\hat{Precision}$ and $\hat{Recall}$ are respectively the precision and recall micro average.

Table 3.2 tabulates the precision, recall, and F-measure of various systems along with the micro and macro averages. In terms of recall, the proposed system has the best performance by the margin of 6% from AML, the second-best performing system, and by the margin of 14% from XMap and LogMap. Regarding precision, XMap, LogMap, and AML have better performance compared to SANOM. It is usually the case that the true positive increases at the expense of more false positives. Concerning F-measure, however, SANOM is superior to those of LogMap and XMap and is quite competitive with AML which has been the best performing alignment system in this track.

We finally compare SANOM and MapPSO, two alignment systems based on the evolutionary algorithm, in terms of the execution time over a computer with a CPU core-i5 and $4GB$ RAM. Table 3.3 shows the execution time in seconds of both systems over the mapping tasks in the conference track. It is evident that SANOM is significantly faster than MapPSO. The overall time required for MapPSO to complete all the tasks is approximately 747 seconds, while SANOM completes them in about 58 seconds. Therefore, SANOM is not only superior from precision and recall metrics but is also remarkably swift with respect to MapPSO.

Table 3.2: The performance scores of the systems over the tasks of the conference track. The metrics for comparison are precision (P), recall (R), and F-measure (F). The overall performance of each system over all tasks is gauged by the micro- and macro- averaging.

	SANOM			AML			LogMap			XMap			LogMapLite			KEPLER			ALIN			Wiki3			MapPSO		
	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R
cmt-conference	0.61	0.74	0.93	0.67	0.59	0.53	0.73	0.62	0.53	0.73	0.62	0.53	0.56	0.42	0.33	0.53	0.56	0.60	1.00	0.42	0.27	0.38	0.36	0.33	0.05	0.09	0.27
cmt-confOf	0.80	0.62	0.50	0.90	0.69	0.56	0.83	0.45	0.31	0.86	0.52	0.38	0.67	0.48	0.38	0.55	0.44	0.38	1.00	0.22	0.13	0.63	0.42	0.31	0.07	0.10	0.19
cmt-edas	0.63	0.69	0.77	0.90	0.78	0.69	0.89	0.73	0.62	0.75	0.72	0.69	0.73	0.67	0.62	0.69	0.69	0.69	1.00	0.47	0.31	0.73	0.67	0.62	0.08	0.13	0.38
cmt-ekaw	0.54	0.58	0.64	0.75	0.63	0.55	0.75	0.63	0.55	1.00	0.71	0.55	0.56	0.50	0.45	0.55	0.55	0.55	1.00	0.43	0.27	0.71	0.56	0.45	0.09	0.15	0.45
cmt-iasted	0.67	0.80	1.00	0.80	0.89	1.00	0.80	0.89	1.00	0.80	0.89	1.00	0.80	0.89	1.00	0.50	0.67	1.00	1.00	0.67	0.50	0.67	0.80	1.00	0.04	0.07	0.50
cmt-sigkdd	0.85	0.88	0.92	0.92	0.92	0.92	1.00	0.91	0.83	1.00	0.91	0.83	0.89	0.76	0.67	0.77	0.80	0.83	1.00	0.50	0.33	0.80	0.73	0.67	0.19	0.31	0.75
conference-confOf	0.79	0.76	0.73	0.87	0.87	0.87	0.85	0.79	0.73	0.79	0.76	0.73	0.90	0.72	0.60	0.56	0.58	0.60	0.83	0.48	0.33	0.73	0.62	0.53	0.15	0.23	0.53
conference-edas	0.67	0.74	0.82	0.73	0.69	0.65	0.85	0.73	0.65	0.83	0.69	0.59	0.75	0.62	0.53	0.48	0.53	0.59	0.83	0.43	0.29	0.64	0.58	0.53	0.02	0.03	0.06
conference-ekaw	0.66	0.70	0.76	0.78	0.75	0.72	0.63	0.55	0.48	0.65	0.52	0.44	0.62	0.42	0.32	0.52	0.50	0.48	0.75	0.36	0.24	0.64	0.46	0.36	0.09	0.13	0.28
conference-iasted	0.88	0.64	0.50	0.83	0.50	0.36	0.88	0.64	0.50	0.83	0.50	0.36	0.80	0.42	0.29	0.63	0.45	0.36	1.00	0.35	0.21	0.67	0.40	0.29	0.03	0.06	0.21
conference-sigkdd	0.75	0.77	0.80	0.85	0.79	0.73	0.85	0.79	0.73	0.82	0.69	0.60	0.80	0.64	0.53	0.71	0.69	0.67	0.86	0.55	0.40	0.67	0.59	0.53	0.09	0.15	0.40
confOf-edas	0.82	0.78	0.74	0.92	0.71	0.58	0.77	0.63	0.53	0.91	0.67	0.53	0.58	0.58	0.58	0.45	0.49	0.53	0.83	0.40	0.26	0.50	0.49	0.47	0.10	0.15	0.32
confOf-ekaw	0.81	0.83	0.85	0.94	0.86	0.80	0.93	0.80	0.70	0.81	0.72	0.65	0.77	0.61	0.50	0.62	0.63	0.65	1.00	0.46	0.30	0.73	0.52	0.40	0.26	0.35	0.55
confOf-iasted	0.71	0.63	0.56	0.80	0.57	0.44	1.00	0.62	0.44	0.80	0.57	0.44	1.00	0.62	0.44	0.36	0.40	0.44	1.00	0.36	0.22	0.57	0.50	0.44	0.08	0.14	0.44
confOf-sigkdd	0.83	0.77	0.71	1.00	0.92	0.86	1.00	0.83	0.71	1.00	0.73	0.57	1.00	0.73	0.57	0.80	0.67	0.57	1.00	0.44	0.29	0.80	0.67	0.57	0.06	0.11	0.43
edas-ekaw	0.71	0.72	0.74	0.79	0.59	0.48	0.75	0.62	0.52	0.79	0.59	0.48	0.59	0.50	0.43	0.65	0.60	0.57	0.63	0.32	0.22	0.63	0.51	0.43	0.04	0.07	0.17
edas-iasted	0.69	0.56	0.47	0.82	0.60	0.47	0.88	0.52	0.37	0.88	0.52	0.37	0.78	0.50	0.37	0.64	0.47	0.37	0.75	0.26	0.16	0.80	0.55	0.42	0.03	0.05	0.16
edas-sigkdd	0.80	0.64	0.53	1.00	0.80	0.67	0.88	0.61	0.47	1.00	0.70	0.53	0.88	0.61	0.47	0.88	0.61	0.47	1.00	0.33	0.20	0.78	0.58	0.47	0.07	0.11	0.27
ekaw-iasted	0.70	0.70	0.70	0.88	0.78	0.70	0.75	0.67	0.60	1.00	0.82	0.70	0.60	0.60	0.60	0.54	0.61	0.70	1.00	0.46	0.30	0.75	0.67	0.60	0.01	0.02	0.10
ekaw-sigkdd	0.89	0.80	0.73	0.80	0.76	0.73	0.86	0.67	0.55	0.88	0.74	0.64	0.88	0.74	0.64	0.78	0.70	0.64	1.00	0.53	0.36	0.88	0.74	0.64	0.05	0.08	0.27
iasted-sigkdd	0.70	0.80	0.93	0.81	0.84	0.87	0.71	0.69	0.67	0.87	0.87	0.87	0.73	0.73	0.73	0.59	0.70	0.87	1.00	0.57	0.40	0.72	0.79	0.87	0.05	0.08	0.20
Macro-Averaging	0.74	0.72	0.73	0.84	0.74	0.67	0.84	0.68	0.59	0.86	0.69	0.59	0.76	0.61	0.53	0.61	0.59	0.60	0.93	0.43	0.29	0.69	0.58	0.52	0.08	0.12	0.33
Micro-Averaging	0.72	0.72	0.72	0.84	0.74	0.66	0.82	0.67	0.57	0.84	0.68	0.57	0.72	0.59	0.50	0.59	0.58	0.57	0.89	0.42	0.27	0.67	0.57	0.49	0.07	0.11	0.31

Table 3.3: The consumed time for MapPSO [22] and SANOM to produce an alignment for each of the tasks in the conference track. The times are in seconds, and the number of each task corresponds to mapping two ontologies displayed in Table 3.2.

Task	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
MapPSO	32.0	26.7	27.2	27.1	32.0	24.5	30.0	40.6	39.3	53.3	34.6	32.1	24.2	2.4	29.6	50.1	78.6	30.8	46.7	32.3	32.0
SANOM	9.4	0.9	1.8	1.2	1.5	1.7	1.7	3.2	2.2	3.4	1.6	2.0	2.3	2.3	1.1	3.0	4.9	2.8	4.3	2.3	4.3

3.5.3. DISEASE AND PHENOTYPE TRACK

SANOM is further applied to the OAEI disease and phenotype track [23], which consists of matching various disease and phenotype ontologies. In particular, we consider the mapping of the human phenotype (HP) to the mammalian phenotype (MP), and aligning the human disease ontology (DOID) and the orphanet and rare diseases ontology (ORDO).

The ontologies in this track contain approximately 15,000 concepts; therefore, the alignment of these ontologies is challenging. Faria et al. [13] investigated the challenges of large biomedical ontologies, and they recommended several ways of dealing with these ontologies. Some of these recommendations, such as using hash tables for storing the similarity of entities, have been used in SANOM, which makes it possible to align even ontologies with this size. MapPSO could not find the alignment of ontologies in this track since it requires a massive amount of memory space. Thus, it cannot be compared with SANOM on this track. For the reference, a *voted* reference alignment has been used which was created based on the outputs of the alignment systems participated in this track for the last three years. A reasoner was also used in order to validate the final alignment.

In comparison to the systems participated in other tracks of the OAEI, fewer systems can generate a reliable alignment for this track. All other participating systems in this track use a biomedical background knowledge. In particular, LogMap uses normalizations and spelling variants the SPECIALIST Lexicon[4], XMAP uses a dictionary of synonyms extracted from the UMLS Metathesaurus [21], and AML has three background resources, one of which is selected automatically [24]. The current version of SANOM, however, does not utilize any sort of background knowledge for the biomedical domain.

Table 3.4: The precision, recall, and F-measure of the systems participated in aligning DOID and ORDO ontologies from the disease and phenotype track.

	Precision	F-measure	Recall
LogMap	0.937	0.848	0.775
AML	0.514	0.646	0.870
LogMapLite	0.988	0.758	0.615
XMap	0.969	0.700	0.548
SANOM	0.975	0.747	0.605

Table 3.4 tabulates the result of the various alignment systems for aligning DOID and ORDO ontologies. According to this table, the precision of SANOM is better than those of

[4]http://wayback.archive-it.org/org-350/20180312141706

AML, LogMap, and XMap, and is competitive with LogMapLite. In terms of recall, on the other hand, AML and LogMap have better outcomes. SANOM is also better than XMap and is competitive with LogMapLite. Regarding F-measure, LogMap is the best system in this track followed by LogMapLite and SANOM. Thus, SANOM outperformed XMap and AML in this track in spite of the fact that it does not use any background knowledge.

Table 3.5 displays the performance of systems on aligning HP and MP ontologies. According to this table, SANOM has excellent performance in terms of precision and outperforms all systems in this sense. Regarding recall, LogMap and AML are the top two systems, and SANOM is better than XMap and is competitive with LogMapLite . Concerning F-measure, LogMap and AML are the best systems following by LogMapLite and SANOM.

Table 3.5: The precision, recall, and F-measure of the systems participated in aligning HP and MP ontologies from the disease and phenotype track.

	Precision	F-measure	Recall
LogMap	0.875	0.855	0.835
AML	0.889	0.843	0.801
LogMapLite	0.993	0.755	0.609
XMap	0.994	0.477	0.314
SANOM	0.995	0.728	0.574

The outcomes of SANOM on the disease and phenotype tracks are acceptable, but interestingly, its precision is high in contrast to other tracks. This gets back to the nature of the ontologies and the fact that SANOM has no use of biomedical background knowledge. Thus, SANOM can consider only the concepts as potential mappings which have a high string or structural similarity. In this case, mapping based solely on these similarity metrics have led to high precision and low recall. Another important topic is that the reference alignment has been created based on the alignments of other systems in previous years. Thus, the participating systems have contributions to the reference alignment which means that it is more likely that they have much more correspondences in common with the reference. SANOM, on the other hand, has not participated in this track before and has therefore no impact on the creation of the reference. Nevertheless, the performance of SANOM is comparable and acceptable.

3.6. CONCLUSION AND DISCUSSION

This chapter presented a new ontology alignment system, called SANOM, which uses the simulated annealing to find the correspondences among two given ontologies. The ontology matching problem was first revised as the minimization of a fitness function, and a compound fitness function was developed using several similarity metrics. The simulated annealing was then adapted to optimize the energy function and consequently derive the final alignment. SANOM has shown acceptable performance in discovering the correspondences of two given ontologies. Further, it is also fast and memory-efficient, especially in comparison to other alignment systems using evolutionary algorithms.

However, there are multiple avenues which SANOM can be improved. SANOM has

already acceptable performance in terms of recall, but its precision is not as good as its recall. This is probably due to the nature of the evolutionary algorithms which compute the *overall* fitness of an alignment or state, increasing the chance of existing false mappings in an intermediate state, while its fitness is still superior to others. Rejecting the false mappings could be done using an alignment repair technique. Thus, one avenue to enhance the precision of SANOM would be the use of an alignment repair technique.

Yet another way of improving SANOM is to use the background knowledge such as UMLS. Most of the tracks of the OAEI lie within the biomedical realm, hence utilization of such background knowledge would increase the performance on those tracks and help us fairly compare it with competing ones over those tracks.

SANOM is both memory- and time-efficient. However, mapping big ontologies, e.g., ontologies with more 50,000 concepts, is another problem. Recently, there are several suggestions to enable the alignment systems matching these ontologies as well [13]. For instance, one can store the concepts of one ontology in a hash table and then search the concepts of the second into this hash table. Since finding in hash tables is of order one, the overall searching of all concepts is of order m, where m is the number of concepts in the second ontology. It was shown that such a simple strategy finds many correspondences in the biomedical domain and decreases the consequent search space quadratically. Such suggestions are left for the future development of SANOM.

REFERENCES

[1] M. Cheatham and P. Hitzler, *String similarity metrics for ontology alignment,* in *International Semantic Web Conference* (Springer, 2013) pp. 294–309.

[2] W. Cohen, P. Ravikumar, and S. Fienberg, *A comparison of string metrics for matching names and records,* in *Kdd workshop on data cleaning and object consolidation*, Vol. 3 (2003) pp. 73–78.

[3] M. Cheatham and P. Hitzler, *The properties of property alignment.* in *OM* (2014) pp. 13–24.

[4] M. Mohammadi, A. Atashin, W. Hofman, and Y.-H. Tan, *Sanom results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 185.

[5] M. Mohammadi, W. Hofman, and Y.-H. Tan, *Sanom-hobbit: Simulated annealing-based ontology matching on hobbit platform,* Knowledge Engineering Review .

[6] M. Mohammadi, W. Hofman, and Y.-H. Tan, *Sanom results for oaei 2018,* in *OM-2018: Proceedings of the Thirteenth International Workshop on Ontology Matching* (2018).

[7] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *Equation of state calculations by fast computing machines,* Journal of Chemical Physics **21**, 1087 (1953).

[8] G. A. Miller, *Wordnet: a lexical database for english,* Communications of the ACM **38**, 39 (1995).

[9] Z. Wu and M. Palmer, *Verbs semantics and lexical selection,* in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (Association for Computational Linguistics, 1994) pp. 133–138.

[10] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, *Feature-rich part-of-speech tagging with a cyclic dependency network,* in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (Association for Computational Linguistics, 2003) pp. 173–180.

[11] M. F. Porter, *An algorithm for suffix stripping,* Program **14**, 130 (1980).

[12] M. Horridge and S. Bechhofer, *The owl api: A java api for owl ontologies,* Semantic Web **2**, 11 (2011).

[13] D. Faria, C. Pesquita, I. Mott, C. Martins, F. M. Couto, and I. F. Cruz, *Tackling the challenges of matching biomedical ontologies,* Journal of biomedical semantics **9**, 4 (2018).

[14] D. Lea, *A java fork/join framework,* in *Proceedings of the ACM 2000 conference on Java Grande* (ACM, 2000) pp. 36–43.

[15] D. Faria, B. S. Balasubramani, V. R. Shivaprabhu, I. Mott, C. Pesquita, F. M. Couto, and I. F. Cruz, *Results of aml in oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 122.

[16] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, *et al.*, *Results of the ontology alignment evaluation initiative 2016,* in *OM: Ontology Matching* (No commercial editor., 2016) pp. 73–129.

[17] E. Jiménez-Ruiz and B. C. Grau, *Logmap: Logic-based and scalable ontology matching,* in *International Semantic Web Conference* (Springer, 2011) pp. 273–288.

[18] M. KACHROUDI, G. DIALLO, and S. B. YAHIA, *Oaei 2017 results of kepler,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 138.

[19] S. Hertling, *Wikiv3 results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 190.

[20] J. da Silva, F. A. Baiao, and K. Revoredo, *Alin results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 114.

[21] O. Bodenreider, *The unified medical language system (umls): integrating biomedical terminology,* Nucleic acids research **32**, D267 (2004).

[22] J. Bock and J. Hettenhausen, *Discrete particle swarm optimisation for ontology alignment,* Information Sciences **192**, 152 (2012).

[23] I. Harrow, E. Jiménez-Ruiz, A. Splendiani, M. Romacker, P. Woollard, S. Markel, Y. Alam-Faruque, M. Koch, J. Malone, and A. Waaler, *Matching disease and phenotype ontologies in the ontology alignment evaluation initiative,* Journal of biomedical semantics **8**, 55 (2017).

[24] D. Faria, C. Pesquita, E. Santos, I. F. Cruz, and F. M. Couto, *Automatic background knowledge selection for matching biomedical ontologies,* PloS one **9**, e111226 (2014).

4

FREQUENTIST APPROACH FOR ALIGNMENT COMPARISON

The statistician cannot evade the responsibility for understanding the process he applies or recommends.

Sir Ronald Fisher

After discovering the alignments, several performance metrics are available to evaluate the alignment systems. The metrics typically require the identified alignment and a reference containing the underlying actual correspondences of ontologies in question. The current trend in the alignment evaluation is to put forward a new metrics (e.g., precision, weighted precision, semantic precision) and compare various alignment systems by juxtaposing the computed scores, or their averages in the case of having multiple benchmarks. However, claiming if one system has better performance than one another cannot be substantiated solely by comparing two figures. In this chapter, we propose statistical procedures which enable us to theoretically favor one system over another for a specific domains. We distinguish the comparison of alignment systems on one or multiple benchmarks, since they are statistically distinct from each other. For comparison over one benchmark, McNemar's test is adopted and its different statistics are revised and studied. For comparison of two systems over multiple benchmarks, Wilcoxon Signed-rank and McNemar's tests are recommended due to their robustness and statistical safety in different circumstances. The Friedman and Quade tests with their corresponding post-hoc procedures are studied for comparison of multiple systems over mutliple benchmarks, and their [dis]advantages are discussed. In the case of having more than two alignment systems for comparison, the family-wise error rate is expected to happen. Thus, the ways of preventing such an error are also discussed. The overall comparison is summarized by using directed graphs and critical difference diagrams for comparison over one and multiple benchmarks, respectively.

4.1. INTRODUCTION

In this chapter, appropriate statistical procedures are empirically and theoretically studied, which allow verifying the claim of significant difference among alignment systems. These methods also enable us to compare robustly the results of alignment systems obtained from one or multiple benchmarks, and determine if one system is superior to another. For the case of comparing multiple systems, the chances are high that they are declared not significantly different; therefore, no single system might be the best as the result of the statistical analysis.

The comparison of alignment systems over one or multiple benchmarks (or matching tasks) is sharply distinguished in terms of statistical inference, while the modus operandi in the literature for both cases suffers from the same pitfalls, i.e., the comparison is based on two figures: A performance score or its average. Currently, the mean of performance scores is the only yardstick toward which various ontology matching systems are compared over multiple benchmarks. However, averages are sensitive to outliers. The existence of outliers is seemingly inevitable in ontology matching, since some systems have poor performance on particular benchmarks due to either their difficulty or the system's incapability to produce a correct alignment. On top of that, the poor performance of a system on one single benchmark would cancel out the fair performances over the remaining benchmarks (and vice versa), thereby influencing the overall average performance. Furthermore, one system is claimed to have superior performance over another either the discrepancy between their averages is small or large. However, the slight difference between averages can be ignored and claiming that the systems are significantly different might be wrong. Also, the sole comparison of averages is not substantiated by any evidence. By the same token, the comparison on one benchmark is also made by the juxtaposition of one performance score, and the system with a higher score, regardless of the difference, is claimed to be superior.

The claim of the superior performance of an alignment system might be refuted based on the no free lunch (NFL) theorem [1, 2]. According to the NFL theorem, there is no single system which performs well in all scenarios [1]. However, there is usually background knowledge available which can distinguish the performance of one system over the rest in one particular domain, e.g., one system performs better on biomedical ontologies and another on logistics ontologies. Therefore, the outcome of this chapter as well as the next two chapters is not in contradiction to that of the NFL theorem as it is sought to find the superior system in a particular domain or on a particular benchmark.

For using the statistical tests, suppose that k systems are tested over N benchmarks. Let P_i^j be the performance score of the j^{th} system on the i^{th} benchmark. The goal is to decide if the systems are different from each other based on their performance scores P_i^j, which inherently indicates that one system is better. Such an approach has been scrutinized in other areas of research [3–9]. Demšar [3] studied the statistical procedures for comparing two or more classifiers over multiple benchmarks. Garcia et al. [4, 5] extended the Demšar work and proposed more advanced non-parametric tests and their corresponding post-hoc procedures for comparison of multiple classifiers. Trawinski et al. [8] compared the regression learning algorithms and utilized various statistical tests to do so. Similar approaches are applied to other areas such as information retrieval [6] and evolutionary algorithms [9].

The performance analysis of alignment systems is different from the areas of research which have already considered statistical inference. First and foremost, the number of benchmarks for matching, especially at the OAEI, is either large enough (roughly speaking more than 30 benchmarks) or very small (less than 10 benchmarks or matching tasks.) In contrast, the number of benchmarks in other areas is usually assumed to be moderate, e.g., more than 10 but less than 30. The assumption on the number of benchmarks is valid due to either the lack of benchmarks or the difficulties of running the methods over a large number of benchmarks. From the statistical point of view, the moderate and small sample size put an obstacle in the way of checking the presumptions of the statistical tests and invalidate the results of the parametric tests. Therefore, the current trend is to favor the non-parametric statistics for comparison. In ontology alignment, on the other hand, it is possible to check the presumption of parametric tests as there are enough benchmarks in several tracks such as *benchmark* and *multifarm*. We further investigate the case that a few benchmarks, e.g., less than ten, are available, and propose utilizing McNemar's test for comparison. For the moderate number of benchmarks, the Wilcoxon Signed-rank test is recommended as it is the case in other fields [3].

Another crucial difference between machine learning problems and ontology alignment is that there is no resampling method, such as k-fold cross-validation, in the latter. The resampling methods would result in having multiple samples from a benchmark so that the comparison can be made according to these samples. For ontology alignment, on the other hand, there are no such samples so that applying statistical tests are not straightforward for comparison over one benchmark. In this regard, we adopt McNemar's test to compare various alignment systems on one benchmark. This test can be applied to the paired nominal data summarized in a contingency table with a dichotomous trait. Interestingly, the outcome of two alignment systems can be viewed as dichotomous (i.e., correct and incorrect correspondences) of two experiments (i.e., two alignment systems). Therefore, McNemar's test suits for comparison of alignments over one benchmark or matching task. Two ways for creating a contingency table based on alignments are discussed and their similarity with recall and F-measure is explained.

The remainder of this chapter is structured as follows. Section 4.2 explains the notions of null hypothesis significance testing and the way we use it for comparing alignment systems. Section 4.3 presents the adaptation of McNemar's test for comparing alignment systems over one benchmark, while Section 4.4 is dedicated to the statistical tests for comparison over multiple benchmarks. The ways of controlling the family-wise error rate are studied in detail in Section 4.5. The comparison of statistical tests together along with applying them to several OAEI tracks are discussed in Section 4.6, and Section 4.7 provides final remarks and conclusions.

4.2. STATISTICAL SIGNIFICANCE TESTING

The null hypothesis significance testing (NHST) is of the essence in the realm of statistical inference. Here, we aim at utilizing NHST to compare alignment systems and identifying systems with superior performance. In this regard, there are various statistical tests which can be used in different circumstances. There are two determining factors for selecting the best statistical tool for comparison. First, the number of benchmarks, that

Table 4.1: The possible use of statistical tests with respect to the number of benchmarks and the number of alignment systems to be compared.

	Two systems	Multiple systems
one benchmark	McNemar's test	Pairwise McNemar's test
Multiple benchmarks	Paired t-test Wilcoxon Singed-rank test McNemar's test	Repeated Measure ANOVA Friedman Quade

is the number of samples for the statistical tests, is crucial and profoundly impacts the tests to be used. Another factor is the number of alignment systems to be compared. In this regard, we have to consider different statistics and prevent the so-called family-wise error rate from happening.

To leverage the hypothesis testing, a null hypothesis is required. The null hypothesis, shown by H_0, states that there is no significant difference between two or more populations according to the available samples. On the other hand, the alternative hypothesis, shown by H_a, is the opposite of the null hypothesis and states that there is a meaningful difference between two or more populations based on the available samples. Thus, it is desirable to reject the null hypothesis and instead, accept the alternative. In the ontology matching case, especially at the OAEI, it is usually the case that the outcome of various systems over a range of benchmarks is available and it is sought to verify which system has better performance than the others. To compare k systems, the null and alternative hypotheses are

$$\begin{aligned} &H_0 : \hat{P}^1 = \hat{P}^2 = ... = \hat{P}^k \\ &H_a : \text{at least one } \hat{P}^i \text{ differs} \end{aligned} \tag{4.1}$$

where $\hat{P}^i$ represents the overall performance of the i^{th} system. The overall performance of an alignment system in the null hypothesis varies considerably in different statistics; hence, they are precisely introduced for each particular test.

This chapter reviews the relevant tests to find the probability of observing the alignments generated by systems given H_0 is correct (this probability is called *p-value*.) If a p-value is less than the nominal significance level α, which must be determined in advance, the null hypothesis is rejected, and the systems in question are declared significantly different. Otherwise, it fails to reject the null hypothesis.

Table 4.1 tabulates the possible use of different statistical tests with respect to the number of benchmarks and number of alignment systems to be compared. We compare the statistical tests in terms of their power and replicability for comparison of alignment systems with a different number of benchmarks (see Section 4.6.1) and shows that which of the tests tabulated in Table 4.1 are the appropriate, especially with respect to the number of benchmarks. According to Table 4.1, if the comparison is based solely on one benchmark, McNemar's test is adopted [10] to compare two alignment systems, and its various statistics are reviewed and compared together. For comparing multiple alignment systems on one benchmark, McNemar's test is again used, where it is applied

Table 4.2: A simple contingency table.

		Exp. 2		
		-	+	sum
Exp. 1	-	n_{00}	n_{01}	$n_{0.}$
	+	n_{10}	n_{11}	$n_{1.}$
	sum	$n_{.0}$	$n_{.1}$	M

to each pair of alignment systems.

For comparison of two alignment systems over multiple benchmarks, the first test considered here is the paired t-test. However, it could be statistically unsafe due to its strong presumptions. Therefore, the non-parametric tests, the Wilcoxon Signed-rank [11] and McNemar's [10] tests, are proposed to be utilized, since they have fewer and easy-to-satisfy presumptions.

The comparison of multiple systems over multiple benchmarks is more challenging. The null hypothesis, in this case, is that all systems perform equally, and if it is rejected, it is drawn that there is at least one system with different performance. However, it cannot determine what systems are significantly different. A post-hoc procedure then follows to indicate where exactly the difference among performance scores occur. The former test is called the *omnibus* test, and the latter is referred to as the *post-hoc* test. The repeated measures ANOVA [12], Friedman [13] and Quade [14] tests and their corresponding post-hoc procedures are discussed in detail. The family-wise error rate (FWER) is a well-known error in statistical inference, which happens when there are multiple systems to be compared, either the number of benchmark is one or more. The ways of preventing such an error are elaborated. The following sections explain the tests in detail and examples are provided if need be.

4.3. COMPARISON OVER ONE BENCHMARK

We now consider the case that there is one benchmark over which multiple alignment systems are compared. To that end, we first review the ways of creating the contingency table to which McNemar's test can be applied.

4.3.1. CONTINGENCY TABLE CONSTRUCTION

McNemar's test is applicable when there are two experiments over M samples [10]. Let the outcome of each test be either positive or negative; then, a simple contingency table would be as Table 4.2. In this table, n_{00} and n_{11} are called the accordant pair and are the number of times both experiments produce positive and negative outcomes, respectively. The discordant pair, i.e., n_{01} and n_{10}, are the number of times the results of experiments are in contradiction; n_{01} is the number of experiments in which the first outcome is negative while the second one is positive, and n_{10} is the other way around.

For ontology matching, the positive or negative outcome can be defined in two ways, each of which has its own merits and is suitable for a particular situation. For two ontologies O and O' in question, let R be the reference alignment containing a set of correct correspondences, and A_1 and A_2 be two alignments retrieved by two different systems.

In the first approach of the contingency table construction, the focus is solely on the true positives and false positives are ignored. Hence, n_{00} and n_{11} are the number of false negatives and true positives jointly identified by both systems, respectively. n_{01} (and similarly n_{10}) is the number of correct correspondences only in A_2 and not in A_1. These elements can be written as:

$$\begin{cases} n_{00} = |R - (A_1 \cup A_2)|, \\ n_{01} = |(A_2 \cap R) - A_1|, \\ n_{10} = |(A_1 \cap R) - A_2|, \\ n_{11} = |A_1 \cap A_2 \cap R|, \end{cases} \tag{4.2}$$

where |.| indicates the cardinality operator. This approach is conceptually similar to *recall* as it does not consider the wrong correspondences in the alignments. We again accent that this approach is distinct from the performance metrics, including recall, as we compare two alignments and do not produce any score indicating the fineness of a system.

An example elaborates on the issue of this approach. Assume that two systems could discover the complete reference alignment, i.e., $A_1 = A_2 = R$. In this case, $n_{01} = n_{10} = 0$ which means that they are equally well (it is discussed in further sections that n_{01} and n_{10} are the only important pair for McNemar's test). Now, suppose that $A_1 = R$ and $A_2 = R+B$, where B is a set of correspondences that are not in R. In this case, n_{01} is the same as n_{10} which again indicates that their performances are indiscernible. However, it is easy to see that A_1 is more reliable as it does not mistakenly discover any correspondences. Statistically speaking, this approach does not take into account false positives and only considers true positives. Nonetheless, such an approach is suitable for occasions where the goal is to have as many correspondences as possible so that the false positives do not have a profound impact.

The second approach of building the contingency table avoids the foregoing pitfall and considers the false positives as well, thereby having a higher complexity for computing the numbers compared to the previous approach. Therefore, it is necessary to explain how to compute each element of the contingency table individually.

For this approach, n_{00} is the number of false positives as well as false negatives of both alignments. Hence, it includes the correspondences that are in R but not in A_1 nor A_2 plus the correspondences which are in both A_1 and A_2 but not in R, i.e., $n_{00} = |R - (A_1 \cup A_2)| + |(A_1 \cap A_2) - R|$. n_{10} is the number of correct correspondences in A_1 which are not in A_2 plus the false correspondences identified only by A_2 and not by A_1, i.e., $n_{10} = |(A_1 \cap R) - A_2| + |A_2 - A_1 - R|$. By the same token, n_{01} can also be obtained. n_{11} is a bit more challenging as the total number of possible correspondences between two ontologies is required. Let this number be W, one possibility for W is to multiply the number of entities of two ontologies, i.e. $W = |O| \times |O'|$, where |.| is the cardinality operator. Thus, $n_{11} = |A_1 \cap A_2 \cap R| + |(W - R) - (A_1 \cup A_2)|$. The statistics considered in this paper only need the discordant pair; making the values of n_{11} and subsequently W unimportant. The elements as mentioned earlier of the contingency table from the second approach

can be summarized as:

$$\begin{cases} n_{00} = |R - (A_1 \cup A_2)| + |(A_1 \cap A_2) - R|, \\ n_{01} = |(A_2 \cap R) - A_1| + |A_1 - A_2 - R|, \\ n_{10} = |(A_1 \cap R) - A_2| + |A_2 - A_1 - R|, \\ n_{11} = |A_1 \cap A_2 \cap R| + |(W - R) - (A_1 \cup A_2)|. \end{cases} \tag{4.3}$$

This way of contingency table construction considers false positives as well. The foregoing example illustrates the advantages of these formulas, where $A_1 = R$ and $A_2 = R + B$, $n_{01} = 0$ and $n_{10} = |B|$. The null hypothesis is thus rejected for large enough of B, and A_1 is claimed to be superior. Therefore, the false positive of B resulted in declaring A to be a better system. Note that this calculation is relative to the other system. In other words, it does not consider all the incorrect correspondences, but the ones that are not in the rival alignment system. As the goal is to compare two alignments together, it is entirely logical to find the *relative false positives* for two systems. This approach can be figuratively viewed as similar to F-measure, since it takes both true and false positives into account.

4.3.2. MCNEMAR'S TEST

McNemar's test is applied to the contingency table constructed in the previous section. The null hypothesis in McNemar's test states that the two marginal probabilities of the contingency table are the same, i.e.,

$$\begin{aligned} p(n_{00}) + p(n_{01}) &= p(n_{00}) + p(n_{10}), \\ p(n_{10}) + p(n_{11}) &= p(n_{01}) + p(n_{11}), \end{aligned} \tag{4.4}$$

where $p(a)$ indicates the probability of occurring cell a in Table 4.2. After canceling out the $p(n_{00})$ and $p(n_{11})$ from equations (4.4), the null and alternative hypotheses become

$$\begin{aligned} H_0 &: \quad p(n_{01}) = p(n_{10}) \\ H_a &: \quad p(n_{01}) \neq p(n_{10}). \end{aligned} \tag{4.5}$$

To compute the p-value of the null hypothesis (4.5), we consider four statistics from McNemar's test and discuss their advantages and pitfalls in the hypothesis testing. The statistics studied here only work with the accordant pair of the contingency table. However, there is also an exact unconditional McNemar's test which takes into account the discordant pair of the contingency table [15]. The exact unconditional test is way more intricate than the other McNemar's tests put forward here, but its power is approximately the same as other tests [16]. Therefore, we do not include the test.

MCNEMAR'S ASYMPTOTIC TEST

McNemar's asymptotic test assumes that n_{01} is binomially distributed with $p = 0.5$ and parameters $n = n_{01} + n_{10}$ under the null hypothesis [10]. McNemar's asymptotic statistic, defined as,

$$\chi^2 = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}}$$

is distributed according to χ^2 with one degree of freedom. This test is undefined for $n_{01} = n_{10} = 0$.

To reject the null hypothesis, this test requires a sufficient number of data ($n_{01} + n_{10} \geq 25$), since it might violate the nominal significance level α for the small sample size.

MCNEMAR'S EXACT TEST

It is traditionally advised to use McNemar's exact test when a small sample size is available in order not to exceed the nominal significance level. In this test, n_{01} is compared to a binomial distribution with parameter $n = n_{01} + n_{10}$ and $p = 0.5$. Thus, the p-value for this test is obtained as:

$$\text{exact-p-value} = \sum_{x=n_{01}}^{n} \binom{n}{x} \left(\frac{1}{2}\right)^2.$$

The two-sided p-value is calculated by multiplication of the one-sided p-value by two. This test guarantees to have type I error rate below the nominal significance level α.

MCNEMAR'S ASYMPTOTIC TEST WITH CONTINUITY CORRECTION

The main drawback of McNemar's exact test, though preserving the nominal significance level, is conservatism: It unnecessarily generates large p-values so that the null hypothesis cannot be rejected. As a remedy to conservatism, Edwards [17] approximated the exact p-value by the following continuity corrected statistic:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}},$$

which is χ^2-distributed with one degree of freedom. This test is also undefined for $n_{01} = n_{10} = 0$.

MCNEMAR'S MID-P TEST

The continuity corrected method is not as conservative as the exact test, but it does not guarantee to preserve the nominal significance level. The mid-p approach propounds a way to trade off between the conservatism of the exact tests and the significance level transgression of the continuity correction approach [18]. To obtain the mid-p-value, a simple modification is required: The mid-p-value equals the exact p-value minus half the point probability of the observed test statistic [16]. Hence, the p-value could be computed as:

$$\text{mid-p-value} = \text{2-sided exact p-value} - \binom{n}{n_{01}} 0.5^n.$$

McNemar's mid-p test resolves the conservatism of the exact test, but it does not guarantee theoretically to preserve the nominal significance level. In a recent study, however, it is investigated that the mid-p test has low type I error and does not violate the significance level. The continuity-corrected test, in contrast, indicated a high type I error, coming from the nature of asymptotic tests, as well as high type II error, inherited from the exact test. Thus, it is rational not to use the continuity-corrected test for the alignment comparison.

Table 4.3: The tests for comparison of two systems over N benchmarks or matching tasks. *Applicability* is roughly the situation that test can be used and its results are valid and *differences* refers to the differences of performance scores.

Test	Presumptions	Applicability
Paired t-test	Normality of differences	N > 30
Wilcoxon Signed-rank	symmetry of differences w.r.t median	N > 10
McNemar's test	-	N < 10

4.4. COMPARISON OVER MULTIPLE BENCHMARKS

For comparing alignment systems over multiple benchmarks, we first need to specify a performance metric according to which alignment systems are compared. We review the appropriate statistical tests for comparing two or multiple systems.

4.4.1. COMPARISON OF TWO SYSTEMS

This section is dedicated to comparing two systems over multiple benchmarks. The tests are summarized in Table 4.3.

PAIRED T-TEST

A common way to detect the difference between two systems is to compute the paired t-test statistic. Let $d_i = P_i^1 - P_i^2$ be the difference between the performance scores of two alignment systems over the i^{th} benchmark. The t statistic is computed as

$$t = \hat{d}/\hat{\sigma_d}, \tag{4.6}$$

where $\hat{d}$ and $\hat{\sigma}_d$ are the average of differences d_i and standard deviation of samples, respectively. This statistic is distributed according to the Student's t-distribution with $N-1$ degrees of freedom, where N is the number of benchmarks. After obtaining the probability of observing the performances given H_0 being true (p-value) according to the Student's t-distribution, H_0 can be rejected if p-value $\leq \alpha$. The rejection of the null hypothesis indicates the superiority of the system with a higher average performance.

The major drawback of using the paired t-test is the imposed assumption on the differences d_i. According to this test, the differences must be normally distributed in order for the obtained results to be reliable. In the case of comparison among alignment systems, however, there is no provision on the normality of the differences. One way to overcome this is to provide the paired t-test with large enough samples (~30 benchmarks) so that the normality can be assumed according to the *central limit theorem*. Another way is to check the normality of distribution using various tests. Ironically, these tests have less power on small samples, making them unlikely to detect abnormalities.

Another pitfall of the paired t-test is the sensitivity to outliers. Outliers can skew the test statistic and increase the estimated standard error which adversely influences the power of the test. The existence of outliers can lower the power of the paired t-test as the averaging operator. In the case of normality violation, the non-parametric tests should be considered due to their robustness against outliers and the fewer presumptions on the sample distribution.

To verify the applicability of the paired t-test for the OAEI, we took pairs of systems from various tracks (e.g., benchmark[1], multifarm, etc.) and applied the normality test [19]. As there are large sample sizes in several tracks, such as benchmark and multifarm, the normality test might have a reliable outcome. Our investigation showed that in less than 7% of cases, the normality assumption holds. On top of that, it is usually the case that some systems fail to produce acceptable alignments for some particular tasks. Therefore, the existence of outliers seems to be inevitable.

WILCOXON SIGNED-RANK TEST

The non-parametric alternative to the paired t-test is Wilcoxon Signed-rank test [11]. This test ranks the absolute values of performance differences between two systems. Then, it compares the average rank of positive and negative differences.

After computing the difference between two systems over the i^{th} benchmark, e.g., d_i, the differences are ranked based on the values of d_i, disregarding its sign. The number of $d_i = 0$ are evenly split between the sum of ranks. Let W^+ and W^- be

$$W^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$
$$W^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i), \tag{4.7}$$

and $T = min(W^+, W^-)$. If fewer than 25 benchmarks are available, then a table consisting of critical values for T must be utilized [12]. If the number of benchmarks exceeds 25, then the statistics,

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}, \tag{4.8}$$

follows the standard normal distribution. If the p-value is less than α, then we reject the null hypothesis and accept that there is a significant difference between the two systems. Consequently, it is drawn that the system with the higher sum of ranks is better.

An example elaborates on the procedure of the test. Table 4.4 shows F-measure of two systems, edna [20] and GMap [21], over 20 tasks from the *benchmark* track along with the difference in their performance scores and the ranks obtained by the Wilcoxon Signed-rank test. We selected 20 tasks from the benchmark track to be able to demonstrate the ranks in a table. According to this test, $T = min(200, 10) = 10$ and $N = 20$; therefore, the p-value is nearly zero and the null hypothesis is rejected with high confidence. As a result, GMap is claimed to have outperformed edna.

This test assumes the symmetry of differences between the performances score concerning its median [22]. This assumption is not as strong as the normality assumption but can decrease the power of the test if not satisfied. The differences in scores are also considered in this test by assigning higher ranks to benchmarks over which the difference between two systems is bigger. In the next section, McNemar's test is used for comparison over multiple benchmarks. McNemar's test does not impose any presumptions for conducting the test that imposes no assumptions on the samples.

[1] At the time doing the research, the benchmark track was still a part of the OAEI.

Table 4.4: The F-measure scores, their differences, and ranks obtained by the Wilcoxon Signed-rank test of two systems, edna [20], GMap [21], over 20 benchmarks from the *benchmark* track.

	edna	**GMap**	d_i	**rank**
1	0.70	0.98	-0.28	13
2	0.02	0.80	-0.78	20
3	0.62	0.95	-0.33	14
4	0.47	0.90	-0.43	17
5	0.31	0.86	-0.55	18
6	0.17	0.83	-0.66	19
7	0.01	0.00	0.01	1
8	0.62	0.87	-0.25	10
9	0.47	0.73	-0.26	12
10	0.31	0.56	-0.25	11
11	0.16	0.33	-0.17	5
12	0.78	0.98	-0.2	7
13	0.77	0.99	-0.22	9
14	0.78	0.98	-0.2	7
15	1.00	0.98	0.02	2.5
16	0.78	0.98	-0.2	7
17	0.55	0.96	-0.41	15.5
18	1.00	0.98	0.02	2.5
19	0.55	0.96	-0.41	15.5
20	1.00	0.96	0.04	4

MCNEMAR'S TEST

The Sign test is usually considered as one of the alternatives for comparing two populations. The main drawback of the Sign test is its conservativeness, making this test be barely used. However, the Sign test is a special case of McNemar's exact test, which is, as discussed in the previous section, the most conservative one in this family, and there are several other statistics from this family which are more powerful and can be used instead of the exact test.

The contingency table construction has been discussed for comparing over single benchmark. For comparison of two systems A_1 and A_2 over N benchmarks, n_{01} is the number of benchmarks over which the performance scores of A_2 are greater than those of A_1. By the same token, n_{10} is the number of benchmarks where the performance scores of A_1 are higher than those of A_2. The cases of equality are not considered in this test. As a result, McNemar's test can also be used for comparison over multiple benchmarks as well.

If the null hypothesis is rejected, then it is concluded that the system which has outperformed the other over more benchmarks is better. The Four McNemar's statistics discussed in the previous chapter can be applied to this contingency table to verify the difference between two systems over multiple benchmarks.

Table 4.5: The tests for comparison of multiple systems over *N* benchmarks. Applicability is roughly the situation that test results are valid.

Test	Presumptions	Applicability
ANOVA	Sphericity	N > 30
Friedman	-	N > 10
Quade	-	N < 10

4.4.2. COMPARISON OF MULTIPLE SYSTEMS

In this section, the simultaneous comparison of multiple alignment systems is discussed. The null hypothesis here is that the performance of all systems are the same and the alternative hypothesis is that there is at least one system with different performance. In statistics, the comparison of multiple populations consists of two phases: The *omnibus* and *post-hoc* tests, the former of which only detects if there is a significant difference among alignment systems, while the latter precisely indicates the alignment systems with different performance. Table 4.5 summarizes the tests of this section.

OMNIBUS TESTS

In this section, three tests, repeated measures ANOVA [12], Friedman [13] and Quade [14] tests, are reviewed and their advantages and drawbacks are discussed in detail.

REPEATED MEASURES ANOVA

The most well-known test for detecting the difference among more than two related samples is the repeated measures ANOVA. The null hypothesis is that all systems perform equally well. In the repeated measures ANOVA, the total variability is divided into variability between systems, variability between benchmarks and the residual error variability [3]. The between systems' variability is a measure between the variances of the means of the alignment systems [12]. The residual variability, on the other hand, is viewed as the variability by chance. The repeated measures ANOVA would reject the null hypothesis if the between-systems' variability was significantly larger than the residual variability.

As any parametric test, the repeated measures ANOVA is predicated on several assumptions whose violation can invalidate the obtained results. The first assumption is that the data are normally distributed. Although there is no guarantee that the data are normally distributed, statisticians do not ignore the ANOVA for abnormality of distribution unless the distribution is bi-modal [3, 23]. The most important assumption of this test is *sphericity*. Sphericity refers to the conditions where the variances of the differences between each possible pair of groups are equal. This assumption is more likely to be violated as there is no guarantee for the parity of differences' variances. The violation of sphericity invalidates the obtained results and consequently influences the post-hoc test.

The well-known test for checking sphericity is Mauchly's test [24]. We have conducted this test over the results of the OAEI in recent years, and the assumption of sphericity was unexceptionally rejected with an extremely-significant p-value. Even if the sphericity assumption is not rejected, Mauchly's test is usually avoided, since it is

not able to detect the transgression against sphericity in small samples and falsely detect it in large samples. As a result, it is recommended to exploit the non-parametric tests for comparison.

FRIEDMAN TEST

The Friedman test [13] is the non-parametric counterpart of the repeated measures ANOVA and is the extension of the binomial Sign test (or McNemar's exact test with p=0.5). Instead of using the scores themselves for computing the statistic, it first ranks the scores and uses them in the calculation of the statistic. The ranking procedure is among the scores of different systems over one specific benchmark in a way that the best performance score takes the rank of 1 and the worst takes the rank of k, where k is the number of alignment systems. The average rank is assigned if the scores tie.

Let r_i^j be the rank of the j^{th} system on the i^{th} benchmark. If two systems perform equally, it is expected that their average ranks across all the benchmarks are the same. The Friedman statistic,

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right], \tag{4.9}$$

is χ^2 distributed with $k-1$ degrees of freedom. It is investigated that the type II error of equation (4.9) is undesirably high; therefore, a better statistic is derived by Iman-Davenport [25],

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \tag{4.10}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. An example in the next section elaborates on the procedure of finding the Friedman statistic.

QUADE TEST

The Friedman test is only predicated on the ranks of systems over each benchmark. The Quade test [26], on the other hand, takes into account the performance variation among benchmarks and is suitable when the number of benchmarks is small (roughly less than 10 benchmarks). The underlying assumption behind the Quade test is that if the scores' variation over a benchmark is larger, then it is a more challenging one to be aligned. Thus, the success of a system over such benchmarks indicates that it is a better system.

To find the ranks of each method, the range of scores over one benchmark is computed by subtracting the maximum score from the minimum one. Then, the minimum and the maximum range takes the rank of 1 and n, respectively. Let $Q_1, Q_2, ..., Q_n$ be the ranks of n benchmarks and r_i^j be the ranks obtained by the Friedman test for each score. The Quade rank of each score is obtained as $S_i^j = Q_i \left(r_i^j - \frac{k+1}{2} \right)$. Finally, the test statistic is:

$$F_{Quade} = \frac{(N-1)\sum_{j=1}^{k}(S^j)^2}{A - \frac{1}{n}\sum_{j=1}^{N}(S^j)^2}, \tag{4.11}$$

where

$$S^j = \sum_i S_i^j \qquad A = \frac{N^2(N+1)(2N+1)k(k+1)(k-1)}{72},$$

and F_{Quade} is distributed according to F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. The next section includes an example of the calculation of this statistic.

We perform the Quade test on the scores in the above table. Table 4.7 displays the benchmarks' ranks and scores' ranks of the Quade test. The test statistic is:

$$F_{Quade} = 10.16$$

which is distributed according to the F-distribution with $(3, 57)$ degrees of freedom. The corresponding p-value is 1.84×10^{-5} which results in rejecting the null hypothesis.

AN EXAMPLE

In this section, the procedure of Friedman and Quade tests are elaborated by an example. In this regard, we select 20 matching tasks from the benchmark track to demonstrate the calculation of ranks in Friedman and Quade tests. Table 4.6 tabulates *F-measure* of

Table 4.6: F-measure scores and the Friedman ranks (in the parenthesis) of four alignment systems over 20 matching tasks from the OAEI *benchmark* track. Each row and each column correspond to a matching task and a system, respectively. The last row shows the average Friedman rank.

	edna	**GMap**	**LogMap**	**XMap**
1	0.70 (**4**)	0.98(**2**)	0.95(**3**)	1 (**1**)
2	0.02 (**2**)	0.80 (**1**)	0.00(**3.5**)	0 (**3.5**)
3	0.62 (**4**)	0.95(**1**)	0.87 (**2**)	0.66 (**3**)
4	0.47 (**4**)	0.90 (**1**)	0.72 (**2**)	0.65 (**3**)
5	0.31 (**4**)	0.86 (**1**)	0.52 (**2**)	0.51 (**3**)
6	0.17 (**3**)	0.83 (**1**)	0.28 (**2**)	0.15 (**4**)
7	0.01 (**1**)	0.00 (**3**)	0.00 (**3**)	0.00 (**3**)
8	0.62 (**4**)	0.87(**1.5**)	0.87 (**1.5**)	0.65 (**3**)
9	0.47 (**4**)	0.73 (**1**)	0.71 (**2**)	0.65 (**3**)
10	0.31 (**4**)	0.56 (**1**)	0.50 (**2**)	0.42 (**3**)
11	0.16 (**4**)	0.33 (**1**)	0.31 (**2**)	0.19 (**3**)
12	0.78 (**4**)	0.98 (**2**)	0.95 (**3**)	1.00 (**1**)
13	0.77 (**3**)	0.99 (**1**)	0.00 (**4**)	0.8 (**2**)
14	0.78 (**4**)	0.98 (**2**)	0.95 (**3**)	1.00 (**1**)
15	1.00 (**1.5**)	0.98 (**3**)	0.94 (**4**)	1.00 (**1.5**)
16	0.78 (**4**)	0.98 (**2**)	0.95 (**3**)	1.00 (**1**)
17	0.55 (**4**)	0.96 (**2**)	0.92 (**3**)	1.00 (**1**)
18	1.00 (**1.5**)	0.98 (**3**)	0.95 (**4**)	1.00 (**1.5**)
19	0.55 (**4**)	0.96 (**2**)	0.92 (**3**)	1.00 (**1**)
20	1.00 (**1.5**)	0.96 (**3**)	0.92 (**4**)	1.00 (**1.5**)
R_j	**3.2750**	**1.7250**	**2.8000**	**2.2000**

Table 4.7: F-measure scores and the Quade ranks (in the parenthesis) of four systems over 20 matching tasks from the OAEI *benchmark* track. Each row corresponds to a benchmark, the second column is the range, and the third is Q_i in the Quade test.

	Range	**Q$_i$**	**edna**	**GMap**	**LogMap**	**XMap**
1	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95(**3.75**)	1(**-11.25**)
2	0.8	19	0.02 (**-9.5**)	0.8 (**-28.5**)	0 (**19**)	0 (**19**)
3	0.33	13	0.62 (**19.5**)	0.95 (**-19.5**)	0.87 (**-6.5**)	0.66 (**6.5**)
4	0.43	14	0.47 (**21**)	0.9 (**-21**)	0.72 (**-7**)	0.65 (**7**)
5	0.55	17	0.31 (**25.5**)	0.86 (**-25.5**)	0.52 (**-8.5**)	0.51 (**8.5**)
6	0.68	18	0.17 (**9**)	0.83 (**-27**)	0.28 (**-9**)	0.15 (**27**)
7	0.01	1	0.01 (**-1.5**)	0 (**0.5**)	0 (**0.5**)	0 (**0.5**)
8	0.25	10	0.62 (**15**)	0.87 (**-10**)	0.87 (**-10**)	0.65 (**5**)
9	0.26	12	0.47 (**18**)	0.73 (**-18**)	0.71 (**-6**)	0.65 (**6**)
10	0.25	11	0.31 (**16.5**)	0.56 (**-16.5**)	0.5 (**-5.5**)	0.42 (**5.5**)
11	0.17	5	0.16 (**7.5**)	0.33 (**-7.5**)	0.31 (**-2.5**)	0.19 (**2.5**)
12	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95 (**3.75**)	1 (**-11.25**)
13	0.99	20	0.77 (**10**)	0.99 (**-30**)	0 (**30**)	0.8 (**-10**)
14	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95 (**3.75**)	1 (**-11.25**)
15	0.06	3	1 (**-3**)	0.98 (**1.5**)	0.94 (**4.5**)	1 (**-3**)
16	0.22	7.5	0.78 (**11.25**)	0.98 (**-3.75**)	0.95 (**3.75**)	1 (**-11.25**)
17	0.45	15.5	0.55 (**23.25**)	0.96 (**-7.75**)	0.92 (**7.75**)	1 (**-23.25**)
18	0.05	2	1 (**-2**)	0.98 (**1**)	0.95 (**3**)	1 (**-2**)
19	0.45	15.5	0.55 (**23.25**)	0.96 (**-7.75**)	0.92 (**7.75**)	1 (**-23.25**)
20	0.08	4	1 (**-4**)	0.96 (**2**)	0.92 (**6**)	1(**-4**)

four systems, namely, edna [20], GMap [21], LogMap [27], and XMap [28], across the OAEI *benchmark* track. The numbers in the parenthesis are the Friedman ranks of each method over the corresponding matching task. Then, the Friedman statistic can be calculated as

$$\textbf{(Friedman)} \quad \chi_F^2 = \frac{12\times 20}{4\times 5}\left(3.275^2 + 1.725^2 + 2.8^2 + 2.2^2 - \frac{4\times 5^2}{4}\right) = 16.575,$$

$$\textbf{(Iman Davenport)} \quad F_F = 7.25.$$

As the experiment consists of four systems over 20 matching tasks, χ_F^2 has χ^2 distribution with $4-1=3$ degrees of freedom and F_F is distributed according to the F-distribution with $4-1=3$ and $(4-1)(20-1)=57$ degrees of freedom. The p-values calculated for the Friedman and Iman-Davenport tests are 8.65×10^{-4} and 3.33×10^{-4}, respectively. Thus, the null hypothesis is rejected in both cases.

4.4.3. POST-HOC ANALYSIS

If the null hypothesis in the omnibus tests is rejected, a post-hoc test will precisely determine the differences among the systems.

The following statistics must be computed for each pair of systems (i, j) for Friedman

and Quade tests:

$$\text{Friedman} \qquad z = \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6N}}},$$
$$\text{Quade} \qquad z = \frac{T_i - T_j}{\sqrt{\frac{k(k+1)(2N+1)(k-1)}{18N(N+1)}}}, \tag{4.12}$$

where R_i is the average ranks in the Friedman test and $T_i = \frac{2\sum_{i,j} Q_i r_i^j}{N(N+1)}$ in the Quade test. The probability of systems i and j having the same performance can be calculated using these statistics which are distributed according to the standard normal distribution. Similar to the comparison of two systems, one can reject the null hypothesis and conclude that the two systems are significantly different provided that the computed probability is less than α. If the null hypothesis is rejected, then the system with lower average rank, in both Friedman and Quade tests, is claimed to be better.

4.5. FAMILY-WISE ERROR RATE AND P-VALUE ADJUSTMENT

When two systems are compared, the null hypothesis will be rejected if the obtained p-value is below the nominal significance level α. If more than two alignment systems are to be compared, the well-known family-wise error rate (FWER) might occur. FWER refers to the increase in the probability of type I error which is likely to violate the nominal significance level α when multiple populations are to be compared. To explain what FWER is, assume that 5 systems are compared with the significance level $\alpha = 0.05$. If it is desired to do all the pairwise comparisons, then there are $q = 5 \times 4/2 = 10$ hypotheses overall. For each of the null hypotheses, the probability of rejection without occurring the type I error is $1 - \alpha = 0.95$. For all comparisons, on the other hand, the probability of not having any type I error in all the hypotheses is $(0.95)^{10} = 0.6$. As a result, the probability of occurring at least one type I error increases to $1 - 0.6 = 0.4$, which is way higher than the nominal $\alpha = 0.05$. This phenomenon is the so-called family-wise error rate.

To prevent such an error, there are two primary approaches. Akin to the preceding example, the first approach is applicable when all the pairwise comparisons are desired. Conducting all pairwise comparisons is suitable for the comparison studies of the existing systems or for a competition like OAEI. Another approach to control FWER is convenient when a new alignment system is proposed and it is to be compared with other existing ones. In the interest of simplicity, the former approach is called $k \times k$ comparison and the latter is called $k \times 1$ comparison that will be discussed in the following.

4.5.1. CONTROLLING FWER: $k \times 1$ COMPARISON

When a new alignment system is proposed, it is usually compared with other existing ontology alignment systems. For comparing k systems (including the proposed one) in this case, $q = k - 1$ comparisons must be performed. There are four methods which can control the family-wise error rate in this case. These methods can be viewed as the p-value adjustment procedures which modify the p-values in a way that the adjusted p-values (APV) can be directly compared with the significance level, while the nominal

significance level is also preserved. Thus, a null hypothesis is rejected if its corresponding adjusted p-value is below the nominal α.

Let $H_i, i = 1,...,q$ be all hypotheses for comparing k systems and p_i be their corresponding p-values for $i = 1,...,q$. Bonferroni's method [29] is the most straightforward way to prevent FWER. In this procedure, all the p-values are compared with the nominal significance level α divided by the total number of comparisons. In other words, the hypothesis H_i is rejected if $p_i < \alpha/q$. Based on this equation, the adjusted p-value for the hypothesis H_i is obtained by multiplying both sides of the above inequality by q, i.e., $APV_i = \min\{q \times p_i, 1\}$. Thus, H_i is rejected if $APV_i < \alpha$. This procedure, though simple, is too conservative: It retains the hypotheses which must be rejected by generating higher APVs.

In contrary to single-step Bonferroni's correction, there are step-up and step-down procedures which sequentially reject the null hypotheses. It is necessary to order p-values for sequential rejective procedures and we denote the ordered p-values as $p_1 \leq p_2 \leq ... \leq p_q$ and their corresponding hypotheses as $H_1, H_2, ..., H_q$.

Holm's procedure [30] is a step-down method which starts with the most significant (or the smallest) p-value p_1. If $p_1 \leq \frac{\alpha}{q}$, then H_1 is rejected, and p_2 is compared with $\frac{\alpha}{q-1}$. If $p_2 \leq \frac{\alpha}{q-1}$, then H_2 is rejected, and p_3 is compared with $\frac{\alpha}{q-2}$. This procedure continues until a hypothesis is retained. In other words, each p_i in Holm's method is compared with $\frac{\alpha}{q+1-i}$, and it is rejected if it is below this value; otherwise, it is not rejected and the rest hypotheses are retained as well. Holm's adjusted p-value is $APV_i = \min\{v_i, 1\}$ where $v_i = max\{(q-j)p_j : 1 \leq j \leq i\}$.

Similar to Holm's procedure, Holland's correction [31] is also a step-down method. Instead of comparing the p-values with $\frac{\alpha}{q+1-i}$, it compares each p_i with $1-(1-\alpha)^{q-i}$. Thus, the adjusted p-value is $APV_i = \min\{v_i, 1\}$ where $v_i = max\{1-(1-p_j)^{q+1-j} : 1 \leq j \leq i\}$. Finner's procedure [32] is almost the same as Holland's technique and compares each p_i with $1-(1-\alpha)^{\frac{q}{i}}$. The Finner's adjusted p-value is $APV_i = \min\{v_i, 1\}$ where $v_i = \max\{1-(1-p_j)^{\frac{q}{j}} : 1 \leq j \leq i\}$.

Hochberg's method [33] as a step-up procedure works in the opposite direction and starts with the largest p-value. It compares the largest p-value with α, the next largest with $\alpha/2$ and it is terminated until a hypothesis is rejected. All the hypotheses with the smaller p-values are then rejected as well. Hochberg's adjusted p-value is $APV_i = \max\{(q-j)p_j : (q-1) \geq j \geq i\}$.

4.5.2. CONTROLLING FWER: $k \times k$ COMPARISON

For performing all the pairwise comparisons when k systems are available, there are $q = k(k-1)/2$ hypotheses overall. Nemenyi's method [34] is exactly Bonferroni's correction with q being assigned to $k \times k$ comparison. Thus, it has a high type II error which results in not detecting the difference among the population when there is a de facto difference. The same modification of q must be applied to other methods so that they are suitable for $k \times k$ comparison case.

There is also another sequential-rejective null hypothesis approach, which is only suitable for $k \times k$ comparison. This approach takes into account the logical relations

between hypotheses. Shaffer [35] discovered that Holm's procedure could be improved when hypotheses are logically interrelated. In many scenarios, it is not feasible to get any combination of true and false hypotheses. In the pairwise comparison, for instance, it is not possible to have $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ but $\mu_1 \neq \mu_3$. Thus, this case need not be protected against FWER.

Correction procedures which take into account the logical relations are similar to Holm's correction: They start with the most significant (or the smallest) p-value but compare it with α/t_1, where t_1 is the maximum number of hypotheses which can be retained at the first step. If $p_1 < \alpha/t_1$, then the corresponding hypothesis H_1 is rejected, and p_2 is compared with α/t_2. If H_2 is rejected, then p_3 is compared with α/t_3 and so on. The procedure terminates at the stage j if H_j cannot be rejected. The remaining hypotheses with bigger p-values than p_j are also retained. The adjusted p-value for the sequential corrective methods is $APV_i = \min\{v_i, 1\}$ where $v_i = \min\{t_i \times p_i, 1\}$.

There are two well-known techniques which consider the logical relations of hypotheses: Shaffer's and Bergmann's. These methods differ in their way to obtain the maximum number of true hypotheses at each level. Holm's procedure simply assigns the maximum number of true hypotheses at the stage j to the number of remaining hypotheses at the j^{th} stage, i.e. $t_j = q - j + 1$.
In Shaffer's method [35], the possible numbers for true hypothesis and consequently t_j is obtained by the following recursive formula

$$S(k) = \bigcup_{j=1}^{k} \{\binom{2}{j} + x : x \in S(k-j)\}$$

where $S(k)$ is the set of all possible numbers of true hypotheses when there are k alignments for comparison and $S(0) = S(1) = 0$. t_j is simply computed based on the set $S(k)$.

Similar to Shaffer's method, Bergmann's method [36] use the logical interrelations between the hypotheses but dynamically estimates the maximum number of true hypotheses at the stage j, given that $j-1$ hypotheses are rejected. To that end, they defined the exhaustive set which is an index set of hypotheses $I \subseteq \{1, ..., q\}$ where exactly all the hypotheses $H_j, j \in I$ can be true. For instance, let A_1, A_2, and A_3 be three alignments under study. If the null hypothesis between A_1 and A_2 is rejected, e.g., $A_1 \neq A_2$, then it is not possible that both hypothesis $A_1 = A_3$ and $A_2 = A_3$ be correct because the performance of A_3 cannot be the same as A_1 and A_2, while A_1 and A_2 have been already declared significantly different.

Having calculated the exhaustive set, any hypothesis H_j is rejected if $j \notin Z$, where Z is the acceptance set which is retained and defined as

$$Z = \bigcup \{\text{I: I exhaustive}, \min\{P_i : i \in I\} > \alpha/|I|\} \tag{4.13}$$

Bergmann's method is one of the most powerful procedures when $k \times k$ comparison is demanded, since it dynamically takes into account the logical relations of hypothesis. However, building the exhaustive set is time-consuming, especially if more than nine systems are available for comparison.

4.6. EXPERIMENTS

In this section, we first compare the statistical tests for comparing alignment systems in terms of their power and replicability, based on which various tests are recommended in different circumstances. Then, McNemar's test is applied to the systems participated in the OAEI *anatomy* track and the outcome is visualized by a directed graph. The tests for comparison of multiple systems are finally applied to the OAEI *benchmark* and *multifarm* tracks, and the corresponding results are visualized by the critical difference (CD) diagrams.

4.6.1. COMPARING STATISTICAL TESTS FOR ALIGNMENT COMPARISON: POWER AND REPLICABILITY

In this section, the statistical tests for comparing alignment systems are compared to each other in terms of their power and replicability. The power of a statistical test is formally defined as the probability of rejecting false null hypotheses. In reality, however, it is impossible to say if the null hypothesis is wrong in advance, making it impractical to gauge the power of statistical tests from the formal definition. Instead, there are two ways to estimate the power of a statistical test. First, the number of rejected null hypotheses in one thousand experiments are counted with a nominal significance level α. Another way is the average p-value in one thousand experiments; the lower the average is, the better the test will be.

For each way of the power estimation, there is a corresponding *replicability* measure. Bouckaert [37] defined the replicability as the probability that two experiments with the same pair of algorithms produce the same results. He estimated this probability as (in n experiments):

$$R(e) = \sum_{1 \le i \le j \le n} \frac{I(e_i = e_j)}{n(n-1)/2}, \tag{4.14}$$

where I is the indicator function, and e_i is the outcome of the i^{th} experiment (0 if the null hypothesis in the i^{th} experiment is rejected, and 1 otherwise.) If the hypothesis is accepted in u and rejected in v experiments, $R(e)$ can be easily computed as:

$$R(e) = \frac{u(u-1) + v(v-1)}{n(n-1)}. \tag{4.15}$$

Instead of using the number of rejected or retained hypotheses, Demšar [3] proposed a robust estimator based on the p-value obtained in each experiment. Demšar defined the replicability $R(p)$ as:

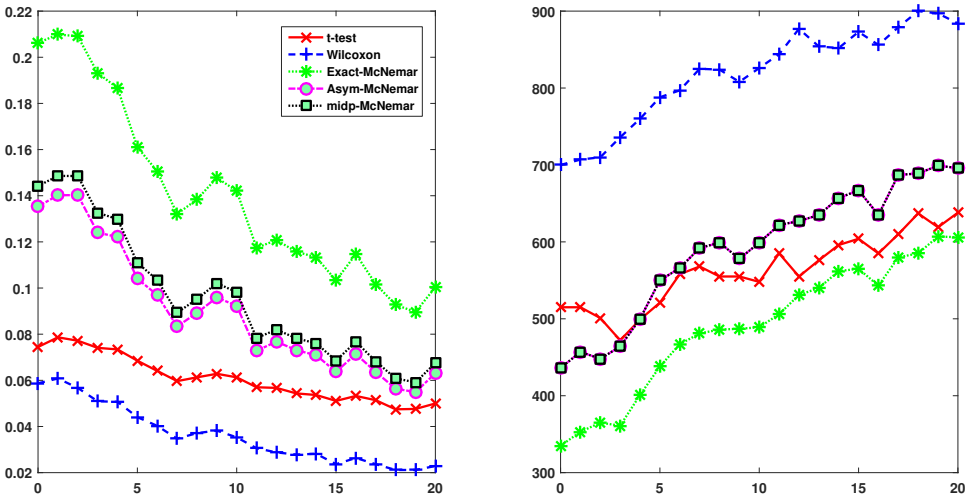
$$R(p) = 1 - 2var(p) = 1 - 2\frac{\sum_i (p_i - \hat{p})^2}{n-1}, \tag{4.16}$$

where $\hat{p}$ is the mean of the p-values and p_i is the p-value of the i^{th} experiment.

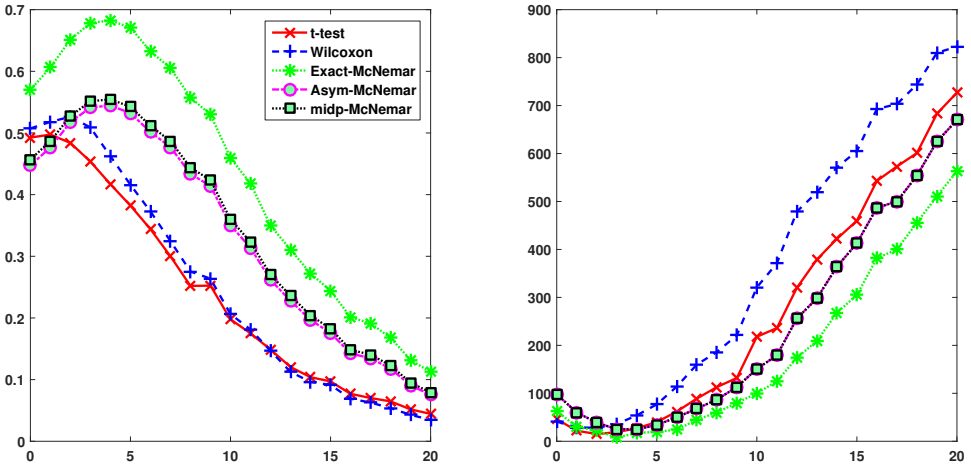
Since no single ontology matching system performs better than others in all scenarios [1, 2], it is usually the case that researchers would like to show the superiority of a system in one specific domain. In this case, there are some systems which perform better than others. To show this in simulation, some matching tasks are randomly selected

from the OAEI *benchmark* track so that the probability of selecting the i^{th} task is proportional to $1/(1+e^{-cd_i})$, where d_i is the difference between the performance scores and c is the bias [3]. For $c = 0$, the probability of selecting all tasks are the same. With higher values of c, it is more likely to pick the sets in favor of one system. This procedure is only considered for comparing the statistical tests, because doing such experiments with benchmarks chosen in favor of one system is, in one way or another, cheating.

First of all, 20 matching tasks are selected from the OAEI 2015 benchmark track with



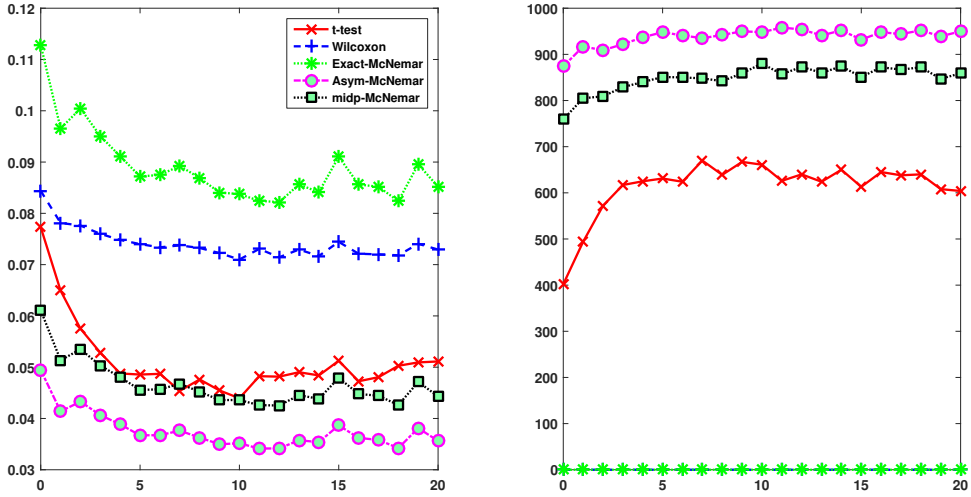
(a) AML2014 vs. AML



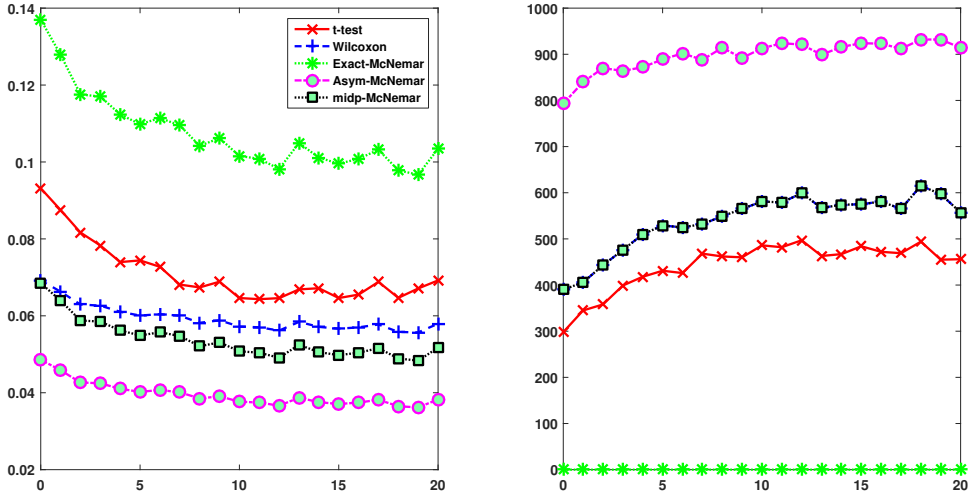
(b) AML vs. LogMap

Figure 4.1: Comparison of the paired t-test, Wilcoxon Signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests in terms of their power in 1,000 experiments over 20 matching tasks from the benchmark track. The x-axis is c and the y-axis is: (a) Left plots: The average p-value; (b) Right plots: The number of rejected null hypotheses.

the procedure mentioned above. The comparison is between the top two systems and two systems with mediocre performance so that various numbers of c will effectively change the selected tasks. Figure 4.1 plots the power estimation defined by the average p-value (left-hand side) and the number of rejected null hypotheses (right-hand side) in one thousand experiments for five statistical tests studied in this paper. The x-axis in all plots is c as defined above, and the y-axis is the average p-value for the left plots and the number of rejected hypotheses for the right ones. McNemar's test with continuity



(a) CroMatcher vs. AML



(b) GMap vs. Lily

Figure 4.2: Comparison of the paired t-test, Wilcoxon Signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests regarding their power in 1,000 experiments over five matching tasks. The x-axis is c and the y-axis is: (a) Left plots: The average p-value; (b) Right plots: The number of rejected null hypotheses.

correction is dismissed because there is no guarantee that its type I error be below the nominal significance level [16].

According to Figure 4.1, the average p-value of the Wilcoxon Signed-rank test is lower than or competitive with that of the paired t-test. This is probably due to the fact that the number of selected matching tasks is relatively high and presumptions of the paired t-test are likely to be satisfied through the *central limit theorem*. However, the number of rejected null-hypotheses in the Wilcoxon Signed-rank test is higher than that in the paired t-test in both cases. Therefore, we suggest using the Wilcoxon Signed-rank test when the comparison of two alignment systems is desired under this circumstance.

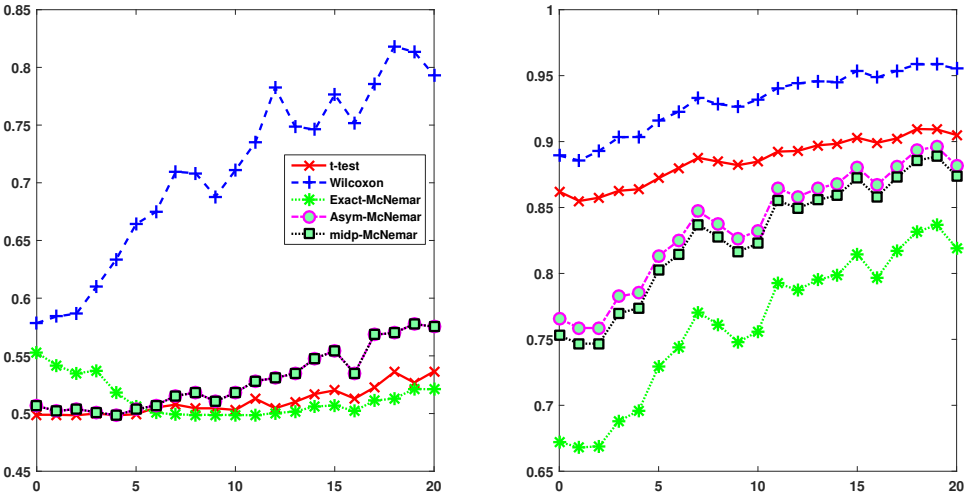
It can also be readily seen that McNemar's exact test (or the Sign test) is the most conservative one; thus, it should be avoided as a means of comparison. Another interesting point is that McNemar's mid-p and asymptotic tests are slightly different regarding the average p-values but almost the same with respect to the number of rejected null hypotheses. Further, these two tests are competitive with the paired t-test, especially in terms of the number of rejected null hypotheses. As McNemar's tests are non-parametric, their utilization is recommended as an alternative to the Wilcoxon Signed-rank test.

For the second scenario, five matching tasks are selected according to the above procedure. Figure 4.2 shows the power estimations when five matching tasks are selected, while the horizontal and vertical axes are the same as those in Figure 4.1. Interestingly, the power of the Wilcoxon Signed-rank test is less than McNemar's asymptotic and mid-p tests. McNemar's asymptotic test shows high power, especially in terms of rejected hypotheses. When few matching tasks or benchmarks are available, McNemar's asymptotic and mid-p tests are preferred over the Wilcoxon Signed-rank test.

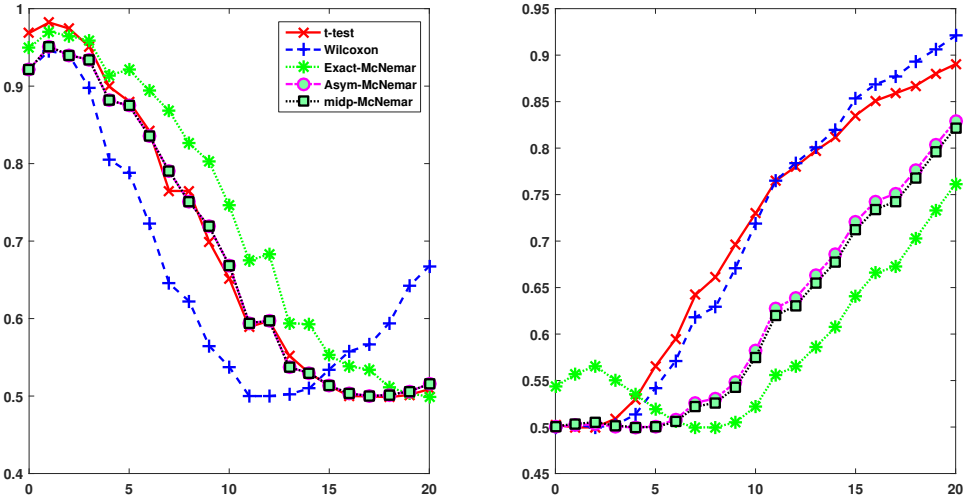
In addition to the power comparison, the statistical tests are compared with respect to the replicability. Figure 4.3 shows $R(e)$ on the right-hand side and $R(p)$ on the left-hand side when 20 matching tasks are selected. Interestingly, the results of two measures are in contradiction. The Wilcoxon Signed-rank test is (slightly) better than other tests regarding $R(p)$ and $R(e)$. While the paired t-test shows better replicability in terms of $R(e)$, McNemar's asymptotic and mid-p tests illustrate better performance in terms of $R(p)$.

For the case of selecting five matching tasks, McNemar's asymptotic test indicates a better replicability rate in terms of both perspectives, while the Wilcoxon Signed-rank test shows less replicability concerning both measures as shown in Figure 4.4. Another interesting point is the paradoxical replicability of McNemar's exact tests that is not able to reject any null hypothesis as can be observed from Figure 4.2, making the corresponding $R(e)$ equal to one in all scenarios. Regarding $R(p)$, on the other hand, its average p-values in one thousand experiments endorse their unreliability in comparison to other tests.

The final scenario is the case when the number of matching tasks or benchmarks is large enough, where we can verify the presumption of the paired t-test. We paired various systems together from benchmark and multifarm tracks and performed Jarque-Bera test [19] to check the normality assumption that is required for the paired t-test. Ironically, the normality assumption is held in less than 7%, making it safer to conduct the Wilcoxon Signed-rank test even if an enough number of benchmarks are available



(a) AML2014 vs. AML

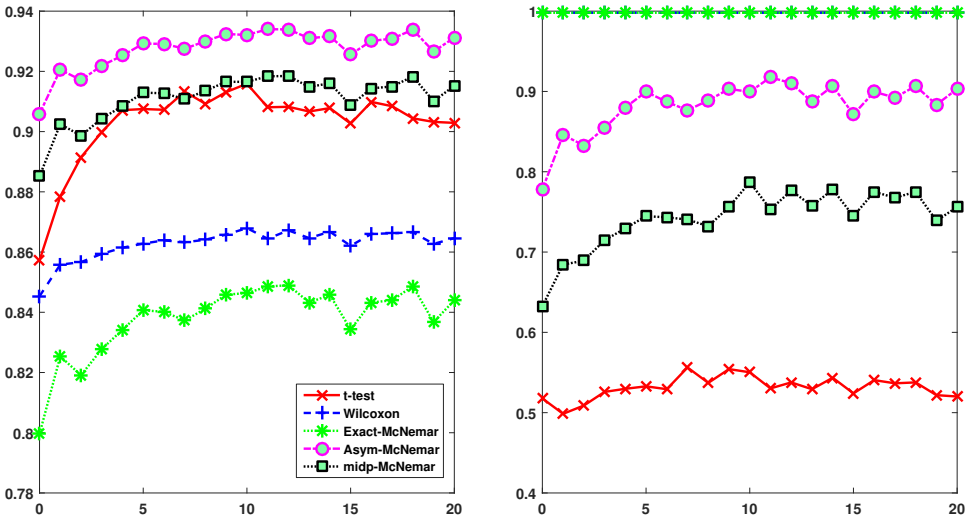


(b) AML2014 vs. LogMap

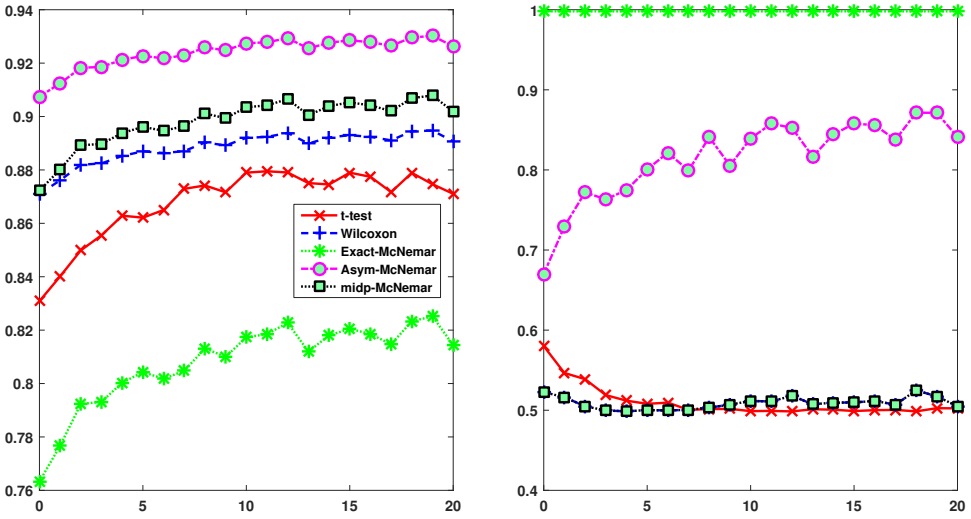
Figure 4.3: Comparison of the paired t-test, Wilcoxon Signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests in terms of replicability in 1,000 experiments over 20 matching tasks from the benchmark track. The x-axis is c, and the y-axis is: (a) Left plots: The replicability estimation $R(p)$; Right plots: The replicability estimation $R(e)$.

for comparison.

Table 4.8 tabulates the comparison of all pairs of systems with $c = 15$. The below diagonal numbers indicate the average p-value and the corresponding replicability measure $R(p)$, and the above diagonal entries show the number of rejected null hypotheses and the corresponding replicability measure $R(e)$. The average p-value of the Wilcoxon Signed-rank test is much lower than those of the other methods in almost all cases. It



(a) CroMatcher vs. AML



(b) GMap vs. Lily

Figure 4.4: Comparison of the paired t-test, Wilcoxon Signed-rank, and McNemar's (exact, asymptotic, and mid-p) tests concerning replicability in 1,000 experiments over five matching tasks from the benchmark track. The x-axis is *c* and the y-axis is: (a) Left plots: The replicability estimation *R(p)*; (b) Right plots: The replicability estimation *R(e)*.

is also recommendable by replicability measure *R(p)*, but *R(e)* prefers other tests with a p-value higher than a critical value of 0.05.

Table 4.8: Comparison of the paired t-test, Wilcoxon Signed-rank and McNemar's (asymptotic and mid-p) tests with $c = 15$. Below diagonal entries are the average p-value in 1,000 experiments and the corresponding $R(p)$; Above diagonal entries are the number of rejected null hypotheses in 1,000 experiments and the corresponding $R(e)$.

(a) Wilcoxon Signed-rank test

	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		621/0.53	1000/1	80/0.85	873/0.78	334/0.57	450/0.51	171/0.74
AML14	0.08/0.84		1000/1	283/0.59	1000/1	143/0.75	1000/1	885/0.80
CMtch	0.01/0.98	0.00/0.99		1000/1	47/0.91	1000/1	1000/1	1000/1
GMap	0.47/0.50	0.26/0.62	0.00/0.99		1000/1	941/0.89	1000/1	1000/1
Lily	0.02/0.95	0.00/0.99	0.48/0.50	0.00/0.99		1000/1	1000/1	1000/1
XMAP	0.19/0.69	0.36/0.54	0.00/0.99	0.01/0.97	0.00/0.99		998/0.99	945/0.90
LogMap	0.16/0.73	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.99		23/0.96
Mamba	0.34/0.55	0.02/0.96	0.00/0.99	0.00/0.99	0.00/0.99	0.01/0.99	0.53/0.50	

(b) Paired t-test

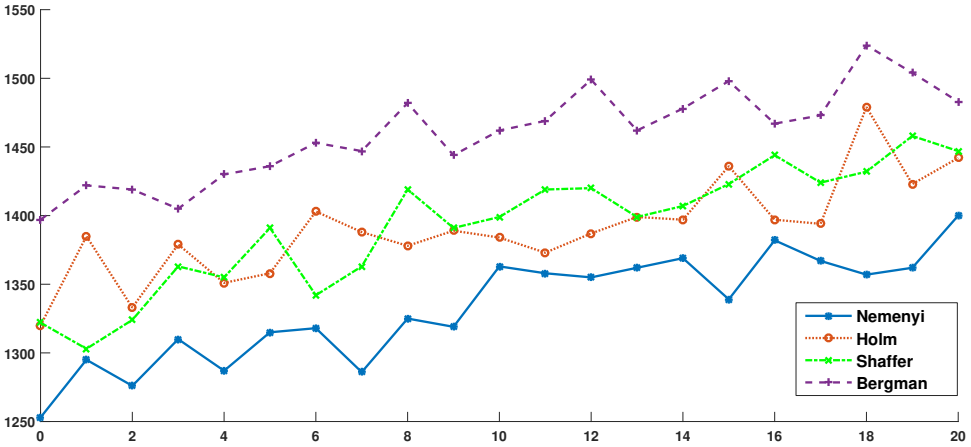
	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		213/0.66	934/0.88	36/0.93	589/0.52	272/0.60	442/0.51	155/0.74
AML14	0.16/0.72		987/0.97	132/0.77	1000/1	158/0.73	1000/1	911/0.84
CMtch	0.02/0.95	0.01/0.98		1000/1	106/0.81	1000/1	1000/1	1000/1
GMap	0.49/0.50	0.46/0.50	0.00 /0.99		1000/1	980/0.96	1000/1	1000/1
Lily	0.05/0.90	0.00/0.99	0.34/0.55	0.00/0.99		1000/1	1000/1	1000/1
XMA	0.25/0.62	0.38/0.53	0.00/0.99	0.01/0.98	0.00/0.99		1000/1	962/0.93
LogMap	0.14/0.76	0.00/0.99	0.00/1	0.00/0.99	0.00/1	0.00/0.99		27/0.95
Mamba	0.36/0.54	0.02/0.96	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.98	0.50/0.50	

(c) McNemar's mid-p test

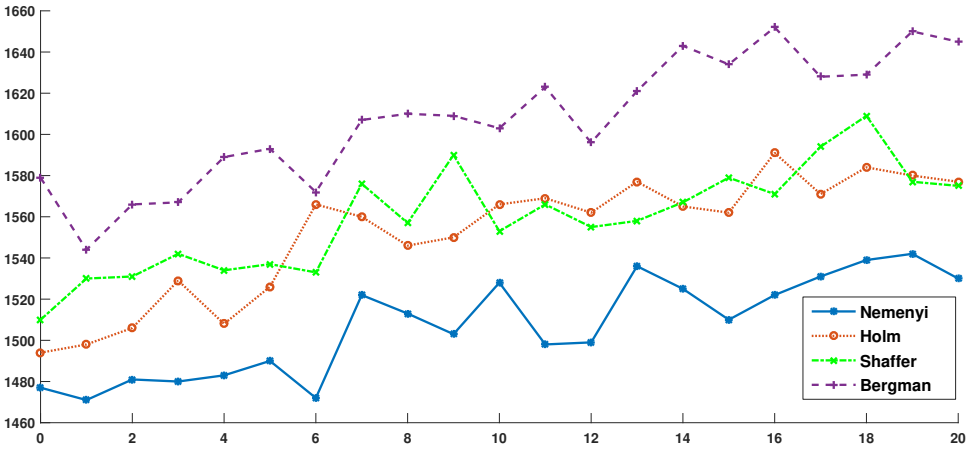
	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		213/0.66	934/0.88	36/0.93	589/0.51	272/0.60	442/0.51	155/0.74
AML14	0.16/0.73		987/0.97	132/0.77	1000/1	158/0.73	1000/1	911/0.84
CMtch	0.02/0.95	0.00/0.98		1000/1	106/0.81	1000/1	1000/1	1000/1
GMap	0.49/0.50	0.46/0.50	0.00/0.99		1000/1	980/0.96	1000/1	1000/1
Lily	0.05/0.90	0.00/0.99	0.33/0.55	0.00/0.99		1000/1	1000/1	1000/1
XMAP	0.25/0.62	0.38/0.53	0.00/0.99	0.01/0.98	0.00/0.99		1000/1	962/0.93
LogMap	0.14/0.76	0.00/0.99	0.00/1	0.00/0.99	0.00/1	0.00/0.99		27/0.95
Mamba	0.36/0.54	0.01/0.96	0.00/0.99	0.00/0.99	0.00/0.99	0.00/0.98	0.50/0.50	

(d) McNemar's asymptotic test

	edna	AML14	CMtch	GMap	Lily	XMAP	LogMap	Mamba
edna		619/0.53	1000/1	419/0.51	875/0.78	335/0.55	249/0.63	335/0.55
AML14	0.10/0.82		1000/1	523/0.50	1000/1	110/0.80	967/0.94	866/0.76
CMtch	0.00/0.98	0.00/0.99		1000/1	29/0.94	1000/1	1000/1	1000/1
GMap	0.17/0.71	0.11/0.80	0.00/1		1000/1	276/0.60	1000/1	1000/1
Lily	0.02/0.95	0.00/1	0.63/0.53	0.00/ 1		1000/1	1000/1	1000/1
XMAP	0.22/0.65	0.44/0.50	0.00/0.99	0.27/0.60	0.00/0.99		967/0.93	881/0.79
LogMap	0.29/0.58	0.01/0.98	0.00/1	0.00/0.99	0.00/1	0.01/0.98		10/0.98
Mamba	0.21/0.66	0.02/0.95	0.00/1	0.00/0.99	0.00/1	0.02/0.95	0.65/0.54	



(a) Comparison of multiple systems with selecting 10 benchmarks



(b) Comparison of multiple systems with selecting 40 benchmarks

Figure 4.5: The comparison of correction methods for the Friedman test in terms of the number of rejected null hypotheses for various numbers of c in x-axis; Two different scenarios: (a) Selection of 10 benchmarks; (b) Selection of 40 benchmarks.

PERFORMANCE OF STATISTICAL TESTS FOR COMPARING MULTIPLE SYSTEMS

In this section, the experiments across multiple alignment systems are studied. First, the power of various post-hoc procedures is reviewed and then the aforementioned multiple comparisons are applied to the OAEI *benchmark* and *multifarm* tracks and the corresponding results are reported.

Figure 4.5 shows the results over the *benchmark* track by the Friedman test and various post-hoc procedures. The x-axis in this figure is the parameter c and the y-axis is the overall number of the rejected hypotheses with respect to a correction method. Bergmann's correction performs better than other methods as its number of rejected null hypothesis is consistently outweigh the number of rejected hypotheses of the others. At

Table 4.9: n_{01} and n_{10} for constructing the contingency table from the first point of view which ignores the false positives(see Eq. (4.2)). For comparing the i^{th} and j^{th} systems, $n_{01} = (i, j)$ and $n_{10} = (j, i)$, where (i, j) is the element at the i^{th} row and the j^{th} column in the table.

	ALIN	AML	KEPLER	LogMap	LogMapLite	SANOM	WikiV3	XMap
Alin	0	2	0	4	3	10	14	0
AML	903	0	301	168	326	161	322	143
KEPLER	608	8	0	28	134	61	106	22
LogMap	766	29	182	0	180	73	213	53
LogMapLite	592	14	115	7	0	31	128	20
SANOM	782	32	225	83	214	0	239	77
WikiV3	610	17	94	47	135	63	0	37
XMap	788	30	202	79	219	93	229	0

Table 4.10: n_{01} and n_{10} for constructing the contingency table from the second point of view which takes into account the false positives (see Eq. (4.3)). For comparing the i^{th} and j^{th} systems, $n_{01} = (i, j)$ and $n_{10} = (j, i)$, where (i, j) is the element at the i^{th} row and j^{th} column in the table.

	ALIN	AML	KEPLER	LogMap	LogMapLite	SANOM	WikiV3	XMap
ALIN	0	74	48	121	47	166	160	103
AML	909	0	338	266	366	292	456	213
KEPLER	608	63	0	118	174	196	239	107
LogMap	766	76	203	0	184	176	346	127
LogMapLite	592	76	159	84	0	161	269	111
SANOM	783	74	253	148	233	0	370	140
WikiV3	616	77	135	157	180	209	0	121
XMap	794	69	238	173	257	214	356	0

the other extreme, Nemenyi's correction is the weakest method and must be ignored. Further, Holm's and Shaffer's methods are competitive with each other.

4.6.2. COMPARISON OF ALIGNMENT SYSTEMS

In this section, the alignment systems are compared by using the statistical tests. First, the anatomy track, which has only one benchmark, is considered, and systems are compared accordingly. Second, we consider *benchmark* and *multifarm* tracks for comparison of alignment systems over multiple benchmark. The following systems are selected for comparison in different tracks: Alin [38], AML and AML2014 [39], KEPLER [40], LogMap and LogMapLite [27], SANOM [41, 42], WikiV3 [43], XMap [28], MapPSO [44], CLONA [45], edna [20], CroMatcher [46], GMap [21], Lily [47], and Mamba [48].

ANATOMY TRACK

The recommended statistical procedures are applied to the OAEI *anatomy* track, and the corresponding results are reported. As an extra experiment, different string similarity metrics for the anatomy track are compared and ranked according to the number of correct discoveries.

We have two ways of obtaining the contingency table, four McNemar's statistics and

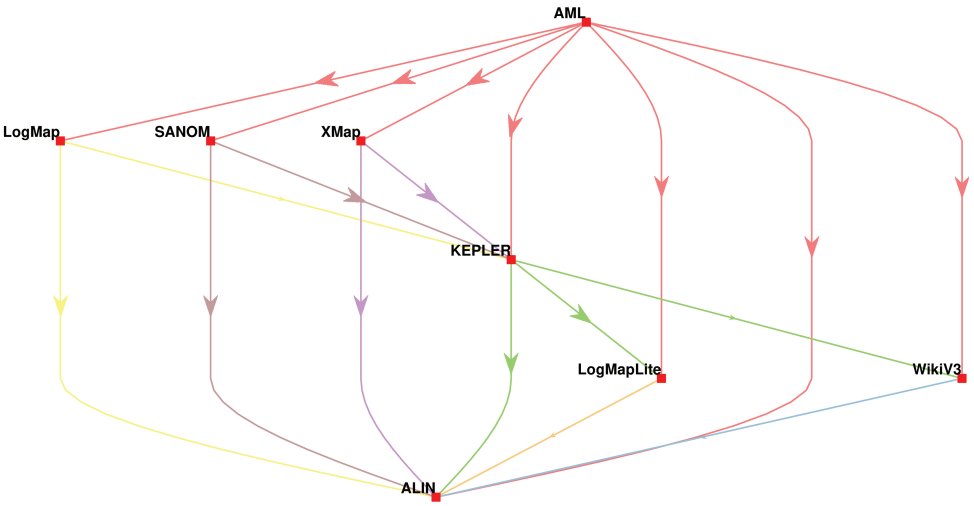


Figure 4.6: Comparison of alignment systems by McNemar's mid-p test with Nemenyi's correction while the false positive is ignored. The edge $A \rightarrow B$ indicates that A outperforms B.

four ways to prevent FWER. Therefore, there are totally 32 experiments for comparison. On account of simplicity (and probably for the exclusion of duplication), we only consider four experiments: the two ways of building the contingency table compared with McNemar's mid-p test and controlling FWER by Nemenyi's and Bergmann's correction techniques, the most conservative and the most powerful methods. The underlying reason behind the mid-p test selection is that it is not as conservative as the exact test and it is less likely to violate the nominal significance level α rather than the asymptotic test.

The contingency table is built by two foregoing methodologies. The values of n_{01} and n_{10} for the first and second way of table construction are arranged in Tables 4.9 and 4.10, respectively. For the interest of simplicity, n_{01} and n_{10} are tabulated in one single table for each perspective in below and upper diagonal entries, respectively. To compare the i^{th} and j^{th} systems in each approach, (i, j) and (j, i) elements of this table are taken as n_{01} and n_{10}, where (i, j) is the element at the i^{th} row and j^{th} column. For instance, let's compare $Alin$ and AML systems. In the first perspective, $n_{01} = 903$ means that there are 903 correspondences discovered by AML but not by Alin. And, $n_{10} = 2$ indicates that there are two correspondences identified by Alin but not by AML. In the second perspective, on the other hands, $n_{01} = 909$ and $n_{10} = 74$. Comparing with the previous view, n_{10} changes from 2 to 74 which means that AML has discovered 72 incorrect correspondences that are not in Alin. The little increase in n_{01} is due to the false positives of Alin ($909 - 903 = 6$ correspondences) in comparison to AML. As a result, it is evident that the false positive rate of Alin is less than that of AML, while the true positives of AML is much higher than those of Alin. If McNemar's test rejects the null hypothesis, AML is thus concluded to have better performance than Alin due to its higher true discovery rate. The comparison of other systems can be conducted likewise which clarifies the difference between the two perspectives.

We conduct all the pairwise comparisons and we take advantage of Nemenyi's and

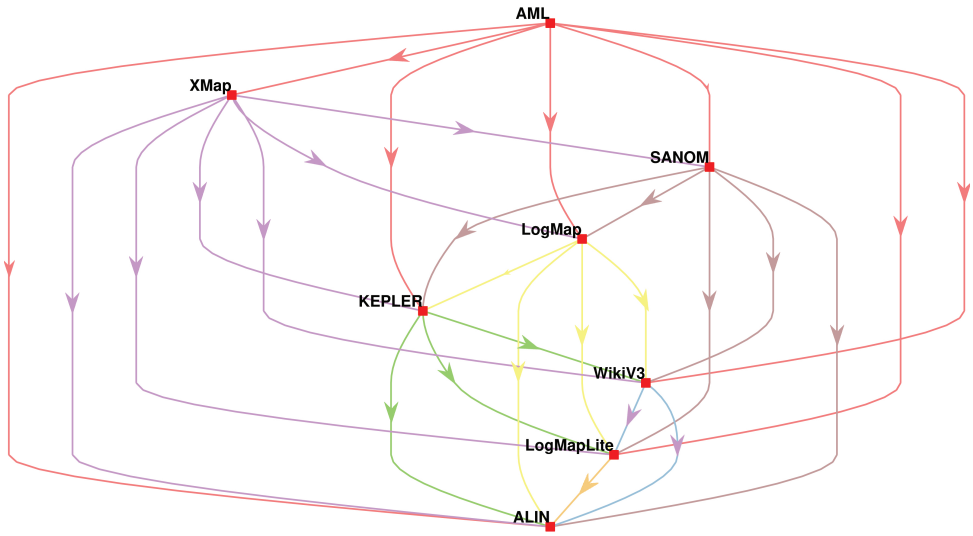


Figure 4.7: Comparison of alignment systems by McNemar's mid-p test with Bergmann's correction while the false positive is ignored. The edge $A \rightarrow B$ indicates that A outperforms B.

Bergmann's correction procedures, the most conservative and most powerful ones, to control the family-wise error rate. A directed graph visualizes the outcome of the pairwise comparison. Four different directed graphs correspond to each perspective and each correction method are displayed in Figures 4.6 - 4.9. The nodes in these graphs are the systems in question and any directed edge $A \rightarrow B$ means that A is significantly better than B. If there is no such an edge, however, there is no significant difference between the corresponding systems.

First, we compare the results obtained from Nemenyi's and Bergmann's correction techniques from each perspective of the contingency table construction. Figures 4.6 and 4.7 are the directed graphs that correspond to the pairwise comparisons of alignment systems obtained by applying Nemenyi's and Bergmann's correction based on the first perspective of contingency table construction, respectively. The results of these two correction methods are varied in several comparisons: Bergmann's correction indicates the significant difference between LogMap, SANOM, and XMap, while Nemenyi's correction cannot detect these differences. In addition, while Bergmann's correction declares LogMapLite and WikiV3 significantly different, Nemenyi's correction fails to detect their difference as significant. Thus, Bergmann's correction is more powerful than Nemenyi's method as the theory suggests. A similar argument holds true for the second way of contingency table construction by considering Figures 4.8-4.9.

We now compare two contingency table construction method with Bergmann's correction. Based on Figures 4.7 and 4.9, AML is the best system in both perspectives. However, when false positives are ignored, SANOM outperforms LogMap, while LogMap outperforms SANOM if false positives are considered. This shows that while the true positives of SANOM is much higher than that of LogMap, the latter system detects much lower false positives compared to the former. The higher false positives of SANOM com-

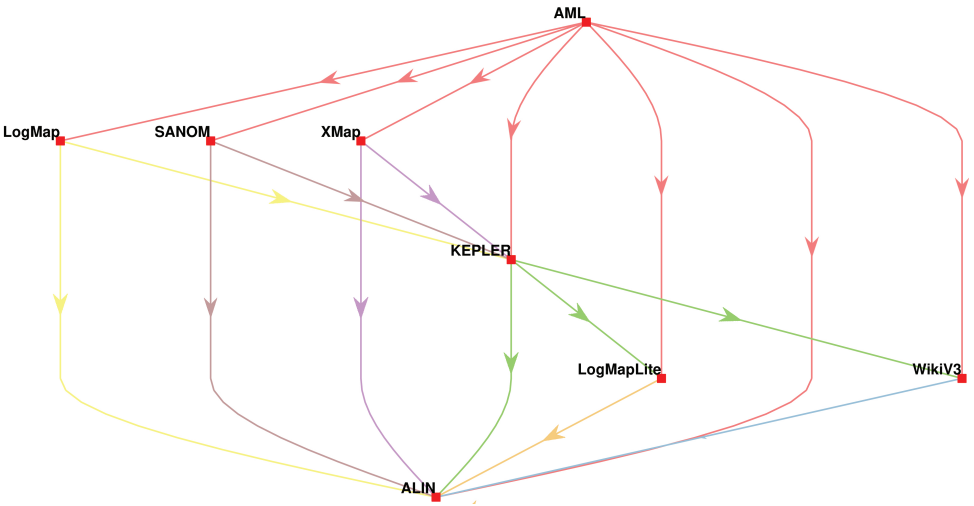


Figure 4.8: Comparison of alignment systems by McNemar's mid-p test with Nemenyi's correction while the false positive is considered. The edge $A \rightarrow B$ indicates that A outperforms B.

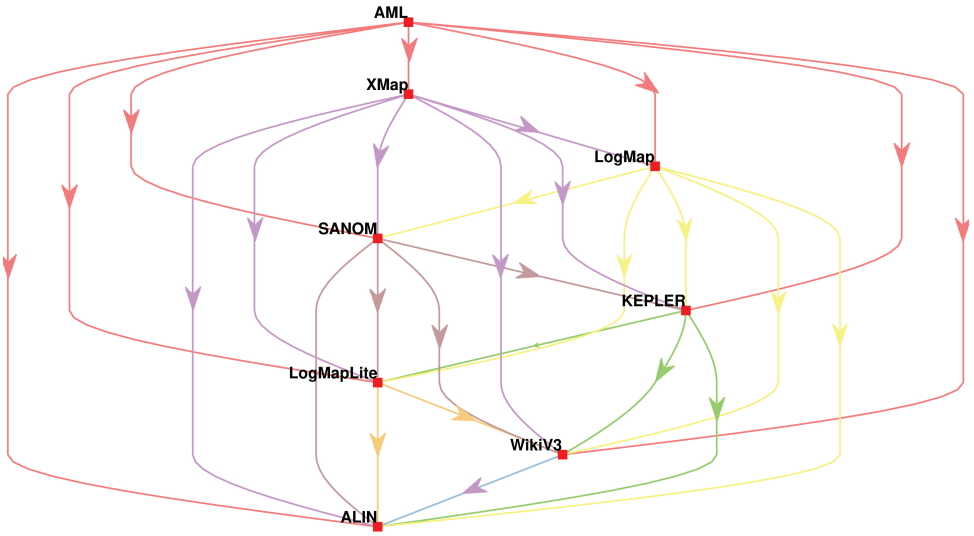


Figure 4.9: Comparison of alignment systems by McNemar's mid-p test with Bergmann's correction while the false positive is considered. The edge $A \rightarrow B$ indicates that A outperforms B.

pared to LogMap can be also seen by considering Tables 4.9 and 4.10. Another difference between the two perspective on contingency table construction is the place of WikiV3 and LogMapLite. If false positives are ignored, WikiV3 outperforms LogMapLite, since WikiV3 has 135 true positives that are not in LogMapLite, but LogMapLite has only 128. When the false positive is considered, LogMapLite has 269 true positives and negatives together that are not correctly detected by WikiV3, while this number is 180 for WikiV3

Table 4.11: Ranking of systems participated in the OAEI 2016 anatomy track from two different perspectives. The first perspective is to ignore the false positive (IFP) and the second is to consider it (CFP). The position of upper rows in this table indicates that it is significantly better than the methods coming in the lower rows. Cells with two methods indicate that the methods are not declared significantly different.

Rank	IFP	CFP
1	AML	AML
2	XMap	XMap
3	SANOM	LogMap
4	LogMap	SANOM
5	KEPLER	KEPLER
6	WikiV3 & LogMapLite	
7	LogMapLite	WikiV3
8	Alin	Alin

compared to LogMapLite. As a result, LogMapLite is declared significantly superior to WikiV3 if false positives are considered.

We rank the systems that participated in the OAEI anatomy track in Table 4.11 based on Bergmann's correction. The columns with labels IFP and CFP correspond to the contingency table construction with ignoring the false positives (IFP) and considering (CFP) it, respectively. In this table, the systems in higher rows are ones that are significantly better than the ones in the lower rows. If two systems are not significantly different, they are placed in the same cell. It can be readily seen that AML and Alin are the best and the worst systems from two perspectives, respectively.

For the final experiment, the string-based similarity measures are compared over the anatomy track. These metrics are of the utmost importance that can discover most of the correspondences of two given ontologies, including the ontologies in the anatomy track [49]. To compare such metrics over the anatomy track, we take advantage of the Shiva framework [50] which converts the ontology mapping into an assignment problem. In this framework, the similarity between each concept from the source ontology is gauged with all the concepts of the target ontology. The similarity score between the concepts of two ontologies constructs a matrix, which can be given to the Hungarian algorithm [51] to find the best match for each entity. We use nine string-based similarity measures to construct the matrix: Levenstein [52], N-gram [53], Hamming [54], Jaro [55], JaroWinkler [56], SMOA [57], NeedlemanWunsch2 [58], Substring distance [54], and equivalence measure. The Hungarian method applies to the resultant matrix to find the best match for each concept.

We consider the case when the false positive is not taken into account. The primary reason is that the selection of the appropriate string similarity measure can enable us to discover most of the potential correspondences [49]. If the right similarity metric is chosen, then the unreliable correspondences could be omitted by applying more strict thresholds.

Similar to the previous ones, Table 4.12 tabulates n_{01} and n_{10} corresponding to different string-based similarity measures while the false positive is ignored. The results are

Table 4.12: n_{01} and n_{10} for constructing the contingency table from the first point of view (ignoring the false positive) across various string-based similarity measures. For the comparison of the i^{th} and j^{th} metrics, $n_{01} = (i, j)$ and $n_{10} = (j, i)$ where (i, j) is the element at the i^{th} row and the j^{th} column in the table.

	Equal	Hamming	Jaro	JaroWinkler	Levenshtein	N-gram	Needleman.	SMOA	SubString
Equal	0	0	2	2	0	0	0	71	0
Hamming	842	0	51	51	32	54	48	258	494
Jaro	888	95	0	0	42	59	60	252	532
JaroWinkler	888	95	0	0	42	59	60	252	532
Levenshtein	966	156	122	122	0	64	50	277	593
N-gram	1041	253	214	214	139	0	174	290	636
Needleman.	932	138	106	106	16	65	0	276	573
SMOA	880	225	175	175	120	58	153	0	552
SubString	422	74	68	68	49	17	63	165	0

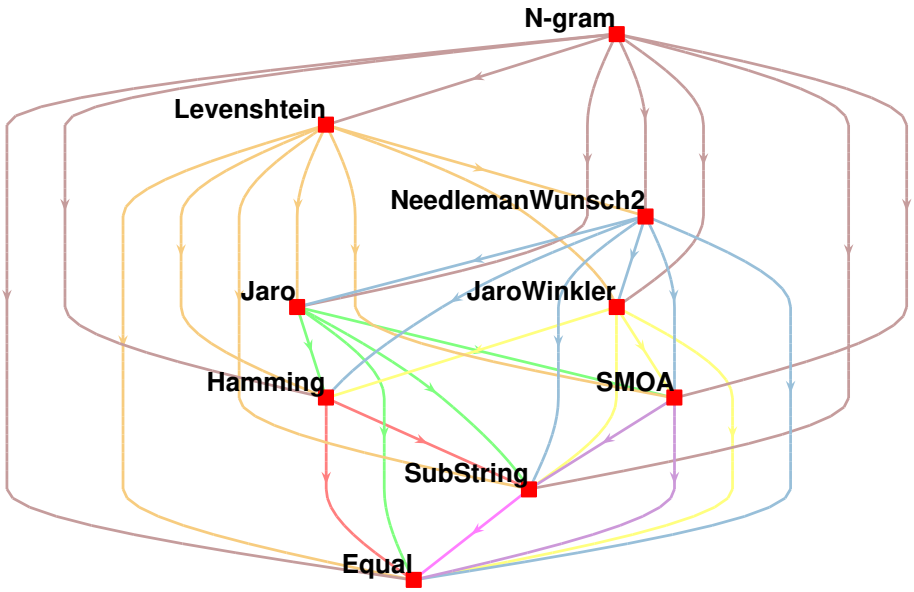


Figure 4.10: comparison of string-based similarity measures for the anatomy track. The arrow $A \rightarrow B$ indicates that A outperforms B.

visualized by a directed graph shown in Fig. 4.10. From this figure, N-gram has shown the best performances and is followed by Levenstein. Further, SMOA and Hamming distances are the ones with the least retrieved correspondences but they are better than Substring and Equivalence measures as expected.

BENCHMARK TRACK

In this section, we apply the statistical analysis to the results obtained from 94 benchmarks generated from the seed ontology *biblio*. The comparison is conducted based on the F-measure as it considers the both false positives and false negatives.

Table 4.13 tabulates the average ranks obtained by Friedman and Quade tests. The Friedman statistic is 385.73 with 7 degrees of freedom; thus the corresponding p-value is 1.8×10^{-10}. The Quade statistic (with (7,651) degrees of freedom) and its p-value are 91.60 and 1.22×10^{-92}, respectively. The null hypothesis which is the equivalence of alignment systems is rejected by both tests.

Tables 4.14 and 4.15 show the adjusted p-values obtained by various correction procedures for Friedman and Quade tests, respectively. Based on these tables, the rejected hypotheses can be simply discovered by the comparing the adjusted p-values with the nominal significance level α, while the FWER is inherently controlled. With $\alpha = 0.05$ and with the Friedman test, the first 18 hypotheses are rejected with Nemenyi's correction, while 19 hypotheses are rejected with more advanced methods.

In the Quade test, on the other hands, the first 12 hypotheses are rejected with all correction methods. As mentioned above, the Quade test is more suitable when fewer benchmarks are available. In the benchmark track, that contains *94* pairs of ontologies, the Friedman test is expected to be more powerful, as can be readily drawn from Tables 4.14 and 4.15. The sequential p-value adjustment methods reject the same number of hypotheses which means that they have the same power with respect to $R(e)$. From the $R(p)$ view, however, Bergmann's correction method is more powerful as it results in smaller adjusted p-values.

To better visualize and understand these results, Fig. 4.11 and 4.12 show the critical difference (CD) plot of the Friedman and Quade tests with various correction methods for $\alpha = 0.05$. The non-significant systems are connected to each other by a line. The results drawn from the table can be easily viewed from the CD diagrams as well. One difference between Nemenyi's and other sequential methods is the fixed critical difference in the former. It means that if the difference between any two methods is less than the critical difference, shown at the top of the plot, then they are not significantly different. This is the reason we distinguish the plot of Nemenyi's correction with the others.

The Quade test with four correction methods indicates that Lily and CroMatcher are together have better performance, and the remaining systems are not significantly dif-

Table 4.13: The average ranks of all systems computed by Friedman and Quade tests over the *benchmark* track.

Algorithm	Friedman	Quade
Lily	1.51	1.37
CroMatcher	1.81	1.75
GMap	4.35	4.29
XMap	4.78	5.18
AML2014	5.37	5.56
Mamba	5.68	5.42
edna	6.09	6.24
LogMapLite	6.41	6.18

Table 4.14: The adjusted p-values by four p-value adjustment methods across the OAEI 2015 *benchmark* track using the Friedman test.

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	Lily vs. LogMapLite	7.08×10^{-43}	1.98×10^{-41}	1.98×10^{-41}	1.98×10^{-41}	1.98×10^{-41}
2	CroMatcher vs. LogMapLite	7.30×10^{-38}	2.04×10^{-36}	1.97×10^{-36}	1.53×10^{-36}	1.53×10^{-36}
3	edna vs. Lily	8.85×10^{-38}	2.48×10^{-36}	2.30×10^{-36}	1.86×10^{-36}	1.86×10^{-36}
4	edna vs .CroMatcher	4.30×10^{-33}	1.20×10^{-31}	1.07×10^{-31}	9.02×10^{-32}	6.44×10^{-32}
5	Lily vs. Mamba	1.49×10^{-31}	4.18×10^{-30}	3.58×10^{-30}	3.14×10^{-30}	2.39×10^{-30}
6	CroMatcher vs. Mamba	2.68×10^{-27}	7.50×10^{-26}	6.16×10^{-26}	5.62×10^{-26}	2.94×10^{-26}
7	AML2014 vs. Lily	3.15×10^{-27}	8.81×10^{-26}	6.92×10^{-26}	6.61×10^{-26}	4.09×10^{-26}
8	AML2014 vs. CroMatcher	2.66×10^{-23}	7.45×10^{-22}	5.59×10^{-22}	5.59×10^{-22}	2.93×10^{-22}
9	Lily vs. XMap	5.40×10^{-20}	1.51×10^{-18}	1.08×10^{-18}	8.64×10^{-19}	7.02×10^{-19}
10	CroMatcher vs. XMap	1.11×10^{-16}	3.11×10^{-15}	2.11×10^{-15}	1.78×10^{-15}	1.22×10^{-15}
11	GMap vs. Lily	1.66×10^{-15}	4.64×10^{-14}	2.98×10^{-14}	2.65×10^{-14}	2.15×10^{-14}
12	CroMatcher vs. GMap	1.24×10^{-12}	3.46×10^{-11}	2.10×10^{-11}	1.98×10^{-11}	1.36×10^{-11}
13	GMap vs. LogMapLite	8.34×10^{-9}	2.34×10^{-7}	1.33×10^{-7}	1.33×10^{-7}	1.33×10^{-7}
14	edna vs. GMap	1.04×10^{-6}	2.93×10^{-5}	1.57×10^{-5}	1.57×10^{-5}	1.15×10^{-5}
15	XMap vs. LogMapLite	4.87×10^{-6}	1.37×10^{-4}	6.82×10^{-5}	6.33×10^{-5}	5.36×10^{-5}
16	GMap vs. Mamba	1.98×10^{-4}	0.0055	0.0025	0.0026	0.0016
17	edna vs. XMap	2.22×10^{-4}	0.0062	0.0027	0.0027	0.0016
18	AML2014 vs. LogMapLite	0.0035	0.099	0.039	0.039	0.028
19	AML2014 vs.GMap	0.0047	0.125	0.0446	0.0446	0.0312
20	XMap vs. Mamba	0.0114	0.318	0.102	0.102	0.056
21	LogMapLite vs. Mamba	0.041	1.159	0.331	0.331	0.206
22	edna vs. AML2014	0.041	1.159	0.331	0.331	0.207
23	AML2014 vs. XMap	0.098	2.756	0.590	0.590	0.295
24	GMap vs. XMap	0.233	1.00	1.00	1.00	0.93
25	edna vs. Mamba	0.245	1.00	1.00	1.00	0.934
26	edna vs. LogMapLite	0.38	1.00	1.00	1.00	1.00
27	AML2014 vs. Mamba	0.38	1.00	1.00	1.00	1.00
28	CroMatcher vs. Lily	0.388	1.00	1.00	1.00	1.00

ferent (with α = .05). The Friedman test also confirms the superiority of Lily and CroMatcher. With Nemenyi's correction, the Friedman test shows that Gmap, XMap, and AML2014 are not significantly different, while GMap indicates better performance in comparison to AML2014 by other sequential-based correction methods. Another difference between Nemenyi's correction and sequentially-corrected methods is the significant difference between AML2014 and LogMapLite: Nemenyi's correction cannot detect any difference between them, whereas they are significantly different when Holm's, Shaffer's, or Bergmann's correction is applied.

The results of this track are in line with the theory. First, Nemenyi's correction is so conservative and detect fewer differences among alignment systems. Further, the Friedman test has more power than the Quade test when a sufficient number of benchmarks is supplied.

Last but not least, the results of this section is compared with the averaging. The average of F-measure for Lily and CroMatcher systems, which are two top systems in the

Table 4.15: The adjusted p-values by four p-value adjustment methods across the OAEI 2015 *benchmark* track using the Quade test.

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	edna vs. Lily	2.65×10^{-10}	7.42×10^{-9}	7.42×10^{-9}	7.42×10^{-9}	7.42×10^{-9}
2	Lily vs. LogMapLite	4.41×10^{-10}	1.23×10^{-8}	1.1×10^{-8}	9.27×10^{-9}	9.27×10^{-9}
3	edna vs. CroMatcher	5.64×10^{-9}	1.58×10^{-7}	1.47×10^{-7}	1.18×10^{-7}	1.18×10^{-7}
4	CroMatcher vs. LogMapLite	9.04×10^{-9}	2.53×10^{-7}	2.26×10^{-7}	1.89×10^{-7}	1.36×10^{-7}
5	AML2014 vs. Lily	5.25×10^{-8}	1.47×10^{-6}	1.26×10^{-6}	1.10×10^{-6}	8.40×10^{-7}
6	Lily vs. Mamba	1.45×10^{-7}	4.06×10^{-6}	3.33×10^{-6}	3.04×10^{-6}	1.89×10^{-6}
7	AML2014 vs. CroMatcher	7.36×10^{-7}	2.06×10^{-5}	1.62×10^{-5}	1.54×10^{-5}	8.09×10^{-6}
8	Lily vs. XMap	7.60×10^{-7}	2.13×10^{-5}	1.62×10^{-5}	1.60×10^{-5}	9.88×10^{-6}
9	CroMatcher vs. Mamba	1.86×10^{-6}	5.21×10^{-5}	3.72×10^{-5}	2.98×10^{-5}	2.05×10^{-5}
10	CroMatcher vs. XMap	8.41×10^{-6}	2.36×10^{-4}	1.60×10^{-4}	1.35×10^{-4}	9.26×10^{-5}
11	GMap vs. Lily	1.48×10^{-4}	0.0041	0.0026	0.0023	0.0019
12	CroMatcher vs. GMap	9.57×10^{-4}	0.0267	0.0163	0.0153	0.0105
13	edna vs. GMap	0.011	0.325	0.186	0.186	0.186
14	GMap vs. LogMapLite	0.0144	0.405	0.217	0.217	0.186
15	AML2014 vs. GMap	0.099	1.00	1.00	1.00	0.793
16	GMap vs. Mamba	0.142	1.00	1.00	1.00	1.00
17	edna vs. XMap	0.170	1.00	1.00	1.00	1.00
18	XMap vs. LogMapLite	0.195	1.00	1.00	1.00	1.00
19	GMap vs. XMap	0.249	1.00	1.00	1.00	1.00
20	edna vs. Mamba	0.290	1.00	1.00	1.00	1.00
21	LogMapLite vs. Mamba	0.327	1.00	1.00	1.00	1.00
22	edna vs. AML2014	0.381	1.00	1.00	1.00	1.00
23	AML2014 vs. LogMapLite	0.426	1.00	1.00	1.00	1.00
24	AML2014 vs. XMap	0.619	1.00	1.00	1.00	1.00
25	CroMatcher vs. Lily	0.623	1.00	1.00	1.00	1.00
26	XMap vs. Mamba	0.754	1.00	1.00	1.00	1.00
27	AML2014 vs. Mamba	0.854	1.00	1.00	1.00	1.00
28	edna vs. LogMapLite	0.937	1.00	1.00	1.00	1.00

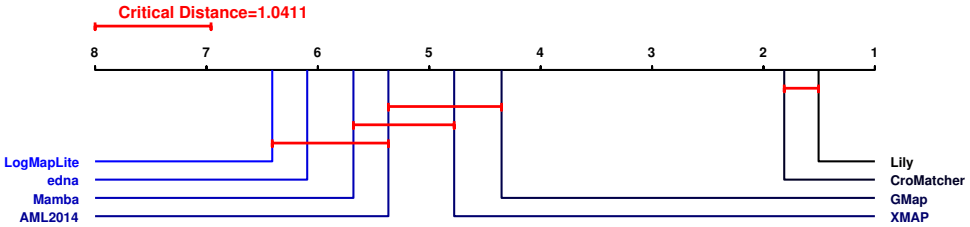
OAEI benchmark track, is 0.90 and 0.88, respectively. These are together the top systems from the statistical analysis as well. At the other extreme, edna and LogMapLite are the worse ones with the average of 0.41 and 0.46, respectively. Similarly, these systems are also the worst ones regarding the statistical analysis.

There are some small difference between the ranking of systems from averaging and the statistical analysis. For instance, AML2014 has a lower rank than Mamba from the statistical analysis, while the latter system is claimed to have outperformed the other one with respect to averaging. However, the major difference between averaging and the statistical analysis is that several differences are declared insignificant. This seems rational, since we cannot indicate the superiority of one system merely if its average is slightly higher than one another.

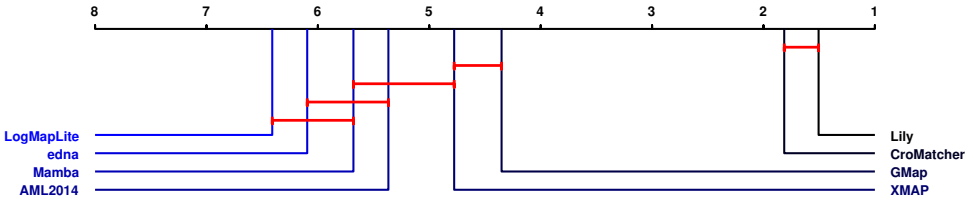
MULTIFARM TRACK

Another track in the OAEI which is considered here is *multifarm*, to which we apply the statistical procedures over F-measure obtained for each benchmark. The ranks computed by the Friedman and Quade tests for four systems are presented in Table 4.16.

The Friedman statistic (with 3 degrees of freedom) and its p-value are 98.80 and 5.80×10^{-11}, respectively. Similarly, the Quade statistic is computed as 138.30 with $(3, 46)$

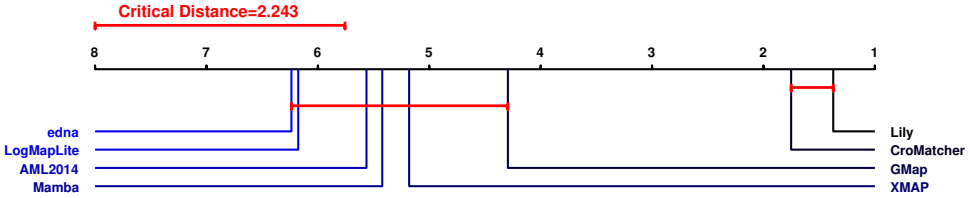


(a) Nemenyi's correction method

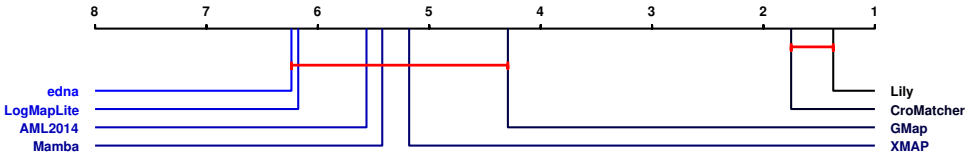


(b) Holm's, Shaffer's, and Bergmann's correction methods

Figure 4.11: The critical difference diagrams for the Friedman test with four p-value adjustment methods on the benchmark track: (a) Nemenyi's correction method, (b) Holm's, Shaffer's, and Bergmann's correction methods. The x-axis is the average rank of each system obtained by the Friedman test.



(a) Nemenyi's correction method



(b) Bergmann's correction

Figure 4.12: The critical difference diagrams for the Quade test with four p-value adjustment methods on the benchmark track: (a) Nemenyi's correction, (b) Holm's, Shaffer's, and Bergmann's correction. The x-axis is the average rank of each system obtained by the Quade test.

degrees of freedom, and the corresponding p-value is approximately zero. Thus, both tests reject the null hypothesis, and it is concluded that there is a significant difference among the systems.

The post-hoc procedure is applied to F-measure of the systems over the benchmarks in the *multifarm* track. The adjusted p-values of various post-hoc procedures are presented in Table 4.17. Based on this table, the systems that are significantly different from each other are simply detected, given the significance level α.

Table 4.16: Average Rankings of systems on the multifarm track computed by Friedman and Quade tests.

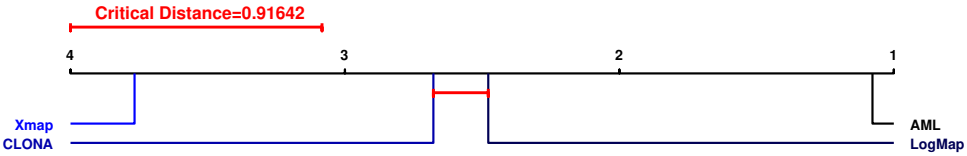
	AML	LogMap	CLONA	XMap
Friedman	1.07	2.48	2.68	3.77
Quade	1.05	2.51	2.56	3.88

Table 4.17: The adjusted p-values by four p-value adjustment methods on the multifarm track for the Friedman test.

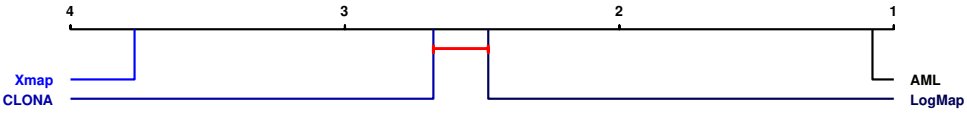
i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	AML vs. XMap	5.10×10^{-23}	3.06×10^{-22}	3.06×10^{-22}	3.06×10^{-22}	3.06×10^{-22}
2	AML vs. CLONA	4.13×10^{-9}	2.48×10^{-8}	2.05×10^{-8}	1.24×10^{-8}	1.24×10^{-8}
3	AML vs. LogMap	2.69×10^{-7}	1.61×10^{-6}	1.07×10^{-6}	8.07×10^{-7}	5.38×10^{-7}
4	LogMap vs. XMap	2.18×10^{-6}	1.31×10^{-5}	6.55×10^{-6}	6.55×10^{-6}	6.55×10^{-6}
5	CLONA vs. XMap	6.31×10^{-5}	3.79×10^{-4}	1.26×10^{-4}	1.26×10^{-4}	6.31×10^{-5}
6	CLONA vs. LogMap	0.462	1.00	0.462	0.462	0.462

Table 4.18: The adjusted p-values by four p-value adjustment methods on the multifarm track for the Quade test.

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	AML vs. XMap	1.52×10^{-13}	9.14×10^{-13}	9.14×10^{-13}	9.14×10^{-13}	9.14×10^{-13}
2	AML vs. CLONA	8.04×10^{-5}	4.83×10^{-4}	4.02×10^{-4}	2.41×10^{-4}	2.41×10^{-4}
3	AML vs. LogMap	1.28×10^{-4}	7.67×10^{-4}	5.10×10^{-4}	3.83×10^{-4}	2.55×10^{-4}
4	LogMap vs. XMap	3.79×10^{-4}	0.0022	0.0011	0.0011	0.0011
5	CLONA vs. XMap	5.77×10^{-4}	0.0034	0.0011	0.0011	0.0011
6	CLONA vs. LogMap	0.91	1.00	1.00	1.00	1.00



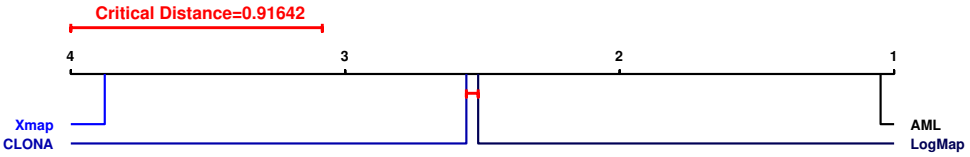
(a) Nemenyi's correction method



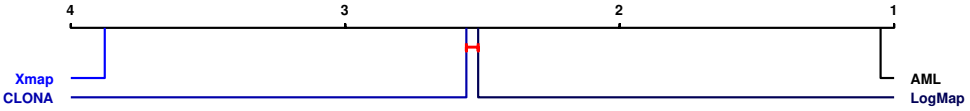
(b) Holm's, Shaffer's, and Bergmann's correction methods

Figure 4.13: The critical difference diagrams for the Friedman test with four p-value adjustment methods on the multifarm track: (a) Nemenyi's correction method; (b) Holm's, Shaffer's, and Bergmann's correction methods. The x-axis is the average rank of each system obtained by the Friedman test.

Similar to the *benchmark* track, we visualize the results obtained over this track. The critical difference diagrams of statistical tests with correction methods are plotted in Figures 4.13 and 4.14, where the x-axis indicates the average rank of each system obtained by Friedman and Quade tests. In this plot, the methods which are not significantly dif-



(a) Nemenyi's correction method



(b) Holm's, Shaffer's, and Bergmann's correction methods

Figure 4.14: The critical difference diagrams for the Quade test with four p-value adjustment methods on the multifarm track: (a) Nemenyi's correction method; (b) Holm's, Shaffer's, and Bergmann's correction methods. The x-axis is the average rank of each system obtained by the Quade test.

ferent are connected to each other by a line. The results of various tests over this track are the same. The Friedman and Quade tests with each method of correction indicate that AML is the best and XMap is the worst system. Further, CLONA and LogMap are not significantly different, but they have better performance than XMap and worse than AML.

4.7. CONCLUSION

The statistical methodologies for comparison of two or more alignment systems were studied in this chapter. McNemar's test is adopted for comparing alignment systems over one benchmarks. For comparison of two systems over multiple benchmarks, three different situations related to the number of benchmarks were considered and an appropriate test was recommended for each of the case. For comparison of multiple systems, the use of ANOVA was avoided due to its severe presumption *sphericity*. Instead, Friedman and Quade tests were proposed for comparison. For comparison of multiple systems, the family-wise error rate and the ways to prevent it are elaborated.

The recommendation for utilization of tests are summarized in Table 4.19 and can be explained as follows:

- If there is one benchmark for alignment, McNemar's test can be adopted for comparing two systems. If there are multiple systems for comparison, one can pair each two systems together and apply McNemar's test. In this case, the family-wise error rate must be controlled as well.

- For comparison of two systems with large enough benchmarks (> 30 benchmarks), the normality test is first conducted to check the normality of differences. If the normality assumption holds, the paired t-test is the most appropriate statistic. Otherwise, the Wilcoxon Signed-rank test is preferred.

- For comparison of two system with a moderate number of benchmarks (less than 30 but above 10), the test of normality is not reliable. Among the non-parametric

Table 4.19: The use of statistical test with respect to the number of benchmarks and the number of alignment systems to be compared.

# Alignment Systems	#Benchmarks ==1	< 10	< 30	> 30	
== 2	McNemar's test	McNemar's test	Wilcoxon Signed-rank test	Assumption satisfied? Yes	No
				Paired t-test	Wilcoxon Signed-rank test
> 2	Pairwise McNemar's test	Quade test	Friedman test	Assumption satisfied? Yes	No
				Repeated Measure ANOVA	Friedman Test

tests, the Wilcoxon Signed-rank test is preferred. In addition, if the number of benchmarks is less than ten, McNemar's asymptotic or mid-p tests are recommended.

- For the case of comparison among multiple systems, the repeated measures ANOVA is not recommended and its use must be avoided. Instead, Friedman and Quade tests are recommended for the moderate or large (more than 10) and the small (less than 10) number of benchmarks, respectively.

- For controlling FWER for comparing all systems together, Bergmann's correction is the most powerful one and is highly recommended. However, it takes a lot of time to conduct the comparison if there are more than 10 alignment systems. If there is any time restriction and there are more than 10 systems, Shaffer's correction is recommended which is powerful and fast. Nemenyi's correction is too conservative, and its use should be avoided.

The null hypothesis significance testing (NHST), though more efficient than averaging, suffers from several drawbacks that invalidate the results of the tests. As an instance, the decision is based on p-value, that is the probability of observing the alignment generated by systems given that the associated null hypothesis is correct. However, the desired probability is the likelihood of equivalence of two systems given their generated alignments. To overcome the shortcomings of NHST, Bayesian statistics should be used. In the next chapter, we discuss in detail the shortcomings of NHST as well as the ways that we can use the Bayesian statistics.

REFERENCES

[1] D. H. Wolpert and W. G. Macready, *No free lunch theorems for optimization,* IEEE transactions on evolutionary computation **1**, 67 (1997).

[2] D. H. Wolpert, *What the no free lunch theorems really mean; how to improve search algorithms,* in *Santa fe Institute Working Paper* (2012) p. 12.

[3] J. Demšar, *Statistical comparisons of classifiers over multiple data sets,* Journal of Machine learning research **7**, 1 (2006).

[4] S. Garcia and F. Herrera, *An extension on"statistical comparisons of classifiers over multiple data sets"for all pairwise comparisons,* Journal of Machine Learning Research **9**, 2677 (2008).

[5] S. García, A. Fernández, J. Luengo, and F. Herrera, *Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power,* Information Sciences **180**, 2044 (2010).

[6] D. A. Hull, *Information retrieval using statistical classification*, Ph.D. thesis, Citeseer (1994).

[7] T. G. Dietterich, *Approximate statistical tests for comparing supervised classification learning algorithms,* Neural computation **10**, 1895 (1998).

[8] B. T. NSKI, M. S. ETEK, Z. Telec, and T. Lasota, *Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms,* Int. J. Appl. Math. Comput. Sci **22**, 867 (2012).

[9] J. Derrac, S. García, D. Molina, and F. Herrera, *A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms,* Swarm and Evolutionary Computation **1**, 3 (2011).

[10] Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages,* Psychometrika **12**, 153 (1947).

[11] F. Wilcoxon, *Individual comparisons by ranking methods,* Biometrics bulletin **1**, 80 (1945).

[12] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures* (crc Press, 2003).

[13] M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance,* Journal of the american statistical association **32**, 675 (1937).

[14] D. Quade, *Using weighted rankings in the analysis of complete blocks with additive block effects,* Journal of the American Statistical Association **74**, 680 (1979).

[15] S. Suissa and J. J. Shuster, *The 2 x 2 matched-pairs trial: Exact unconditional design and analysis,* Biometrics , 361 (1991).

[16] M. W. Fagerland, S. Lydersen, and P. Laake, *The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional,* BMC medical research methodology **13**, 1 (2013).

[17] A. L. Edwards, *Note on the "correction for continuity" in testing the significance of the difference between correlated proportions,* Psychometrika **13**, 185 (1948).

[18] H. Lancaster, *Significance tests in discrete distributions,* Journal of the American Statistical Association **56**, 223 (1961).

[19] C. M. Jarque and A. K. Bera, *Efficient tests for normality, homoscedasticity and serial independence of regression residuals,* Economics letters **6**, 255 (1980).

[20] J. Euzenat, A. Ferrara, L. Hollink, A. Isaac, C. Joslyn, V. Malaisé, C. Meilicke, A. Nikolov, J. Pane, M. Sabou, *et al.*, *Results of the ontology alignment evaluation initiative 2009,* in *Proceedings of the 4th International Conference on Ontology Matching-Volume 551* (CEUR-WS. org, 2009) pp. 73–126.

[21] W. Li and Q. Sun, *Gmap: results for oaei 2015,* Ontology Matching **1**, 150 (2015).

[22] A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, and F. Ruggeri, *A bayesian wilcoxon signed-rank test based on the dirichlet process.* in *ICML* (2014) pp. 1026–1034.

[23] J. H. Drew, *Modern data analysis: A first course in applied statistics,* Technometrics **33**, 487 (1991).

[24] J. W. Mauchly, *Significance test for sphericity of a normal n-variate distribution,* The Annals of Mathematical Statistics **11**, 204 (1940).

[25] R. L. Iman and J. M. Davenport, *Approximations of the critical region of the fbietkan statistic,* Communications in Statistics-Theory and Methods **9**, 571 (1980).

[26] J. Hodges, E. L. Lehmann, *et al.*, *Rank methods for combination of independent experiments in analysis of variance,* The Annals of Mathematical Statistics **33**, 482 (1962).

[27] E. Jiménez-Ruiz and B. C. Grau, *Logmap: Logic-based and scalable ontology matching,* in *International Semantic Web Conference* (Springer, 2011) pp. 273–288.

[28] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, *et al.*, *Results of the ontology alignment evaluation initiative 2016,* in *OM: Ontology Matching* (No commercial editor., 2016) pp. 73–129.

[29] O. J. Dunn, *Multiple comparisons among means,* Journal of the American Statistical Association **56**, 52 (1961).

[30] S. Holm, *A simple sequentially rejective multiple test procedure,* Scandinavian journal of statistics , 65 (1979).

[31] B. S. Holland and M. D. Copenhaver, *An improved sequentially rejective bonferroni test procedure,* Biometrics , 417 (1987).

[32] H. Finner, *On a monotonicity problem in step-down multiple test procedures,* Journal of the American Statistical Association **88**, 920 (1993).

[33] Y. Hochberg, *A sharper bonferroni procedure for multiple tests of significance,* Biometrika **75**, 800 (1988).

[34] P. Nemenyi, *Distribution-free multiple comparisons*, Ph.D. thesis, Princeton University (1963).

[35] J. P. Shaffer, *Modified sequentially rejective multiple test procedures,* Journal of the American Statistical Association **81**, 826 (1986).

[36] B. Bergmann and G. Hommel, *Improvements of general multiple test procedures for redundant systems of hypotheses,* in *Multiple Hypothesenprüfung/Multiple Hypotheses Testing* (Springer, 1988) pp. 100–115.

[37] R. R. Bouckaert, *Estimating replicability of classifier learning experiments,* in *Proceedings of the twenty-first international conference on Machine learning* (ACM, 2004) p. 15.

[38] J. da Silva, F. A. Baiao, and K. Revoredo, *Alin results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 114.

[39] D. Faria, B. S. Balasubramani, V. R. Shivaprabhu, I. Mott, C. Pesquita, F. M. Couto, and I. F. Cruz, *Results of aml in oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 122.

[40] M. KACHROUDI, G. DIALLO, and S. B. YAHIA, *Oaei 2017 results of kepler,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 138.

[41] M. Mohammadi, A. Atashin, W. Hofman, and Y.-H. Tan, *Sanom results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 185.

[42] M. Mohammadi, W. Hofman, and Y. Tan, *Simulated annealing-based ontology matching,* (2018).

[43] S. Hertling, *Wikiv3 results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 190.

[44] J. Bock and J. Hettenhausen, *Discrete particle swarm optimisation for ontology alignment,* Information Sciences **192**, 152 (2012).

[45] M. El Abdi, H. Souid, M. Kachroudi, and S. B. Yahia, *Clona results for oaei 2015,* .

[46] M. Gulic and B. Vrdoljak, *Cromatcher-results for oaei 2013,* in *Proceedings of the 8th International Conference on Ontology Matching-Volume 1111* (CEUR-WS. org, 2013) pp. 117–122.

[47] P. Wang and B. Xu, *Lily: Ontology alignment results for oaei 2008,* in *Proceedings of the 3rd International Conference on Ontology Matching-Volume 431* (CEUR-WS. org, 2008) pp. 167–175.

[48] C. Meilicke, *Mamba-results for the oaei 2015,* Ontology Matching , 181 (2015).

[49] M. Cheatham and P. Hitzler, *String similarity metrics for ontology alignment,* in *International Semantic Web Conference* (Springer, 2013) pp. 294–309.

[50] I. Mathur, N. Joshi, H. Darbari, and A. Kumar, *Shiva: A framework for graph based ontology matching,* arXiv preprint arXiv:1403.7465 (2014).

[51] J. Munkres, *Algorithms for the assignment and transportation problems,* Journal of the society for industrial and applied mathematics **5**, 32 (1957).

[52] L. Yujian and L. Bo, *A normalized levenshtein distance metric,* IEEE transactions on pattern analysis and machine intelligence **29**, 1091 (2007).

[53] G. Kondrak, *N-gram similarity and distance,* in *International Symposium on String Processing and Information Retrieval* (Springer, 2005) pp. 115–126.

[54] J. Euzenat, P. Shvaiko, *et al.*, *Ontology matching*, Vol. 18 (Springer, 2007).

[55] M. A. Jaro, *Probabilistic linkage of large public health data files,* Statistics in medicine **14**, 491 (1995).

[56] W. E. Winkler, *The state of record linkage and current research problems,* in *Statistical Research Division, US Census Bureau* (Citeseer, 1999).

[57] G. Stoilos, G. Stamou, and S. Kollias, *A string metric for ontology alignment,* in *International Semantic Web Conference* (Springer, 2005) pp. 624–637.

[58] S. B. Needleman and C. D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins,* Journal of molecular biology **48**, 443 (1970).

5

BAYESIAN MODELS FOR ALIGNMENT EVALUATION AND COMPARISON

Intuition is a poor guide when facing probabilistic evidence.

Sir Dennis Lindley

The null hypothesis significance testing has been extensively studied in the previous chapter. However, the decision based on the null hypothesis testing is fallacious due to its several inherent drawbacks. Although these drawbacks can be addressed by using Bayesian statistics, the aim of this chapter is not to consider the Bayesian counterpart of the null hypothesis-based tests studied in the previous chapter, but to show that the evaluation based on scores like precision is also a frequentist inference. In this chapter, we introduce risk for an ontology alignment system, which is directly related to their errors that is defined as a function of false positives and false negatives. Based on the error definition, we develop a Bayesian model to estimate the risk of an alignment systems by a distribution (and not a score) based on the correspondences in alignments generated over single or multiple benchmarks. Based on the estimated risk, a Bayesian test is also devised to compute the extent to which one alignment system is preferred to another. We particularly study precision, recall, and F-measure, for each of which we compute a distribution for each alignment system by the proposed Bayesian model. The model presented in this chapter eliminates the drawbacks of the null hypothesis significance tests for both evaluation and comparison of alignment systems.

5.1. INTRODUCTION

In the previous chapter, the issues in the evaluation and comparison of alignment systems, which are based on two figures (performance scores or their averages), were discussed and the null hypothesis testing was considered to compare various alignment systems over single or multiple benchmarks. Although they can provide more evidence for comparison, in contrast to the sole comparison of scores, they suffer from various drawbacks. First, the inference is based on p-values, which is the probability of observing two alignments given the null hypothesis (i.e., the equivalence of two alignment systems) is correct. The decision based on p-values is fallacious, since the p-value is not the probability of interest, i.e., the probability of the null hypothesis given the alignments [1, 2].

In addition, the statement of significance using the null hypothesis testing would not necessarily mean that the alignment systems are significantly different in practice [3]. The null hypothesis will also be rejected if a large sample size, that is the number of benchmark in alignment comparison, is provided as well, regardless of the differences between systems. Thus, two alignment systems might be declared significantly different even though they have practically the same performance. Another break point of the null hypothesis testing is that it does not provide any information if the null hypothesis is not rejected [2], since retaining the null hypothesis does mean that it is accepted. In this case, one cannot claim any statement about the equivalence of two given alignment systems nor a significant difference among them. There is also no principled way to decide the value of significance level α, based on which a p-value would be claimed as significant [1].

Bayesian statistics can address the drawbacks of the null hypothesis significant testing. One straightforward way to use Bayesian statistics is to use the Bayesian counterpart tests to those introduced in the previous chapter. Recently, the Bayesian Wilcoxon Signed-rank [4] and Bayesian Friedman [5] tests are developed that can be used instead of using the frequentist tests. However, using Bayesian tests also need the performance scores of alignment systems over benchmarks, according to which they declare that the systems are significantly different. Instead, in this chapter, we first demonstrate that the ontology alignment evaluation is a statistical inference problem. In this regard, the notion of *risk* for an ontology alignment system is introduced, that is the probability that an alignment systems makes an *error*. The error is a function of false positives and false negatives that can be arbitrarily defined. In this chapter, we particularly focus on the definition of errors for precision, recall, and F-measure.

Since the alignment risk cannot be computed based merely on its definition, we first show that it follows a distribution and then use two statistical strategies, maximum likelihood estimation (MLE) and Bayesian estimation, to approximate it. We show that the MLE of risk with respect to a performance metric, e.g., precision risk, is equivalent to the complement of the same performance score, e.g., precision risk = 1 - precision. The byproduct of estimation of the precision risk, as a result, is that the precision is obtained as well. We also prove that the MLE of risk regarding a performance metric in the case that there are multiple benchmarks is tantamount to the complement of its micro-average. These results corroborate that the evaluation of alignment systems is indeed a statistical problem.

The underlying idea behind the MLE is that there is an unknown parameter which has a *precise* probability value, and the goal is to estimate that value in a way that it maximizes the odds of observing the data (here, the performance of alignments). The notion of having a precise probability is why it provides little information regarding the performance of alignment systems and is thus the source of pitfalls in the current practice of alignment evaluation and comparison. The Bayesian paradigm, on the other hand, would estimate the unknown parameter using a distribution, which is its crucial difference with the MLE. Approximation of the alignment risk using a distribution not only contains the MLE's precise value as its central tendency (e.g., mean, median, or mode), but it also takes into account the uncertainty that the observed performance might entail. A Bayesian model is developed to approximate the risk distributions in the presence of single or multiple datasets. Similar to the MLE, the estimation of the performance distribution could be easily obtained based on the estimation of the risk distribution, e.g., if the precision risk distribution is estimated, then the precision distribution is its complement.

As a result of the Bayesian model, we have a distribution with respect to each performance metric instead of having a score for representing the performance of an alignment. Such distributions take into account the uncertainty of the alignment system performance; hence, the precision of two alignment systems with the same ratio of true positives to true negatives would have different distributions if the number of true positive alters. Consider, for example, that two alignment systems have a precision of 0.5, while they discover four and a hundred correspondences, respectively. The reason that these two systems are deemed equivalent by the conventional evaluation, which we refer to as MLE in this chapter, gets back to the nature of the used statistical strategy, i.e., MLE. The Bayesian estimation, on the other hand, gives a probability distribution so that two systems with the same precision ratio (or any other metric than precision) would have totally distinct distributions if its number of correspondences is different.

In addition, a new Bayesian test is devised based on the estimated risk to compare different alignment systems. The test computes the probability that the performance of one alignment system is better than that of another hinged on their estimated risk. In particular, the probability that System A is superior to System B is the probability that the risk of System A is less than that of System B. The probability can be computed as the mathematical expectation of their risk differences. We can further use the region of practical equivalence (rope) [6], and consider that two alignment systems are identical if their risk difference is less than the rope length. The Bayesian test does not suffer from the pitfalls of decisions based on p-values, since it computes the probability of interests for inference, i.e., the probability that two systems have distinct performance given their alignments over single or multiple benchmarks. The Bayesian tests also avoid other pitfalls of the p-values. Another advantage of the proposed Bayesian model is that it can also be used for the evaluation, in contrast to the null hypothesis testing which can only be used for comparison.

In contrast to the Bayesian Wilcoxon Signed-rank and Friedman tests, for which we need to summarize the performance of alignment systems over each benchmark by a score and then applying the test, the proposed test takes all the correspondence from multiple benchmarks as the input, and calculate the overall performance without any

summarization. Therefore, the proposed test can better indicate the difference between given alignment systems rather than the tests in [1], in which the difference in average or median of performance scores is tested. Another drawback of such tests is that they cannot be applied to make the comparison on one benchmark, while the proposed test can be readily used for comparing systems on one benchmark as well. In addition, the proposed Bayesian model can be used for evaluation as well, while the current Bayesian models are developed solely for comparison. However, note that Bayesian Wilcoxon Signed-rank and Friedman tests can be used for comparison of alignment systems with respect to the scores that are not based on true positives and true negatives. For instance, the ontology alignment risk cannot be estimated for execution time, while we can apply other tests to verify if the difference between the execution time of alignment systems are significantly different.

Finally, we visualize the outcomes of the Bayesian analysis. Precision, recall, and F-measure distributions are displayed for evaluation, and the results of the Bayesian test for comparing alignment systems are visualized by a weighted directed graph. The proposed statistical analysis of alignment systems are applied to the OAEI anatomy and conference tracks, and the participating systems are evaluated and compared accordingly.

5.2. RISK OF ONTOLOGY ALIGNMENT SYSTEMS

The section begins by presenting the formal definition of the alignment risk. We then discuss the potential MLE and Bayesian estimation along with their advantages/pitfalls.

The risk is related to the error (i.e., false positives and false negatives) of a system, which can be seen as a complement to a performance metric. For instance, *silence* is the complement to recall; thus, the recall risk is indeed equivalent to silence. The following definition concisely presents the core of the alignment risk.

Definition 6 (Alignment Risk) *The risk of an ontology alignment system is the probability that the system makes an error.*

Definition 6 is broad enough to accommodate different performance metrics, since “error" can have distinct interpretations in different circumstances. We consider the error of a given alignment with respect to a performance metric. For instance, if precision is the sought metric, then the precision risk is the probability of having a false positive. If the comparison is based on recall, then the recall risk is the probability of having a false negative. F-measure would be a little intricate, but the F-measure risk could be defined as the probability of having a false positive or a false negative, which will be explained later on in this chapter.

The risk of an ontology alignment system is not an observed variable, but it is a parameter to be estimated based on the outcomes of a system over single or multiple benchmarks. The estimation of such a parameter would seem formidable at the beginning, but the following critical yet straightforward observation would pave the way for doing so.

For a moment, we focus on the estimation of the precision risk. Assume that we know the precision risk τ_{Pr} of an alignment system, hence the probability that one correspondence in a given alignment A is false would be τ_{Pr}. Besides, the probability of a

correspondence being true is $1-\tau_{Pr}$. As a result, it is a Bernoulli trial with the failure probability τ_{Pr} and success probability $1-\tau_{Pr}$. Thus, the probability of having K false positives out of N correspondences in the alignment would follow the binomial distribution.

Definition 7 *Given the alignment A with the risk τ, the probability of observing K errors out of N trials would follow a binomial distribution, i.e.,*

$$P(K,N;\tau) = \left(\begin{array}{c} N \\ K \end{array} \right) \tau^K (1-\tau)^{N-K}.$$

The number of errors would vary from one performance score to another. For precision, the number of trials is the number of correspondences in the alignment, i.e., N = $|A|$, and the number of errors is the false positives. For recall, on the other hand, the number of trials is the number of correspondences in the reference, i.e., $N = |R|$, and the number of errors is the false negatives.

For F-measure, we have:

$$\begin{aligned} Risk_F(A,R) &= 1 - \text{F-measure(A,R)} \\ &= 1 - \frac{2|A \cap R|}{|A| + |R|} \\ &= \frac{|A-R| + |R-A|}{|R| + |A|}. \end{aligned} \tag{5.1}$$

According to this equation, the number of trials for F-measure is the sum of correspondences in A and R, i.e., N = $|A| + |R|$, and the number of errors is the sum of false positives ($|A-R|$) and false negatives ($|R-A|$).

Having known the number of trials and errors, one can estimate the risk of an alignment based on Definition 7. A straightforward way of estimating the risk is to use the maximum likelihood estimation (MLE). The risk estimation using the MLE would be the fraction K/N, e.g.,

$$\tau = \frac{K}{N} \qquad 1-\tau = \frac{N-K}{N} \tag{5.2}$$

For the precision risk, for instance, $N-K$ is the number of true correspondences in the alignment and N is the total number of correspondences. Thus, $1-\tau$ is exactly the precision score. A Similar argument follows for recall and F-measure. The MLE also reveals the fact that any estimation would bear both precision and the precision risk, if the precision is the desired criterion. Thus, one can simply consider $1-\tau$ to compute directly the desired performance score, and not its risk.

The MLE in equation (5.2) is merely for one single benchmark. For multiple benchmarks, the micro-average is usually used, which is defined as:

$$\hat{Pr} = \frac{\sum_{i=1}^{q} |TP_i|}{\sum_{i=1}^{q} |A_i|}, \qquad \hat{Re} = \frac{\sum_{i=1}^{q} |TP_i|}{\sum_{i=1}^{q} |R_i|}, \tag{5.3}$$

where TP_i, A_i, and R_i are the true positives, identified alignment, and the reference of the i^{th} benchmark, respectively, and $\hat{Pr}$ and $\hat{Re}$ are the micro-average precision and recall. The following theorem proves that the MLE of risk for a specific metric on multiple benchmarks is equivalent to the complement of the micro-average of the same score.

Theorem 8 *Let S be the alignment system operated on q benchmarks, and the alignments $A_{1:q}$ are identified. The MLE of $1-\tau$ with respect to various scores is tantamount to micro-averaging of the same score over q benchmark.*

Proof. Let $R_{1:q}$ be the reference alignments with respect to q benchmarks and assume that the system S has independently discovered the alignments $A_{1:q}$. The MLE entails

$$arg\max_{\tau} log[p(\tau; A_{1:q}, R_{1:q})]$$

where log is the logarithm function, and $p(\tau; A_{1:q}, R_{1:q})$ is the likelihood of τ based on q benchmarks, and is defined as

$$p(\tau; A_{1:q}, R_{1:q}) = \prod_{i=1}^{q} \binom{N_i}{K_i} \tau^{K_i} (1-\tau)^{N_i - K_i}$$

where K_i and N_i are the numbers of errors and trials for the i^{th} alignment, respectively. It follows

$$\begin{aligned} & arg\max_{\tau} log\, p(\tau; A_{1:q}, R_{1:q}) \\ &= arg\max_{\tau} \quad \sum_{i=1}^{q} K_i log(\tau) + (N_i - K_i) log(1-\tau) \\ &= arg\max_{\tau} \quad log(\tau) \left(\sum_{i=1}^{q} K_i \right) + log(1-\tau) \left(\sum_{i=1}^{q} N_i - K_i \right). \end{aligned}$$

The point τ^* is the optimal value of the above minimization if and only if its derivative with respect to τ is zero. Thus,

$$\begin{aligned} & \frac{\partial}{\partial \tau} log(p) = 0 \Rightarrow \frac{\sum_{i=1}^{q} K_i}{\tau} - \frac{\sum_{i=1}^{q} N_i - K_i}{1-\tau} = 0 \\ & \Rightarrow \left(\sum_{i=1}^{q} K_i \right) \left(\frac{1}{\tau} + \frac{1}{1-\tau} \right) = \left(\sum_{i=1}^{q} N_i \right) \frac{1}{1-\tau} \\ & \Rightarrow \left(\sum_{i=1}^{q} K_i \right) \left(\frac{1}{\tau(1-\tau)} \right) = \left(\sum_{i=1}^{q} N_i \right) \frac{1}{1-\tau} \\ & \Rightarrow \tau = \frac{\sum_{i=1}^{q} K_i}{\sum_{i=1}^{q} N_i} \qquad \text{and} \quad 1-\tau = \frac{\sum_{i=1}^{q} N_i - K_i}{\sum_{i=1}^{q} N_i}. \end{aligned}$$

For precision, $N_i - K_i = |TP_i|$ and $N_i = |A_i|$, and for recall $N_i - K_i = |TP_i|$ and $N_i = |R_i|$. Similarly, the MLE of F-measure will follow. Thus, the MLE of $1-\tau$ with respect to a particular score over multiple benchmarks is precisely identical to the micro-average of the same score, and the proof is complete. □

So far, it is shown that the evaluation of alignment systems is a statistical inference problem, and the current evaluation using various performance scores could indeed obtain by the MLE, thanks to the notion of risk. It is further discussed that the pitfalls regarding the current evaluation approach are coming from the nature of the MLE.

In the MLE, the parameters of interest are assumed to be fixed but unknown, and the optimization procedure would find the optimal values as the precise point estimate. Thus, the evaluation and comparison using the MLE boil down to one figure for the former and the juxtaposition of two figures for the latter. In the Bayesian estimation, on the other hand, the parameters are not assumed to be fixed but rather a random variable. Thus, the outcome of the Bayesian analysis would result in a distribution instead of a point. Having such a distribution would enable us to take into account the uncertainty regarding the alignment system performance and compare various systems more meaningfully by inferring over the risk posterior distribution.

One Bayesian approach for the risk estimation is to use the beta-binomial conjugate. In this conjugate, the beta prior with parameters a and b, *beta(a,b)*, is considered, and the posterior for a given alignment with K errors out of N trials is computed as

$$p(\tau|N,K) = beta(a+K, b+N-K). \tag{5.4}$$

The mode of the posterior distribution is

$$Mode = \frac{a+K-1}{a+b+N-2}.$$

If the uninformative prior $beta(1,1)$ is selected, then the mode of beta-binomial would be equivalent to the MLE estimate, i.e., *Mode = K/N*. However, the variance of the beta distribution would be different for larger values of N and K. Such uncertainty is not reflected by the MLE. This simple example shows that the Bayesian estimation not only contains the MLE estimate as the central tendency, but also provides more information regarding the uncertainty of the alignment system performance.

The simple beta-binomial distribution would suffice if there were only one benchmark for evaluation. However, the generalization to multiple benchmarks cannot be performed by using merely this model. In the next section, we develop a Bayesian hierarchical model to estimate the risk based on the outcome of an alignment system across multiple benchmarks.

5.3. RISK APPROXIMATION: A BAYESIAN HIERARCHICAL MODEL

The risk of a system is a latent variable that must be approximated using a methodology. The MLE and a simple Bayesian model were discussed in the previous section, and their drawbacks were explained. In this section, we develop a Bayesian hierarchical model to estimate the risk of an alignment system for q benchmarks. Further, the model would estimate the final risk of an alignment system based on its risk over multiple benchmarks.

Assume that an ontology alignment system has been applied to q benchmarks, and we obtain N_i and K_i for $i = 1,...,q$. We show the set of all N_i and K_i obtained from q benchmarks as $N^{1:q}$ and $K^{1:q}$, respectively. The objective is to estimate the risk of a system over every benchmark, shown by $\hat{\tau}_i$, and the overall risk τ^*. Therefore, the Bayes rule follows

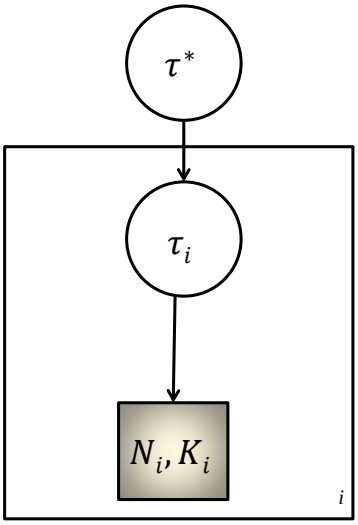


Figure 5.1: The graphical representation of the Bayesian model for estimating the risk.

$$P(\tau^{1:q}, \tau^* | N^{1:q}, K^{1:q}) \propto P(N^{1:q}, K^{1:q} | \tau^{1:q}, \tau^*) P(\tau^{1:q}, \tau^*)$$
$$= P(\tau^*) \prod_{i=1}^{q} P(N_i, K_i | \tau_i) P(\tau_i | \tau^*), \tag{5.5}$$

where the last equality holds true, since the results of different benchmarks are independent of each other. The graphical model associated to equation (5.5) is depicted in Figure 5.1. The rectangular shape denotes the observed variables, and the circles depict the random variables which needs to be approximated. As a convention, the variables τ_i, N_i and K_i are contained in a plate which means that the same model is repeated for different benchmarks.

We now need to specify the distribution of all elements in equation (5.5). So far, the number of errors has been modeled as the binomial distribution, i.e.,

$$K_i \sim binomial(\tau_i, N_i). \tag{5.6}$$

The parameter τ_i is unknown and must be estimated, hence we need to model it as another distribution. The τ_i distribution could be

$$\tau_i \sim beta(a, b), \tag{5.7}$$

where $beta(.,.)$ is the beta distribution, and a and b are its corresponding shape parameters. To make the model more meaningful, we reformulate the beta distribution with two other parameters. Let τ^* be the mean of this beta distribution and the concentration

parameter be $\gamma = a + b$, we have:

$$\begin{aligned} \tau^* &= \frac{a}{a+b} \quad \& \quad \gamma = a + b \\ \Rightarrow a &= \tau^* \gamma \quad \& \quad b = (\gamma - 1)\tau^* \\ \Rightarrow \tau_i &\sim beta\big(\tau^*\gamma, \tau^*(\gamma - 1)\big). \end{aligned} \tag{5.8}$$

Equation (5.8) means that the risk τ_i follows a beta distribution whose mean is τ^*. Thus, the values of τ_i are at the neighborhood of τ^*, and their proximity is controlled by the parameter γ. The parameters γ and τ^* are also unknown, hence we again model them as a distribution. According to equation (5.8), the values of γ must be greater than one, since the concentration parameter cannot be negative. There are many distributions for non-negative variables, and we use here the gamma distribution for γ as

$$\gamma - 1 \sim gamma(\alpha, \beta), \tag{5.9}$$

where *gamma* is the gamma distribution, and α_i and β_i are its shape and rate parameters, respectively. The τ^* is yet another parameter to be estimated. Thus, we model it as a beta distribution as well

$$\tau^* \sim beta(a^*, b^*). \tag{5.10}$$

The final step is to identify the remaining parameters. For the gamma distributions, we need to specify α_i, β_i, α^*, and β^*. The parameters can be stated in a way to be completely uninformative. However, we let the data learn the parameters. Thus, we model them as the uniform distribution

$$\alpha, \beta \sim uniform(l, u),$$

where $uniform(.,.)$ is the uniform distribution with the parameters l and u. We set $l = 0$ and $u = 1000$ to cover a broad spectrum of values.

Finally, the parameters a^* and b^* must be specified. We assign $a^* = b^* = 0.1$, since it is an uninformative prior distribution. The specified model should be solved using Markov-chain Monte Carlo (MCMC) techniques [7]. The model was written in JAGS [8], and the required sampling process was performed accordingly.

The Bayesian model has been intuitively developed based on the assumption that the risk of the alignment system on one benchmark is in the neighborhood of the overall alignment risk. We further validate the model by comparing distributions with the scores computed by the traditional way, i.e., the MLE. The experimental investigation supports the reasonable outcome of the approximated distributions, since the distributions are centered around the MLE in all cases.

5.4. A RISK-BASED BAYESIAN TEST

Having estimated the risk distributions of two alignment systems, it is also possible to compare their performance using a Bayesian test. The risk distributions allow us making the comparison more meaningfully, since we can compute the probability (or confidence) that one system is better than one another. Thus, the comparison is not drawn

based solely on the juxtaposition of two scores. Furthermore, we can define the region of practical equivalence (rope) to identify the systems with identical performance.

There is no principled way to determine the length of the rope, shown by r, and it is an expert decision to assess. The idea of the rope is quite simple: If the difference between posterior risk distributions of two alignment systems is less than r, then the systems in question are practically equivalent. Based on this notion, one can compute the probability that two systems are practically equal. If one is interested in determining the better systems even with a subtle difference, then $r = 0$ and the outcome of the test would indicate the superiority of one system over one another.

The probability that alignment A_1 with risk $\hat{\tau}_1^*$ is better than alignment A_2 with risk $\hat{\tau}_2^*$ can be computed as:

$$\begin{aligned} P(A_1 > A_2) &= P(\hat{\tau}_1^* < \hat{\tau}_2^*) \\ &= \int\int \mathscr{I}_{\hat{\tau}_2^* - \hat{\tau}_1^* > r} P(\hat{\tau}_1^* | data) P(\hat{\tau}_2^* | data) d\hat{\tau}_1^* d\hat{\tau}_2^*, \end{aligned} \tag{5.11}$$

5

where $P(\hat{\tau}_i^* | data)$ is the posterior risk distribution of the i^{th} system, and $\mathscr{I}$ returns one if the condition specified in its subscript is satisfied, and zero otherwise.

Similarly, one can compute the probability that A_2 is better than A_1 and the probability that they are equivalent as:

$$\begin{aligned} P(A_1 < A_2) &= \int\int \mathscr{I}_{\hat{\tau}_1^* - \hat{\tau}_2^* > r} P(\tau | \hat{data}_1^*) P(\hat{\tau}_2^* | data) d\hat{\tau}_1^* d\hat{\tau}_2^*, \\ P(A_1 = A_2) &= \int\int \mathscr{I}_{|\hat{\tau}_1^* - \hat{\tau}_2^*| < r} P(\tau | \hat{data}_1^*) P(\hat{\tau}_2^* | data) d\hat{\tau}_1^* d\hat{\tau}_2^*. \end{aligned} \tag{5.12}$$

The above-mentioned probabilities could also be obtained from the MCMC samples. As an instance, equation (5.11) is estimated by t samples of the MCMC chains as follows:

$$P(A_1 > A_2) = \frac{1}{t} \sum_{i=1}^{t} \mathscr{I}_{\hat{\tau}_2^{*i} - \hat{\tau}_1^{*i} > r}, \tag{5.13}$$

where $\hat{\tau}_j^{*i}$ is the i^{th} sample of $\hat{\tau}_j^*$ drawn by the MCMC, and $j = 1, 2$. Other probabilities are also computed in a similar way.

5.5. EXPERIMENTAL RESULTS

This section is dedicated to the experiments regarding the proposed Bayesian hierarchical model. We consider the results of conference and anatomy tracks from the OAEI to display the applicability of the Bayesian model.

The experiments on each track are twofold. The first one is the evaluation of each alignment system in which we display the distribution of precision, recall, and F-measure, and show that the obtained distributions are meaningful, since they are centered around the MLE. The second part is the comparison of alignment systems, where we apply the proposed Bayesian test and visualize the overall outcome by a weighted directed graph.

Table 5.1: Precision, recall, and F-measure of various systems on the OAEI anatomy track. The maximum likelihood estimation (MLE) is equivalent to that of the traditional way of reporting scores, and the other one is the mean of the distribution obtained by the proposed Bayesian hierarchical model (BHM).

System	Precision		F-measure		Recall	
	MLE	BHM	MLE	BHM	MLE	BHM
AML	0.95	0.95	0.943	0.943	0.936	0.936
XMap	0.926	0.925	0.893	0.893	0.863	0.862
KEPLER	0.958	0.951	0.836	0.833	0.741	0.741
LogMap	0.918	0.911	0.88	0.877	0.846	0.846
LogMapLite	0.962	0.954	0.829	0.826	0.728	0.728
SANOM	0.888	0.888	0.870	0.870	0.853	0.852
WikiV2	0.883	0.882	0.802	0.801	0.734	0.734
Alin	0.996	0.984	0.506	0.504	0.339	0.339

We use the Alignment API [9] to find the numbers required for the proposed hierarchical model. In particular, the numbers K and N for the precision risk of alignment A could be obtained by the Alignment API as:

$$K_{PR} = nbFound - nbCorrect, \qquad N_{PR} = nbFound, \tag{5.14}$$

where $nbFound = |A|$, $nbCorrect = |A \cap R|$, and the subscript PR represents the precision risk. The functions $nbCorrect$ and $nbFound$ in equation (5.14) are provided by functions with identical names in the Alignment API. Similarly, these numbers could be obtained for recall and F-measure as follows:

$$\begin{aligned} K_{RR} &= nbExpected - nbCorrect, \\ N_{RR} &= nbFound, \\ K_{FR} &= nbExpected + nbFound - 2 \times nbCorrect, \\ N_{FR} &= nbExpected + nbFound, \end{aligned}$$

where $nbExpected = |R|$, and subscripts RR and FR represent the recall risk and F-measure risk, respectively. We considered the results of two OAEI tracks and compared the participating systems together. The systems which were evaluated are Alin [10], AML [11], KEPLER [12], LogMap and LogMapLite [13], SANOM [14, 15], WikiV3 [16], and XMap [17].

5.5.1. ANATOMY TRACK

We now apply the Bayesian methodology to the anatomy track. On account of the clarity of results, the distribution of $1 - \tau$ was considered, since it could be directly related to the performance scores themselves. We refer to this distribution as the *performance distribution* of alignments as opposed to the risk distribution.

We first compare the outcomes obtained from the Bayesian model to those of the traditional way of reporting results. To this end, the means of the performance distributions were compared with performance scores. The traditional performance scores were referred to as the MLE, since we showed that they are the MLE of the risk (see Section 5.2).

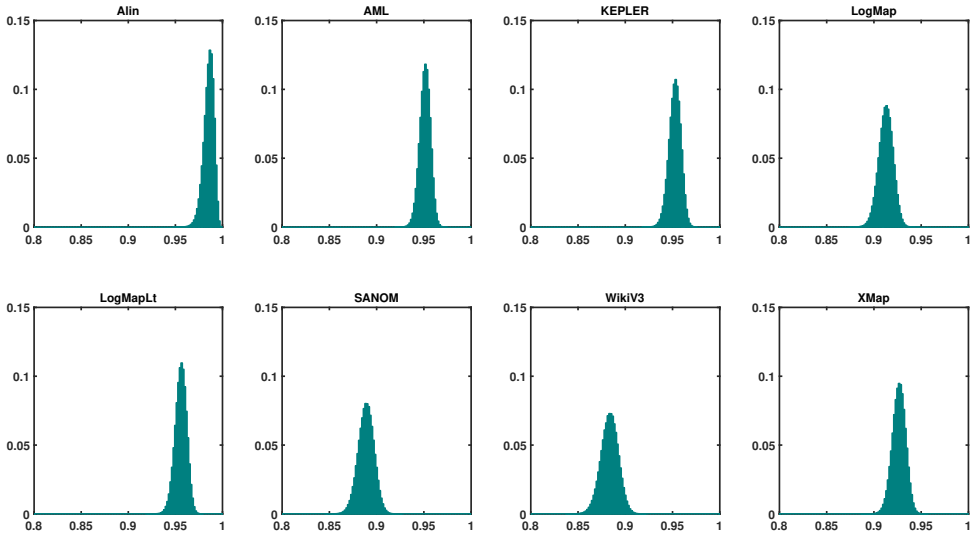


Figure 5.2: The estimation of the precision performance distribution $1-\tau$ of eight systems on the OAEI anatomy track using the Bayesian hierarchical model.

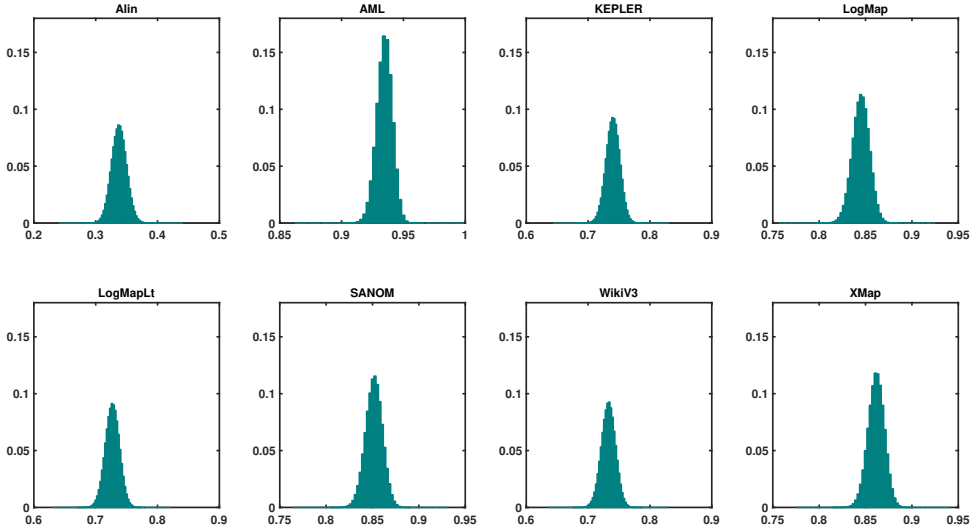


Figure 5.3: The estimation of the recall performance distribution $1-\tau$ of eight systems on the OAEI anatomy track using the Bayesian hierarchical model.

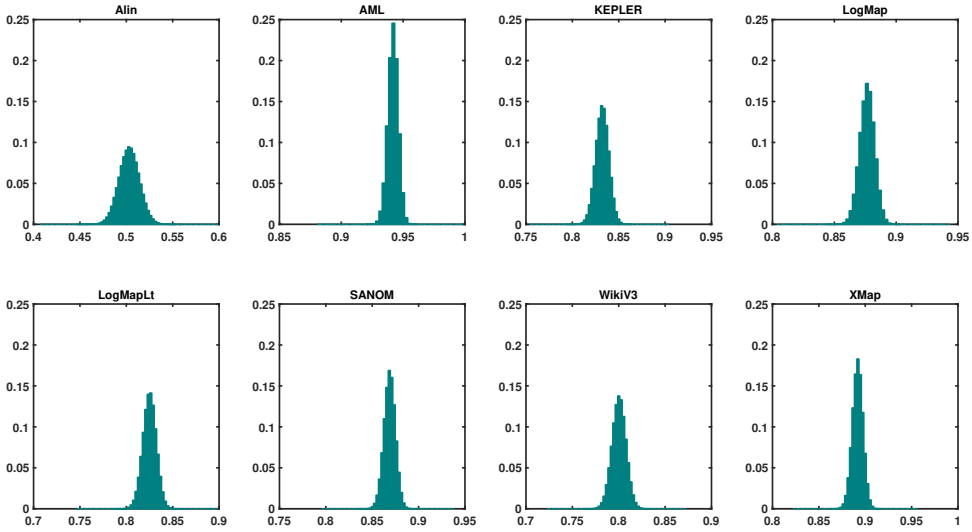


Figure 5.4: The estimation of the F-measure performance distribution $1-\tau$ of eight systems on the OAEI anatomy track using the Bayesian hierarchical model.

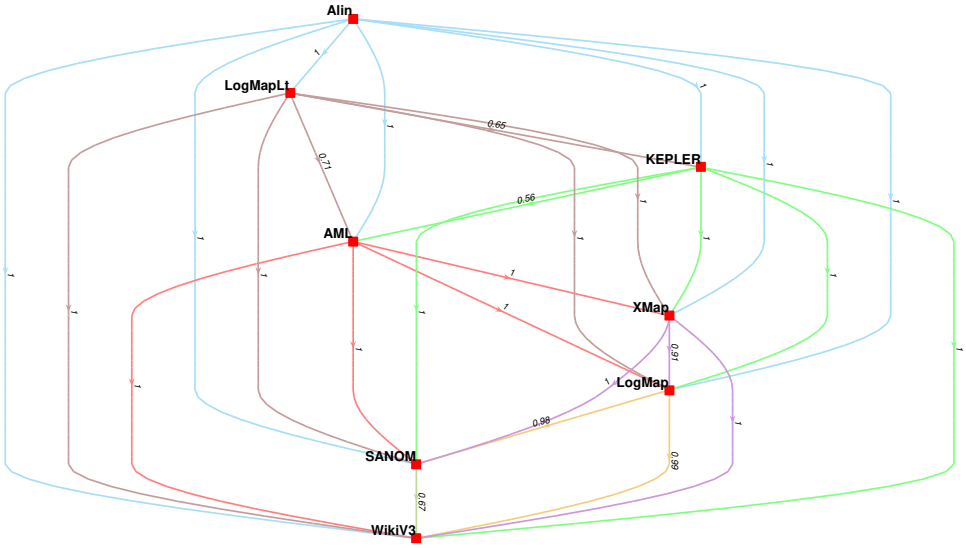


Figure 5.5: Comparison of eight alignment systems with respect to their precision on the OAEI anatomy track using the proposed Bayesian test.

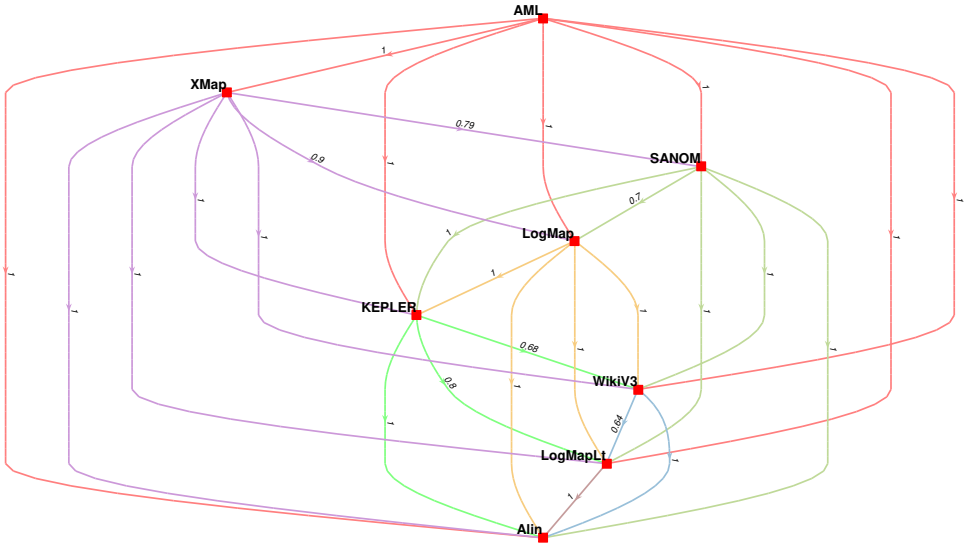


Figure 5.6: Comparison of eight alignment systems with respect to their recall on the OAEI anatomy track using the proposed Bayesian test.

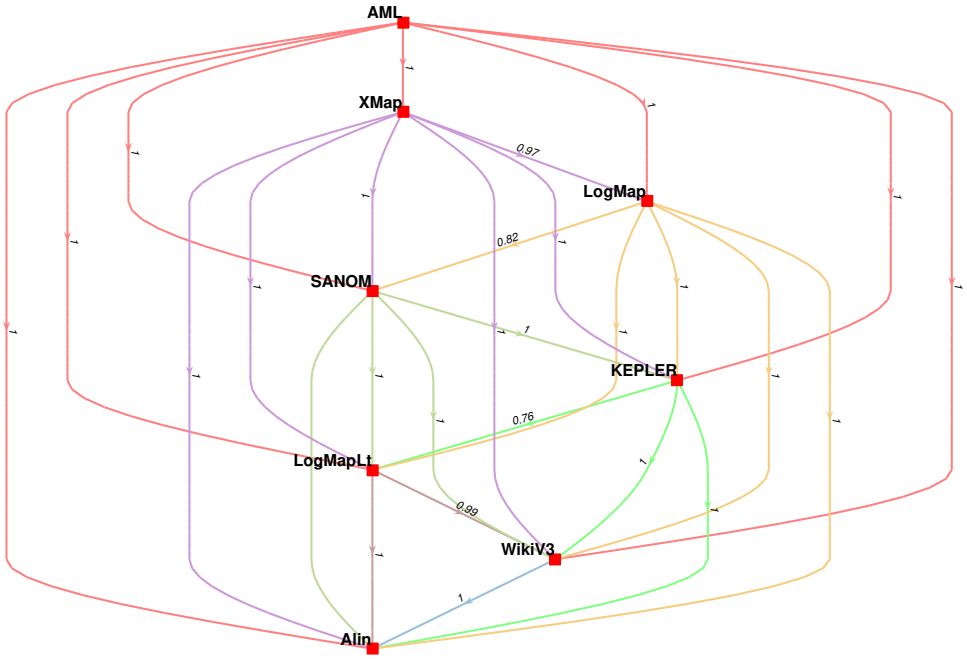


Figure 5.7: Comparison of eight alignment systems with respect to their F-measure on the OAEI anatomy track using the proposed Bayesian test.

Table 5.2: Comparison of alignment systems on the OAEI conference track. The performance scores with the subscript MLE are the scores obtained by the average over all benchmarks, while those with the subscript BHM are the means of the estimated distributions by the proposed Bayesian hierarchical model (BHM). We further tabulate the standard deviations (SD) of the performance metrics, that help us analyze the estimated distributions. The acronyms $\hat{Pr}$, $\hat{F}$, and $\hat{Re}$ stand for precision, F-measure, and recall, respectively.

	$\hat{Pr}_{MLE}$	SD	$\hat{Pr}_{BHM}$	$\hat{F}_{MLE}$	SD	$\hat{F}_{BHM}$	$\hat{Re}_{MLE}$	SD	$\hat{Re}_{BHM}$
ALin	0.93	0.228	0.933	0.43	0.105	0.418	0.29	0.173	0.271
AML	0.84	0.170	0.853	0.74	0.123	0.764	0.67	0.343	0.703
KEPLER	0.61	0.260	0.591	0.59	0.105	0.584	0.60	0.329	0.591
LogMap	0.84	0.197	0.841	0.68	0.118	0.692	0.59	0.317	0.585
LogMapLite	0.76	0.278	0.750	0.61	0.127	0.600	0.53	0.326	0.504
SANOM	0.74	0.188	0.730	0.72	0.085	0.728	0.73	0.307	0.771
Wiki3	0.69	0.221	0.682	0.58	0.124	0.576	0.52	0.349	0.501
XMap	0.86	0.200	0.865	0.69	0.122	0.699	0.59	0.338	0.590

Table 5.1 tabulates the MLE and the mean of Bayesian hierarchical model (BHM) estimation for each of the three performance metrics. This table proves that the MLE and the mean of the BHM estimation are very close to each other. Thus, the proposed model would yield the information provided by the traditional way of the evaluation.

The difference of two approaches, however, is that the BHM estimation would suggest more insights about the performance of the systems in question. In particular, we plot the *performance* distributions for each of the scores. Figures 5.2-5.4 display the performance distributions of precision, recall, and F-measure, respectively. It is readily seen that the peaks of the distributions are over the corresponding MLE with some variations.

We further compare the systems over the anatomy track using the Bayesian test introduced in Section 5.4. For each pair of systems, the probability of one system being superior to another is computed with the size of rope equals to zero. Thus, the equivalence of two systems is not considered in this experiment.

The comparison is drawn from three points of view, each related to precision, recall, and F-measure. Figures 5.5-5.7 are the weighted directed graphs demonstrating the outcomes of comparison. The nodes in these graphs are the systems in question, and each edge $A \xrightarrow{w} B$ means that A is superior to B with the probability w.

Based on Figure 5.5, Alin is the best system in terms of precision, followed by LogMapLite and KEPLER. At the other extreme, SANOM and WikiV3 have poor performance concerning precision. Regarding recall, however, AML, XMap, and SANOM are the systems with superior performance, thanks to Figure 5.6. In contrast to precision, Alin has poor performance with respect to recall.

As a combination of both precision and recall, one can compare the systems in terms of F-measure using Figure 5.7. From this figure, one can realize that the overall performance of AML and XMap are superior, followed by LogMap and SANOM.

5.5.2. CONFERENCE TRACK

In the OAEI conference track, there are usually 21 mapping tasks for matching seven ontologies together. The evaluation and comparison of the OAEI conference track are

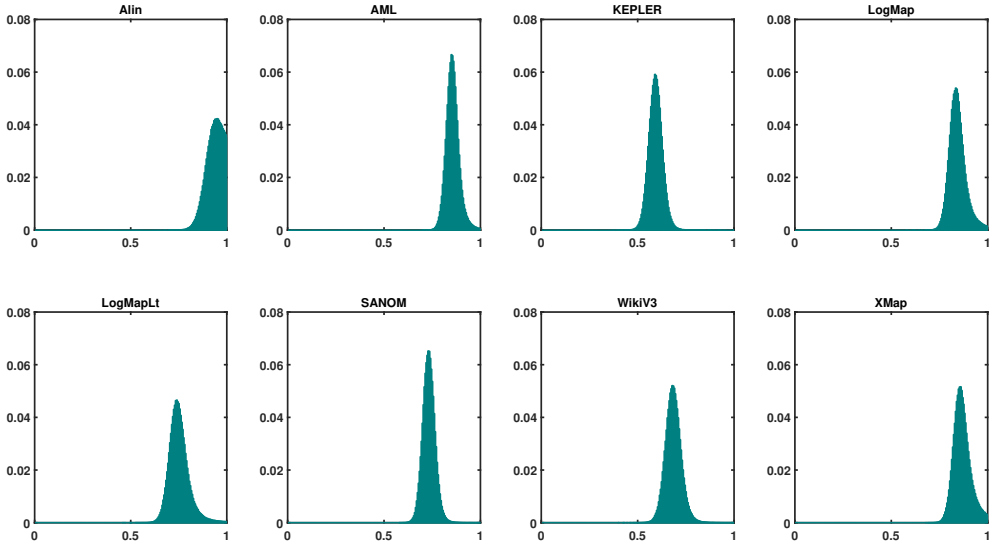


Figure 5.8: The estimation of the precision distribution $1-\tau$ of systems on the OAEI conference track using the proposed Bayesian hierarchical model.

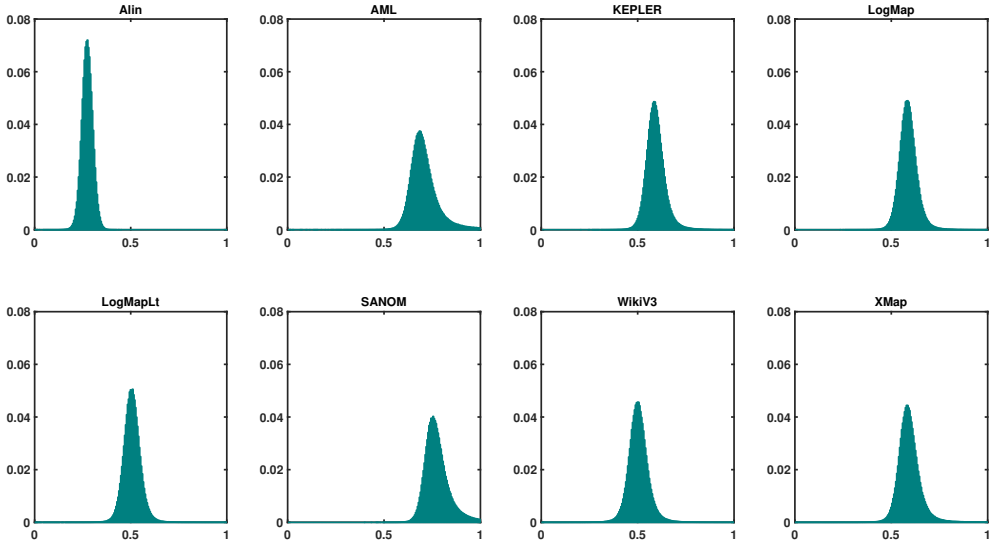


Figure 5.9: The estimation of the recall distribution $1-\tau$ of systems on the OAEI conference track using the proposed Bayesian hierarchical model.

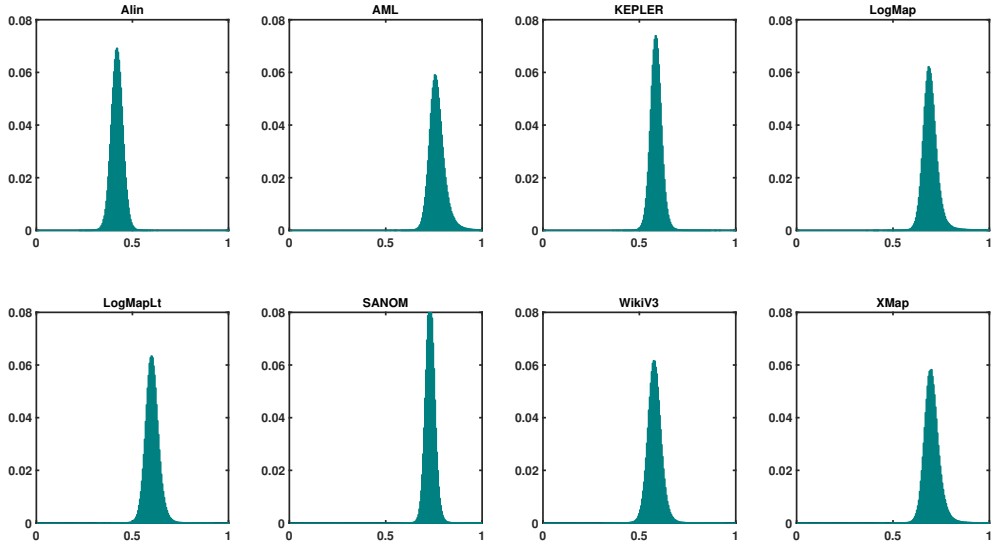


Figure 5.10: The estimation of the F-measure distribution $1 - \tau$ of systems on the OAEI conference track using the proposed Bayesian hierarchical model.

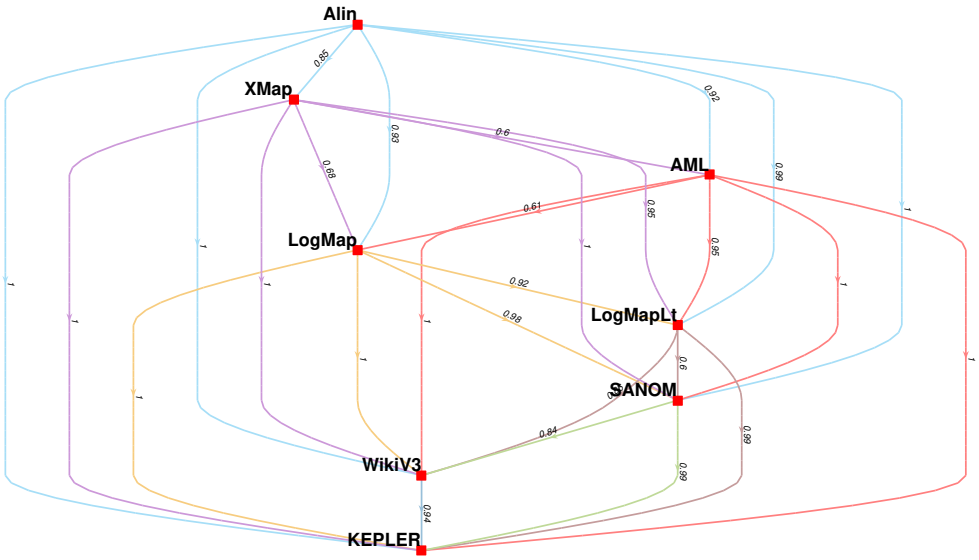


Figure 5.11: Comparison of alignment systems with respect to their precision on the OAEI conference track using the proposed Bayesian test.

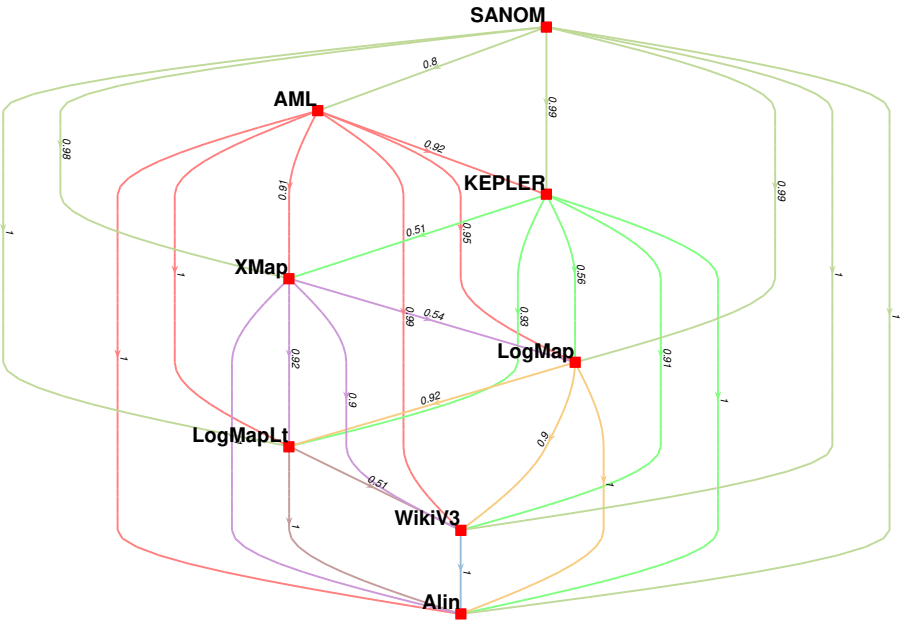


Figure 5.12: Comparison of alignment systems with respect to their recall on the OAEI conference track using the proposed Bayesian test.

different from the anatomy track, since there are multiple benchmarks to conduct the comparison. This would help show the performance of the proposed hierarchical model with respect to the traditional way of the evaluation and comparison.

Table 5.2 displays the evaluation of eight systems on the OAEI conference track. The scores with the subscript MLE are the averages of performance scores over all benchmarks, which is the traditional way of evaluating the alignment systems. We also place the standard deviations (SD) of each score over multiple benchmarks which will yield benefits for the interpretation of the estimated distributions by the proposed model. Besides, the averages of the estimated distributions are also shown for the interest of comparison. The acronyms $\hat{Pr}$, $\hat{F}$, and $\hat{Re}$ represent precision, F-measure, and recall, respectively, and their subscripts indicate if they either the MLE or the Bayesian estimation (BHM).

It is readily seen from Table 5.2 that the means of the estimated distributions are mostly close to the average performance. However, there are some discrepancies as well. For instance, the average F-measure of AML is 0.74, while the mean of its F-measure distribution is around 0.764. We further compute the median, another measure of central tendency, which is known to be more robust in dealing with outliers. Interestingly, the median of F-measures for AML is around 0.762, which is close to what is estimated by the proposed model. The same argument holds for the AML precision estimation, and for other systems with other performance scores, i.e., Alin recall, KEPLER precision, SANOM recall. This experiment supports the validity of the proposed Bayesian model, since the mean of the estimated distributions is at the proximity of the averages or medians of the

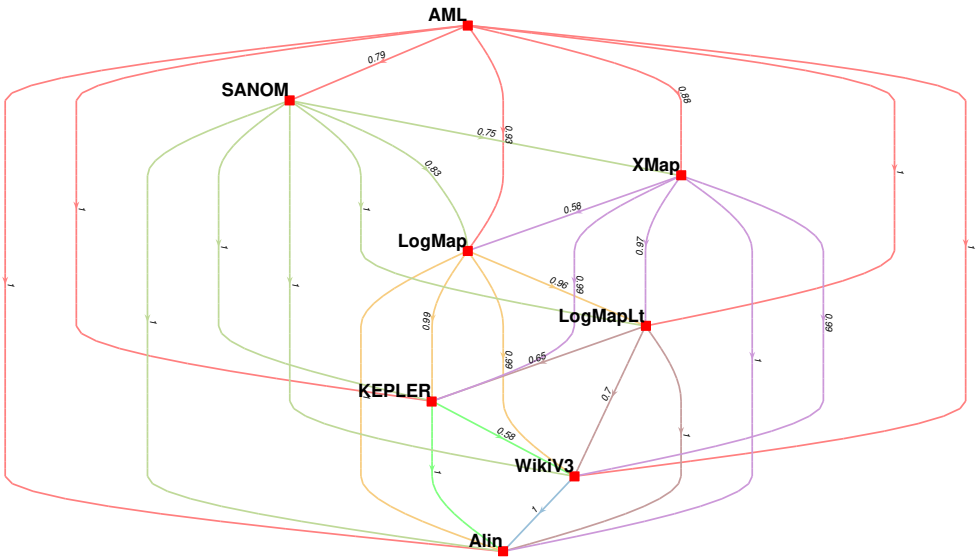


Figure 5.13: Comparison of alignment systems with respect to their F-measure on the OAEI Conference track using the proposed Bayesian test.

associated scores. The experiment also confirms the sensitivity of averaging to outliers, which is one of the most important drawbacks, and corroborates the robustness of the proposed Bayesian model against outliers.

We further plot the estimated distributions by the proposed Bayesian model on the OAEI conference track. Figures 5.8-5.10 display the estimated performance distributions of precision, recall, and F-measure, respectively. Table 5.2 confirms that the central tendencies of distributions are in the proximity of the mean or median of the performance scores. The standard deviations of these distributions are proportionate to the standard deviations of the scores, and to the number of false positives and false negatives. As an instance, the standard deviation of AML precision is less than that of Alin, thanks to Table 5.2. Similarly, the standard deviation of the AML precision distribution is evidently less than that of Alin, according to Figure 5.8. As a result, if the performance scores had little variation over various benchmarks, then the resulting estimated distribution would have a lower standard deviation.

As another example, consider the precision performance distribution of Alin and WikiV3 whose scores' standard deviations are approximately identical (see Table 5.2). However, the performance distribution of WikiV3 is more focused than that of Alin. The reason is that WikiV3 has discovered 222 correspondences overall, of which 149 are correct, while Alin has identified 93 correspondences over all tasks, 83 of which are correct. It is thus expected that the performance distribution of WikiV3 precision is more concentrated than that of Alin. Similar arguments hold for those of other performance scores and other systems.

Having conducted the evaluation of systems, we can now compare them with respect to various performance metrics using the proposed Bayesian test. Figures 5.11-

5.13 show the graphs summarizing the comparison of various systems on the OAEI conference track regarding precision, recall, and F-measure, respectively.

Figure 5.11 indicates that Alin is the best performing system in terms of precision while KEPLER and WikiV3 are those with poor precision. Figure 5.12 supports that SANOM is the top system concerning recall, followed by AML and KEPLER, while Alin and WikiV3 are at the other extreme. The comparison concerning F-measure is summarized in Figure 5.13. According to this figure, AML and SANOM are the top systems, while Alin and WikiV3 are at the other end of the graph.

5.6. CONCLUSION AND FUTURE WORKS

This chapter presented a new way for both evaluation and comparison of alignment systems. The traditional way of the evaluation was to summarize the system performance in a figure, which was a score or its average over multiple benchmark, and the comparison was made based on the juxtaposition of these two figures or using frequentist statistics studied in the previous chapter. This chapter introduced the notion of risk and showed that the MLE of risk with respect to a performance score is exactly the same as the complement of the same score. Instead, we presented a new Bayesian model to estimate a distribution for each of performance metrics. Such a model would give more information about the alignment system performance and would help compare the alignment systems more meaningfully. In fact, the evaluation and comparison of alignment systems were performed by considering the correspondences of alignments generated by systems, and not their performance scores over different benchmarks. This way, the estimation of the difference between alignment systems is more precise compared to the estimation based on the scores computed for each benchmark. We applied the proposed model to the OAEI anatomy and conference tracks and contrasted the results with those of the traditional way. We further compared the systems in those tracks and summarized the overall outcome using a weighted directed graph.

One of the drawbacks of the proposed methodology is that it does not consider the uncertainty regarding each correspondence. Right now, there is an uncertain version for the conference track to which the proposed model cannot be applied, since the correspondences are considered to be only true or false, e.g., the confidence value is one for each discovered correspondence. It is an interesting avenue for improving the proposed model to enable it to estimate the risk of alignment systems in the presence of uncertain correspondences.

Another important problem is that the proposed Bayesian test based on the alignment risk does not consider that the alignment systems are applied over the same benchmark. Thus, the Bayesian test proposed in this chapter overestimates the difference between alignment systems. One way to address this issue is to work with the contingency table discussed in Chapter 4 and develop a new Bayesian model for comparing over single or multiple benchmarks.

In addition, the proposed Bayesian model as well as the frequentist tests in the previous chapter can only consider one performance metric for comparison. For evaluation and comparison, however, it is necessary that multiple performance metrics are taken into account. In this regard, we study the use of different multi-criteria decision-making methods in the next chapter, where alignment systems are compared and ranked based

on multiple performance metrics.

REFERENCES

[1] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon, *Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis,* The Journal of Machine Learning Research **18**, 2653 (2017).

[2] E.-J. Wagenmakers, *A practical solution to the pervasive problems of p-values,* Psychonomic bulletin & review **14**, 779 (2007).

[3] G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon, *Statistical comparison of classifiers through bayesian hierarchical modelling,* Machine Learning **106**, 1817 (2017).

[4] A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, and F. Ruggeri, *A bayesian wilcoxon signed-rank test based on the dirichlet process.* in *ICML* (2014) pp. 1026–1034.

[5] A. Benavoli, G. Corani, F. Mangili, and M. Zaffalon, *A bayesian nonparametric procedure for comparing algorithms,* in *International Conference on Machine Learning* (2015) pp. 1264–1272.

[6] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis* (Chapman and Hall/CRC, 2013).

[7] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice* (CRC press, 1995).

[8] M. Plummer, *Jags: Just another gibbs sampler,* (2004).

[9] J. David, J. Euzenat, F. Scharffe, and C. Trojahn dos Santos, *The alignment api 4.0,* Semantic web **2**, 3 (2011).

[10] J. da Silva, F. A. Baiao, and K. Revoredo, *Alin results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 114.

[11] D. Faria, B. S. Balasubramani, V. R. Shivaprabhu, I. Mott, C. Pesquita, F. M. Couto, and I. F. Cruz, *Results of aml in oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 122.

[12] M. KACHROUDI, G. DIALLO, and S. B. YAHIA, *Oaei 2017 results of kepler,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 138.

[13] E. Jiménez-Ruiz and B. C. Grau, *Logmap: Logic-based and scalable ontology matching,* in *International Semantic Web Conference* (Springer, 2011) pp. 273–288.

[14] M. Mohammadi, A. Atashin, W. Hofman, and Y.-H. Tan, *Sanom results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 185.

[15] M. Mohammadi, W. Hofman, and Y. Tan, *Simulated annealing-based ontology matching,* (2018).

[16] S. Hertling, *Wikiv3 results for oaei 2017,* in *OM-2017: Proceedings of the Twelfth International Workshop on Ontology Matching* (2017) p. 190.

[17] M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, *et al.*, *Results of the ontology alignment evaluation initiative 2016,* in *OM: Ontology Matching* (No commercial editor., 2016) pp. 73–129.

6

ONTOLOGY ALIGNMENT RANKING WITH RESPECT TO MULTIPLE METRICS

It's not about making the right choice, it's about making a choice and making it right

J.R. Rim

In the previous chapters, we discussed the comparison of alignment systems with respect to one performance score only. However, the evaluation and comparison based solely on one performance score would not reflect the overall performance of an ontology alignment system. Another important point in evaluation is to accommodate the experts' preferences and provide the rankings of alignment systems for a particular task. The latter is particularly essential for the OAEI, where different systems compete on several standard benchmarks. In this chapter, we consider the use of multi-criteria decision-making (MCDM) methods for considering the preferences of experts for ontology alignment evaluation, as well as ranking the systems with respect to multiple performance metrics. More in detail, we study different OAEI tracks, for each of which a set of performance metrics is selected based on the literature and experts' opinions. In addition, the best-worst method (BWM) is further extended for group decision-making to calibrate the preferences of multiple experts over the performance metrics for various OAEI tracks. Based on their aggregated preferences and the performance of systems at the OAEI, the alignment systems are ranked by applying different MCDM outranking methods. There are different outranking methods that rank the alignment systems in different and potentially conflicting ways. To resolve this conflict, a compromising ensemble method is also developed to compute an aggregated final ranking for each OAEI track. We apply the overall methodology to five OAEI tracks and report the associated results. We also describe the importance of different performance metrics for each OAEI track based on the preferences of multiple experts.

6.1. INTRODUCTION

In previous chapters, ontology alignment systems were evaluated and compared using different performance metrics such as precision, recall, F-measure, by using two primary statistical schools of thought, frequentist and Bayesian. However, the main drawback of the statistical techniques is that they only consider one performance score for comparing alignment systems, making them unable to take into account different facets of an alignment system measured by several metrics. For instance, an important metric for alignment is execution time, which has to be included in an evaluation and comparison. One way to consider two performance metrics together is to use graphical models such as precision-recall and receiver operating characteristic (ROC) curves [1]. However, these models can only take two performance metrics into account. On top of that, although they provide a broad picture of the performance of systems, they do not rank the systems or compare them systematically.

The use of multi-criteria decision-making (MCDM) methods for ranking and comparing different alignment systems is particularly studied in this chapter. To that end, the comparison of alignment systems is modeled as an MCDM problem, in which different performance metrics and different alignment systems are served as criteria alternatives, respectively, and the ontology alignment experts are the decision-makers (DMs). To use the MCDM methods for comparing alignment systems, we first elicit the preferences of experts over multiple performance metrics for different OAEI tracks. To this end, a survey is designed to extract the preferences of ontology alignment experts according to the best-worst method (BWM) [2, 3]. A major drawback of the BWM (and also many other MCDM methods) is that it can only consider one expert (or DM) at once. To resolve this problem, we extend the model for group decision-making by using Bayesian statistics and proposing Bayesian BWM. Based on the Bayesian BWM, we can calibrate the importance of different performance metrics for each of the OAEI tracks and compute the extent to which one metric is more important than another based on the preferences of multiple experts.

For ranking the alignment systems, We consider another class of MCDM techniques, outranking methods, that rank the alignment systems for each OAEI track with respect to various performance metrics and their computed importance. In addition, the outranking methods encompass the weights that are obtained based on experts' preferences by using the Bayesian BWM. For outranking methods, we review and use three different but appropriate MCDM methods which are able to rank the alignment systems in the presence of multiple criteria/performance metrics. The outranking methods rank alignment systems in different and potentially conflicting ways, thereby making a compromising method for finding an aggregated ranking required. Therefore, a new compromising ensemble method is developed to aggregate the rankings computed by different MCDM methods and compute an overall ranking for all alignment systems in question.

In summary, the contributions of this chapter can be itemized as follows:

- The comparison of ontology alignment systems and incorporating the experts' preferences are formulated as an MCDM problem, where alignment systems and performance metrics are served as the alternatives and criteria, respectively.
- Different performance metrics are considered for multiple OAEI tracks based on

the literature of the OAEI and the opinions of the ontology alignment experts.

- The Bayesian BWM is developed, which is a probabilistic extension of the original BWM for group decision-making problems. The preferences of multiple experts on the performance metrics are first extracted by creating a survey for five OAEI tracks and then aggregated by using the Bayesian BWM. The outcome of Bayesian BWM is the calibration of importance of each performance metric for each OAEI track, as well as the extent to which a group of experts prefer one performance metric over another.
- Three MCDM outranking methods are used to rank alignment systems with respect to multiple performance metrics and their computed importance. Since these methods might produce conflicting rankings, a new compromising ensemble method is developed to find the overall aggregated rankings.

In this chapter, the alignment systems are referred to as A_i, $i = 1,2,...,q$, while the performance metrics are denoted by c_j, $j = 1,2,...,n$. Thus, there are q alignment systems and n performance metrics, that are evaluated by K experts. Furthermore, the matrix containing all performance scores are shown by X, and $X_{i.}$, $X_{.j}$, X_{ij} referring to the i^{th} row, the j^{th} column, and the element at the i^{th} row and the j^{th} column, respectively. Also, we show the Euclidean norm with $\|e\| = \sqrt{\sum_i e_i^2}$. Rankings of the alignment systems, computed by the m^{th} MCDM method, is shown by R_m, $m = 1,...,M$.

This chapter is organized as follows. Section 6.2 describes the research methodology and the steps taken in this study to rank the alignment systems by using MCDM methods. Section 6.3 is dedicated to the BWM and its extension, Bayesian BWM, and Section 6.4 contains the outranking methods used to rank alignment systems. The ensemble method is explained in Section 6.5, and experiments are presented in Section 6.6. Finally, this chapter is concluded in Section 6.7.

6.2. MCDM-BASED COMPARISON AND EVALUATION: METHODOLOGY

In this section, we discuss the steps required to apply the MCDM methods to ontology alignment ranking. Figure 6.1 displays the workflow of using MCDM methods for ranking ontology alignment systems. Although we specifically study the OAEI tracks, the proposed workflow can be used for other ontology alignment benchmarks and/or applications as well. In the following, these steps are explained in details.

STEP 1: SELECTING PERFORMANCE METRICS

First of all, we need to specify the appropriate performance metrics for each OAEI track. To that end, we inspected the OAEI website to accumulate the related performance metrics. Based on this inspection, we created a list of performance metrics for each track and ask the ontology alignment experts, who were mainly the OAEI organizers, for their suitability. After the solicitation, the list of performance metrics is completed, which is tabulated in Table 6.1 for five OAEI tracks. The explanations regarding these performance metrics were presented in Section 2.2.6 on page 25.

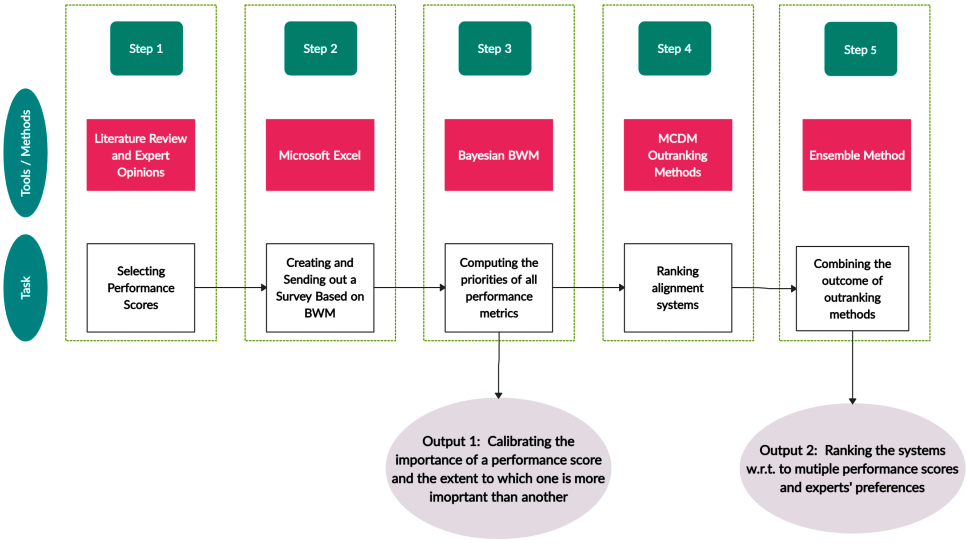


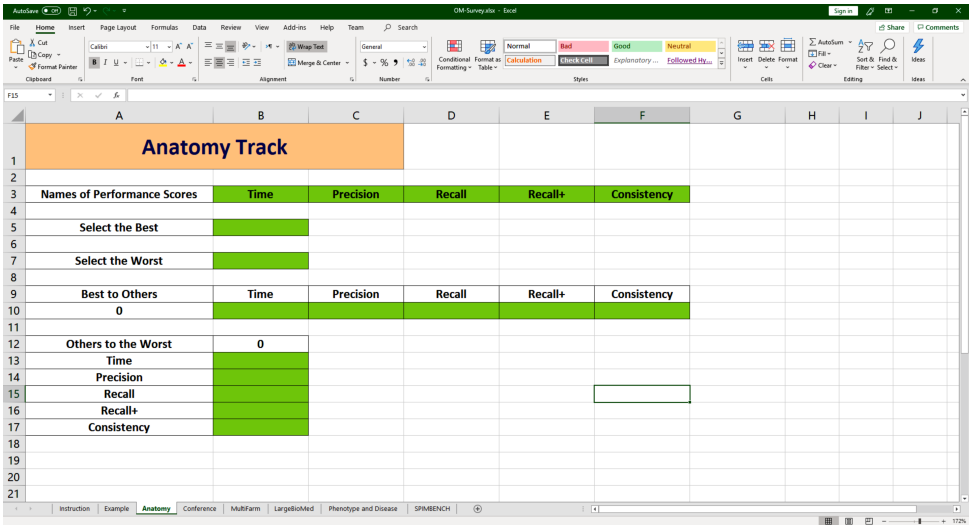
Figure 6.1: The workflow of applying the MCDM methods for comparing ontology alignment systems. The outputs of such evaluation are twofold; 1) The importance of different performance metrics are calibrated based on the experts' preferences, as well as the extent to which one performance metric is more important than one another; 2) The alignment systems are ranked based on experts' opinions and multiple performance scores.

Table 6.1: The selected performance metrics of five tracks of the OAEI.

OAEI track	Performance measures/indicators
Anatomy	time, precision, recall, recall+, consistency
Conference	precision, recall, conservativity, consistency
LargeBioMed	time, precision, recall, consistency
Disease and Phenotype	time, precision, recall
SPIMBENCH	time, precision, recall

STEP 2: CREATING AND SENDING OUT A SURVEY

After determining the performance metrics, we need to elicit the preferences of different experts in the domain in order to specify the importance of these metrics with respect to each other. In this regard, a survey was designed in Microsoft Excel based on best-worst method (BWM) so that experts can specify their preferences for different OAEI tracks. The survey contained an instruction and an example describing the way the experts can correctly evaluate different performance scores. Figure 6.2 plots the survey for the anatomy track. The experts were asked to fill out only the survey of the tracks that they are familiar with. Overall, 12 experts participated in this study, each expressed their preferences for at least one of the tracks.



	A	B	C	D	E	F
1	Anatomy Track					
2						
3	Names of Performance Scores	Time	Precision	Recall	Recall+	Consistency
4						
5	Select the Best					
6						
7	Select the Worst					
8						
9	Best to Others	Time	Precision	Recall	Recall+	Consistency
10	0					
11						
12	Others to the Worst	0				
13	Time					
14	Precision					
15	Recall					
16	Recall+					
17	Consistency					
18						
19						
20						
21						

Figure 6.2: An image of the survey that is used to elicit the preferences of the OAEI experts based on the BWM. The survey designed in Microsoft Excel that contained different sheets for different OAEI tracks. The first two sheets were dedicated to instructions and an example of the BWM.

STEP 3: COMPUTING THE PRIORITIES OF CRITERIA

Since the survey was created based on the BWM, we use it to compute the priorities of different performance metrics for each expert (or decision-maker). The BWM can solely consider one expert at a time. To extend BWM to be applicable for multiple experts, we propose Bayesian BWM, which is able to take the preferences of multiple experts into account and provide a final aggregated priorities reflecting the group opinions. Besides, we can calibrate the extent to which a group of experts prefers one performance metric or criterion to another. As a result, the first outcome of this chapter is the importance of different performance metrics for five OAEI tracks based on experts' preferences. The Bayesian BWM is discussed in detail in Section 6.3.

STEP 4: RANKING ALIGNMENT SYSTEMS

After having the priorities of all performance metrics based on the preferences of all experts, we use another class of MCDM methods, outranking methods, to rank the alignment systems with respect to multiple performance scores and their computed importance. There are a handful number of outranking methods that can be used for ranking the alignment systems. We use three well-known outranking methods, TOPSIS [4], VIKOR [5], and PROMETHEE [5], to rank the alignment systems. These methods can provide rankings of alignment systems with respect to multiple performance scores and their importance. MCDM outranking methods are discussed in Section 6.4.

STEP 5: COMBINING THE OUTCOME OF OUTRANKING METHODS

Different outranking methods may generate different and sometimes conflicting ranking for alignment systems. Therefore, we need a method for aggregating their results.

In this regard, we develop an ensemble method based on the half-quadratic (HQ) theory [6] and find an aggregated final ranking. The ranking of alignment systems is the second outcome of applying MCDM to ontology alignment evaluation, which is of the utmost importance to the OAEI competition. The ensemble method for the aggregation is discussed in Section 6.5.

6.3. BAYESIAN BEST-WORST METHOD

In this section, we propose a novel method for group MCDM that is particularly presented for the BWM. To this end, the input vectors associated with each expert in the BWM is modeled using the multinomial distribution, while we show that the underlying idea of the original BWM is persevered. The proposed method is called Bayesian BWM which can solve the group MCDM problem. The inputs to the Bayesian BWM are identical to those of the original BWM, which are the pairwise comparisons. The output is, on the other hand, the optimal aggregated final weights reflecting the total preferences of all the experts along with the confidence level for ranking the criteria.

In the remainder of this section, we first review the original best-worst method and the corresponding optimization problem to obtain the optimal weights of the performance metrics (or, criteria in general) for one expert or DM only. Further, we provide the probabilistic interpretation of the BWM inputs and outputs and justify that such an interpretation preserves the underlying ideas in the original BWM. Then, a Bayesian hierarchical model is presented that can find the aggregated weights of a group of experts. In addition, we introduce the credal ranking that can calibrate the extent to which one criterion or performance metric is more important than the other. This is particularly useful because it helps describe the importance of various performance metrics for each OAEI track.

6.3.1. BEST-WORST METHOD

The BWM is one of the latest MCDM methods [2, 3] that is based on the pairwise comparison of criteria. The steps required for the original BWM are as follows [2]:

Step 1: The expert needs to provide a set of decision criteria $C = \{c_1, c_2, ..., c_n\}$. These criteria are the performance metrics for ontology alignment evaluation, which are determined by inspecting the literature and experts' opinions.

Step 2: The expert selects the best (c_B) and the worst (c_W) performance metrics from C.

In this step, the expert only selects the best and the worst from set C identified in the first step. The expert does not conduct any pairwise comparison in this stage. Based on the expert's preference, the best performance metric is the most important or the most desirable, while the worst performance metric is the least important or the least desirable criterion among others.

Step 3: The expert conducts the pairwise comparison between the best (c_B) and the other metrics from C.

In this step, the expert calibrates his/her preferences of the best performance metric to the other metrics by a number between one and nine, where *one* means

equally important and *nine* means extremely more important. The pairwise comparison leads to the "Best-to-Others" vector A_B as

$$A_B = (a_{B1}, a_{B2}, ..., a_{Bn}), \tag{6.1}$$

where a_{Bj} represents the preference of the best (c_B) to $c_j \in C$.

Step 4: The expert conducts the pairwise comparison between the worst (c_W) and the other metrics from C.

Similar to Step 3, the expert needs to calibrate his/her preferences of the other performance metrics over the worst by a number between one and nine. The result of this step is the "Others-to-Worst" vector A_W as

$$A_W = (a_{1W}, a_{2W}, ..., a_{nW})^T, \tag{6.2}$$

where a_{jW} represents the preference of $c_j \in C$ over the worst (c_W).

Step 5: Obtaining the optimal weights $w = (w_1, w_2, ..., w_n)$.

Given A_B and A_W, a weight vector w must be computed. The weight vector must be in the neighborhood of equations $w_B/w_j = a_{Bj}$ and $w_j/w_W = a_{jW}$ for $j = 1,2,..,n$. Thus, one can minimize the maximum absolute differences $|\frac{w_B}{w_j} - a_{Bj}|$ and $|\frac{w_j}{w_W} - a_{jW}|$ for all $j = 1,2,...,n$. Besides, the non-negativity and unit-sum property of the weight vector must be fulfilled. As a result, the following optimization problem can find the optimal weight vector w [2]:

$$\begin{aligned} &\min_w \max_j \left\{ \left| \frac{w_B}{w_j} - a_{Bj} \right|, \left| \frac{w_j}{w_W} - a_{jW} \right| \right\} \\ &s.t. \quad \sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0 \quad \forall j = 1,2,...,n. \end{aligned} \tag{6.3}$$

Similarly, the weight vector can also be calculated by the following problem [3]:

$$\begin{aligned} \min_{\xi, w} \quad & \xi \\ s.t. \quad & \left| \frac{w_B}{w_j} - a_{Bj} \right| \leq \xi \quad \forall j = 1,2,...,n \\ & \left| \frac{w_j}{w_W} - a_{jW} \right| \leq \xi \quad \forall j = 1,2,...,n \\ & \sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0 \qquad \forall j = 1,2,...,n. \end{aligned} \tag{6.4}$$

6.3.2. PROBABILISTIC INTERPRETATION OF BWM

We now provide a probabilistic interpretation of the BWM inputs and outputs, and then review two schools of thoughts in the probability estimation, frequentist and Bayesian, in the context of the BWM.

MODELING INPUTS AND OUTPUTS: MULTINOMIAL AND DIRICHLET DISTRIBUTIONS

The typical outcome of MCDM methods is a weight vector $w = [w_1, ..., w_n]$ such that $w_j \geq 0, \sum_{j=1}^{n} w_j = 1$. The magnitude of each w_j indicates the importance of the corresponding performance metric c_j.

From a probabilistic perspective, the performance metrics are seen as the random events, and their weights are thus their occurrence likelihoods. Mathematically speaking, such an interpretation is in line with the MCDM, since $w_j \geq 0$ and $\sum_{j=1}^{n} w_j = 1$ according to the probability theory as well. It is further of the essence to illustrate that probabilistic modeling makes sense from a decision-making point of view.

For the probabilistic reasoning, one needs to model all the inputs and the outputs as the probability distributions. First, consider A_B and A_W which are the inputs to the BWM. From a mathematical point of view, the multinomial distribution can model the vectors, since all of their elements are integers. The probability mass function (PMF) of the multinomial distribution for a given A_w is [7]

$$P(A_W|w) = \frac{(\sum_{j=1}^{n} a_{jW})!}{\prod_{j=1}^{n} a_{jW}!} \prod_{j=1}^{n} w_j^{a_{jW}} \tag{6.5}$$

where w is the discrete probability distribution.

In the multinomial distribution, the weight vector is the discrete probability distribution and A_W contains the number of occurrence of each event. Apparently, it is completely different from what is expected for the BWM represented in Section 6.3.1. We show that modeling with multinomial would fulfill the underlying idea of the BWM.

Based on the multinomial distribution, the probability of event j is proportionate to the number of occurrence of the event to the total number of trials, i.e.,

$$w_j \propto \frac{a_{jW}}{\sum_{i=1}^{n} a_{iW}} \quad \forall j = 1, ..., n. \tag{6.6}$$

Similarly, one can write the same equation for the worst performance metric as

$$w_W \propto \frac{a_{WW}}{\sum_{i=1}^{n} a_{iW}} = \frac{1}{\sum_{i=1}^{n} a_{iW}} \tag{6.7}$$

Using equations (6.6) and (6.7), one obtains

$$\frac{w_j}{w_W} \propto a_{jW}, \quad \forall j = 1, ..., n, \tag{6.8}$$

which is precisely the relation we seek in the original BWM presented in Step 5 of Section 6.3.1.

Similarly, A_B can be modeled using the multinomial distribution. However, A_B is different from A_W: The former represents the preferences of the best over the other performance metrics, while the latter denotes the preferences of the others over the worst. Thus, A_B yields the inverse of the weight, i.e.,

$$A_B \sim multinomial(1/w), \tag{6.9}$$

where w is the probability distribution, and / represents the element-wise division operator. Identical to the worst performacne metric, one can write

$$\begin{aligned} \frac{1}{w_j} \propto \frac{a_{Bj}}{\sum_i a_{Bi}}, \quad \frac{1}{w_B} \propto \frac{a_{BB}}{\sum_{i=1}^{n} a_{Bi}} = \frac{1}{\sum_j a_{Bi}} \\ \Rightarrow \frac{w_B}{w_j} \propto a_{Bj}, \quad \forall j = 1, ..., n, \end{aligned} \tag{6.10}$$

which is again the exact relation we seek in the BWM.

So far, we showed that the multinomial distribution could meaningfully model the inputs to the BWM. The problem of finding the weights in the MCDM problem is thus transferred to the estimation of a probability distribution. Therefore, one can use the statistical inference techniques to find w in the multinomial distribution.

A weight vector for the MCDM must satisfy the non-negativity and unit-sum properties. Therefore, an appropriate distribution to model the weights is the Dirichlet distribution, which is also the Bayesian prior to the multinomial distribution. Given a parameter $\alpha \in R^n$, the Dirichlet distribution of the weights w is defined as [7]

$$Dir(w|\alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^{n} w_j^{\alpha_j - 1}. \tag{6.11}$$

The distribution has only a vector parameter α, and w meets the constraints of an optimal weight vector of MCDM, since it is a discrete probability distribution.

ESTIMATION OF THE WEIGHT VECTOR: STATISTICAL INFERENCE

For a moment, assume that there is only A_W in the BWM, then we consider two widely-accepted inference techniques: frequentist and Bayesian. The underlying idea of the frequentist approach is that there is a precise yet unknown optimal point, and the effort is to estimate it based on the observations. As a result, the outcome of the frequentist inference is a precise weight vector for a set of criteria. The maximum likelihood estimation (MLE) is arguably the most popular inference technique which finds the optimal weight vector using the following optimization

$$w = arg \max_{w, \sum_{j=1}^{n} w_j = 1} P(A_W|w). \tag{6.12}$$

The optimum of (6.12) yields at

$$w_j^* = \frac{a_{jW}}{\sum_{i=1}^{n} a_{iW}}, \quad \forall j = 1, ..., n, \tag{6.13}$$

which is indeed the normalized A_W. The same solution will be obtained by the BWM if the preferences of the DM are fully consistent. Thus, equation (6.13) shows that the MLE bears the same result as the BWM under specific circumstances.

The second approach is the Bayesian estimation, in which the parameters are approximated by using a distribution rather than a precise point as is in the MLE. Thus, we first need to specify a prior distribution for the weight vector. In the Bayesian inference, the Dirichlet distribution is used as the prior to the multinomial. The Dirichlet distribution can represent the weight vector, since it satisfies both its non-negativity and unit-sum properties. Using Dirichlet as the prior and multinomial as the likelihood, the posterior distribution would also be Dirichlet with the posterior parameter $\alpha_{post} = \alpha + A_W$. Since the prior should be uninformative to have a minimal impact on the posterior, we set the prior parameter $\alpha = 1$.

As a result of the Bayesian estimation, the values of w is shown by a Dirichlet distribution. The mode of the posterior distribution $\mu \in R^n$ with the parameter α_{post} is:

$$\begin{aligned}\mu_j &= \frac{\alpha_{post_j} - 1}{\sum_{i=1}^{n} \alpha_{post_i} - n} \\ &= \frac{1 + a_{jW} - 1}{\sum_{i=1}^{n} (a_{iW} + 1) - n} \\ &= \frac{a_{jW}}{\sum_{i=1}^{n} a_{iW}}, \quad \forall j = 1, ..., n. \end{aligned} \tag{6.14}$$

Thus, the mode of the posterior distribution would provide the exact MLE. As a result, the Bayesian paradigm would yield more information regarding the events in question, since its outcome is a distribution, not a point. The standard deviation of such a distribution, for instance, is an indicator of uncertainty regarding the inference problem, which can have distinct interpretations with respect to the problem under study.

So far, we merely considered A_W for estimating the weights; however, it is critical to use both A_B and A_W according to the BWM. The MLE inference containing both A_B and A_W does not bear an analytical solution due to the complexity of the corresponding optimization problem. Further, the simple Dirichlet-multinomial conjugate cannot encompass the A_B and A_W together. The problem compounds when it comes to having the preferences of multiple experts. Considering these issues, a Bayesian hierarchical model is presented in the next section to estimate the optimal weight of the criteria considering both A_B and A_W of multiple experts.

6.3.3. BAYESIAN BEST-WORST METHOD

This section presents a Bayesian hierarchical model to find the optimal weights of a set of performance metrics based on the preferences of multiple experts using the best-worst framework.

GROUP DECISION-MAKING: A JOINT PROBABILITY DISTRIBUTION

Assume that the k^{th} expert, $k = 1, ..., K$, evaluates the criteria $c_1, ..., c_n$ by providing the vectors A_B^k and A_W^k. We show the set of all vectors of K experts by $A_B^{1:K}$ and $A_W^{1:K}$. From now on, the superscript $^{1:K}$ would indicate the total of all vectors in the base. We also represent the overall optimal weight by w^*.

The estimation of w^* entails using several auxiliary variables. In particular, w^* is computed based on the optimal weights of K experts shown by w^k, $k = 1, ..., K$. Thus,

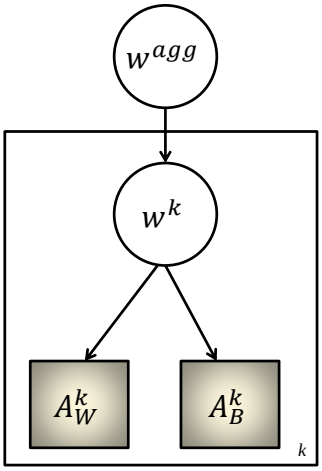


Figure 6.3: The probabilistic graphical model of the Bayesian BWM.

the proposed Bayesian model would simultaneously compute w^* and $w^{1:K}$. Prior to conducting any statistical inference, it is required to write the joint probability distribution of all random variables given the available data. In the group decision-making within the BWM, $A_B^{1:K}$ and $A_W^{1:K}$ are given, and $w^{1:K}$ and w^* must be estimated accordingly. Thus, the following joint probability distribution is sought

$$P\left(w^*, w^{1:K} \,\middle|\, A_B^{1:K}, A_W^{1:K}\right). \tag{6.15}$$

If the probability in (6.15) is computed, then the probability of each individual variable can be computed using the following probability rule,

$$P(x) = \sum_y P(x, y), \tag{6.16}$$

where x and y are two arbitrary random variables.

BAYESIAN HIERARCHICAL MODEL

To develop a Bayesian model, we first need to identify the independence and conditional independence of variables. Figure 6.3 plots the graphical model corresponding to the proposed method. The nodes in the graph are the variables. As a convention, the rectangles are the observed variables, which are the inputs to the original BWM, and the circular nodes are the variables that must be estimated. Also, arrows denote that the node in origin is dependent on the node at the other end. That is to say, the value of w_k is dependent on A_W^k and A_B^k, and the value of w^* is also dependent on w^k. The plate, which covers a set of variables, means that the corresponding variables are iterated for each expert, and w^* is not in the plate, since there is only one w^* for all experts.

The conditional independence between various variables is clear based on Figure 6.3. For instance, A_W^k is independent of w^* given w^k, i.e.,

$$P\left(A_W^k \mid w^*, w^k\right) = P\left(A_W^k \mid w^k\right) \tag{6.17}$$

Considering all independence among different variables, applying the Bayes rule to the joint probability (6.15) follows:

$$\begin{aligned} P\left(w^*, w^{1:K} \mid A_B^{1:K}, A_W^{1:K}\right) &\propto P\left(A_B^{1:K}, A_W^{1:K} \mid w^*, w^{1:K}\right) P\left(w^*, w^{1:K}\right) \\ &= P(w^*) \prod_{k=1}^{K} P\left(A_W^k \mid w^k\right) P\left(A_B^k \mid w^k\right) P\left(w^k \mid w^*\right), \end{aligned} \tag{6.18}$$

where the last equality is obtained using the probability chain rule and the conditional independence of different variables, and the fact that each expert provides his/her preferences independently. Since the estimation of the parameters in equation (6.18) is reliant on the estimation of other variables, there is a chain between different parameters. The existence of the chain is the reason that the model is called *hierarchical*.

We now need to specify the distributions of each element in equation (6.18). We have already shown that A_B and A_W can be modeled using the multinomial distribution in the sense that it preserves the underlying idea of the BWM. There is only one difference between A_B and A_W, since the former shows the preference of all the performance metrics over the worst, while the latter contains the preference of the best over all the others. Thus, one can model them as

$$\begin{aligned} A_B^k \mid w^k &\sim multinomial(1/w^k), && \forall k = 1, ..., K, \\ A_W^k \mid w^k &\sim multinomial(w^k), && \forall k = 1, ..., K. \end{aligned} \tag{6.19}$$

Given w^*, one can expect that each and every w^k be in its proximity. To this end, we reparameterize the Dirichlet distribution with respect to its mean and a concentration parameter. The models of w^k given w^* are

$$w^k \mid w^* \sim Dir(\gamma \times w^*), \quad \forall k = 1, ..., K, \tag{6.20}$$

where w^* is the mean of the distribution and γ is the concentration parameter. The equation in (6.20) says that the weight vector w^k associated with each expert must be in the proximity of w^*, since it is the mean of the distribution, and their closeness is governed by the non-negative parameter γ. The concentration parameter also needs to be modeled using a distribution. A reliable option is the gamma distribution which satisfies the non-negativity constraints, i.e.,

$$\gamma \sim gamma(a, b), \tag{6.21}$$

where a and b are the shape parameters of the gamma distribution.

We finally supply the prior distribution over w^* using an uninformative Dirichlet distribution with the parameter $\alpha = 1$ as

$$w^* \sim Dir(\alpha). \tag{6.22}$$

The specified model does not bear a closed-form solution. As a result, Markov-chain Monte Carlo (MCMC) techniques [8] must be used to compute the posterior distribution. For the MCMC sampling, we use the "just another Gibbs sampler" (JAGS) [9], which is one of the best available probabilistic languages to date, to sample and compute the posterior determined in (6.18). The useful outcome of the model is the posterior distribution of weights for every expert and the aggregated w^*.

The proposed Bayesian model will replace Step 5 of the original BWM explained in Section 6.3.1. In fact, the optimization problem is substituted with a probabilistic model, while the inputs to both methods are identical. However, the proposed model, not only encompasses the preferences of multiple experts, but it also provides more information regarding the confidence of the relation between each pair of the criteria. The extra information is obtained by devising a new Bayesian test based on the approximated distribution from the model, which is explained in the next section.

6.3.4. CREDAL RANKING

The modus operandi in the MCDM is to say one criterion is more important than one another merely if its weight, or the weight average for the group decision-making, is higher than one another. The notion of credal ranking is now introduced, which can calibrate the degree to which one criterion or performance metric is superior to the other one. This is done by using the samples obtained from the MCMC sampling from the proposed Bayesian model. We first define the credal ordering, which is the building-block of credal ranking.

Definition 9 (Credal Ordering) *For a pair of criteria or performance metrics c_i and c_j, the credal ordering O is defined as*

$$O = (c_i, c_j, R, d) \tag{6.23}$$

where

- *R is the relation between the metrics c_i and c_j, i.e., $<$, $>$, or $=$;*
- *$d \in [0,1]$ represents the confidences of the relation.*

Definition 10 (Credal Ranking) *For a set of criteria or performance metrics $C = (c_1, c_2, ..., c_n)$, the credal ranking is a set of credal orderings which includes all pairs (c_i, c_j), for all $c_i, c_j \in C$.*

The confidence in the credal ordering can provide the experts with more information regarding the extent to which one performance metric is preferred over another. We now devise a new Bayesian test based on which we can find the confidence of each credal ordering. The test is predicated on the posterior distribution of w^*. The confidence that c_i being superior to c_j is computed as

$$P(c_i > c_j) = \int I_{(w_i^* > w_j^*)} P\left(w^*\right), \tag{6.24}$$

where $P(w^*)$ is the posterior distribution of w^* and I equals to one if the condition in the subscript holds, and zero otherwise. This integration can be approximated by the

samples obtained via the MCMC. Having S samples from the posterior distribution, the confidence can be computed as

$$P(c_i > c_j) = \frac{1}{S}\sum_{s=1}^{S} I(w_i^{*s} > w_j^{*s}),$$
$$P(c_j > c_i) = \frac{1}{S}\sum_{s=1}^{S} I(w_j^{*s} > w_i^{*s}), \tag{6.25}$$

where w^{*s} is the s^{th} sample of w^* from the MCMC samples. Thus, for each pair of criteria, one can compute the confidence that one is superior to another. The credal ranking can be easily changed into the traditional ranking. In this regard, it is evident that $P(c_i > c_j) + P(c_j > c_i) = 1$. Therefore, c_i is more important than c_j if and only if $P(c_i > c_j) > 0.5$. As a result, the traditional ranking of criteria is obtainable by applying a threshold of 0.5 to the credal ranking.

6.4. MCDM OUTRANKING METHODS

In this study, we use three different MCDM methods, TOPSIS, VIKOR, and PROMETHEE. These methods are used to rank the alignment systems with respect to several performance metrics and experts' preferences. We use the mean of w^*, shown as $\bar{w}^*$, as the weight of performance metrics used in the outranking methods.

6.4.1. TECHNIQUE FOR ORDER PREFERENCE BY SIMILARITY TO IDEAL SOLUTION (TOPSIS)

TOPSIS is one of the popular MCDM methods for ranking alternatives (alignment systems) with respect to a set of criteria (performance metrics) [4]. It first identifies the positive-ideal and negative-ideal solutions and then ranks the alignment systems based on their distances to the two solutions. The alignment systems are ranked based on their closeness to the positive-ideal and their distance from the negative-ideal solution.

TOPSIS has many variations and extensions [10–12], in this study, we adopt the original version proposed in [13]. The ranking process in TOPSIS includes the following steps:

Step 1: First, the performance matrix should be normalized. The element of the normalized matrix $\hat{X}$ is calculated as,

$$\hat{X}_{ij} = \frac{X_{ij}}{\|X_{.j}\|}, \quad i = 1,2,...,q, \quad j = 1,2,...,n. \tag{6.26}$$

Step 2: Find the positive-ideal solution $S^+ = (S_1^+, S_2^+, ..., S_n^+)$, where $S_j^+ = \max_i \hat{X}_{ij}\bar{w}_j^*$ for benefit criteria, e.g., profit, and $S_j^+ = \min_i \hat{X}_{ij}\bar{w}_j^*$ for cost criteria, e.g., time.

Step 3: Find the negative-ideal solution $S^- = (S_1^-, S_2^-, ..., S_n^-)$, where $S_j^- = \min_i \hat{X}_{ij}\bar{w}_j^*$ for the benefit criteria, and $S_j^- = \max_i \hat{X}_{ij}\bar{w}_j^*$ for cost criteria.

Step 4: Calculate the Euclidean distance to the positive-ideal and negative-ideal solutions for each alternative. For the k^{th} alternative, the distance to the ideal solution D_i^+

and to the negative-ideal solution D_i^- is computed as

$$D_i^+ = \|\hat{X}_{i.} - S^+\|, \qquad D_i^- = \|\hat{X}_{i.} - S^-\|. \tag{6.27}$$

Step 5: Calculate the ratio L_i for each alternative as

$$L_i = \frac{D_i^-}{D_i^+ + D_i^-}, \quad i = 1, ..., q. \tag{6.28}$$

Step 6: Rank the alternatives according to their ratios L_i.

6.4.2. VLSEKRITERIJUMSKA OPTIMIZACIJA I KOMPROMISNO RESENJE (VIKOR)

VIKOR is another MCDM method that ranks the alternatives based on a set of possibly conflicting criteria. The procedure used in VIKOR can be summarized as follows [5, 14]:

Step 1: Find the best f^+ and the worst f^- values among the alignment systems for all the criteria (performance metrics). For the benefit criteria, we have

$$\begin{aligned} f_j^+ &= \max_i X_{ij}, \quad j = 1, 2, ..., n, \\ f_j^- &= \min_i X_{ij}, \quad j = 1, 2, ..., n, \end{aligned} \tag{6.29}$$

where the minimum and maximum are substituted if it is the cost criteria.

Step 2: For each alignment system, compute S_i and R_i as

$$\begin{aligned} S_i &= \sum_{j=1}^{n} \bar{w}_j^* \frac{f_j^+ - X_{ij}}{f_j^+ - f_j^-}, \\ R_i &= \max_j \left\{ \bar{w}_j^* \frac{f_j^+ - X_{ij}}{f_j^+ - f_j^-} \right\}. \end{aligned} \tag{6.30}$$

Step 3: For each alignment system, calculate Q_j as

$$\begin{aligned} Q_i &= \nu \frac{S_i - S^+}{S^- - S^+} + (1 - \nu) \frac{R_i - R^+}{R^- - R^+}, \\ S^+ &= \min_i S_i, \qquad S^- = \max_i S_i, \\ R^+ &= \min_i R_i, \qquad R^- = \max_j R_i, \end{aligned} \tag{6.31}$$

where $\nu \in [0, 1]$ is a trade-off parameter. It is the common practice to set $\nu = 0.5$.

Step 4: Ranking the alignment systems based on their corresponding Q_j in descending order.

Step 5: For two alignment system A_i and A_j, A_i is given a better rank than A_j if: (a) $Q_i - Q_j > 1/q - 1$; and (b) A_i has a better rank according to S_i and/or R_i.

6.4.3. PREFERENCE RANKING ORGANIZATION METHOD FOR ENRICHMENT OF EVALUATIONS (PROMETHEE)

PROMETHEE uses a pairwise comparison between different alignment systems to establish a ranking. While PROMETHEE I [15] conducts a partial pairwise comparison and computes the ranking accordingly, PROMETHEE II [16] uses a complete pairwise comparison, which makes it more suitable to rank the alignment systems. The ranking procedure used by PROMETHEE II is as follows:

Step 1: For $i, i' = 1, 2, ..., q$, compute the function $\pi_{ii'}$ as the number of criteria in which A_i has better performance than $A_{i'}$, e.g.,

$$\pi_{ii'} = \sum_{j=1}^{n} \bar{w}_j^* I(X_{ij} > X_{i'j}), \qquad i, i' = 1, 2, ..., q, \tag{6.32}$$

where I is the Dirac function which is one when the condition in the parenthesis is satisfied, and 0 when it is not.

Step 2: Calculate the positive ϕ^+ and negative ϕ^- outranking flow for each alternative as,

$$\phi^+(A_i) = \frac{1}{q-1} \sum_{i'=1}^{q} \pi_{ii'}, \qquad \phi^-(A_i) = \frac{1}{q-1} \sum_{i'=1}^{q} \pi_{i'i}. \tag{6.33}$$

Step 3: Rank in decreasing order the alignment systems based on the difference between their positive and negative outranking flow.

6.5. AN ENSEMBLE OF MCDM OUTRANKING METHODS

MCDM outranking methods may provide different rankings for the same problem because they use different mechanisms, making it hard to provide sufficient support for the ranking of one MCDM outranking method compared to the others. As such, in this section, a compromise method is developed to estimate the final rankings of all alignment systems based on the rankings of different MCDM methods. The proposed method utilizes the half-quadratic (HQ) theory which results in estimating a weight for each of the MCDM methods. The weights obtained by the method satisfy the non-negativity and unit-sum properties, which are necessary for the MCDM methods. Another important property of the proposed method is that, in contrast to averaging, it is insensitive to outliers, owing to the use of the robust HQ functions. In aggregating MCDM rankings, outliers are the outranking methods whose rankings deviate from the majority of rankings, which is expected that they contribute less to the final aggregated rankings. In addition to the aggregated rankings, a consensus index and a trust level are calculated for the aggregated rankings. In the following, we first review the fundamentals of the HQ theory.

6.5.1. HALF-QUADRATIC MINIMIZATION

In this section, we review the fundamental theory of the HQ minimization, introduce the appropriate HQ functions and look at the minimization procedure of HQ programming.

Table 6.2: Different M-estimators and their corresponding minimizer function $\delta(.)$ based on the HQ multiplicative form. β is a positive constant and σ is the parameter of the HQ functions.

estimators	l1-l2	fair	log-cosh	Welsch	Huber
HQ function g(t)	$\sqrt{\beta+\frac{t^2}{\sigma^2}}-1$	$\frac{\lvert t\rvert}{\beta}-log(1+\frac{\lvert t\rvert}{\beta})$	$log(cosh(\beta t))$	$1-exp(-\frac{t^2}{\sigma^2})$	$\begin{cases}\frac{t^2}{2} & \lvert t\rvert\le\sigma\\ \sigma\lvert t\rvert-\frac{\sigma^2}{2} & \lvert t\rvert>\sigma\end{cases}$
Minimizer Function $\delta(t)$	$\frac{1}{\sqrt{\beta+t^2}}$	$\frac{1}{\beta(\beta+\lvert t\rvert)}$	$\frac{\beta}{t}tanh(\beta t)$	$exp(-\frac{t^2}{\sigma^2})$	$\begin{cases}1 & \lvert t\rvert\le\sigma\\ \frac{\sigma}{\lvert t\rvert} & \lvert t\rvert>\sigma\end{cases}$

The Euclidean norm is arguably the most popular loss function used in various circumstances such as regression, while least square fitting is the most popular regression technique that utilizes the Euclidean norm as the loss function. Although it is simple and also yields a closed-form solution, it is highly sensitive to outliers and shows diminished performance in noisy environments. A viable way to solve that sensitivity is to use various robust estimators. In robust statistics, M-estimator is a family of the robust estimators, by which the HQ functions are inspired. Although these functions are not convex, their optimum can be obtained using HQ minimization with guaranteed convergence. Table 6.2 tabulates the HQ functions along with their minimizer functions $\delta()$ that are used in the optimization procedure.

Consider the following minimization,

$$\min_{s}\sum_{j}g(s_j) \tag{6.34}$$

where $g(.)$ is one of the HQ functions tabulated in Table 6.2. To solve problem (6.34), there are two forms of the HQ programming (multiplicative [17] and additive [18]) that can efficiently find a local optimal solution. Both forms have been applied to different areas, including robust estimation [19, 20], image processing [21, 22], machine learning [23, 24], and bioinformatics [25, 26]. Here, we use the multiplicative form, since its optimization procedure can be interpreted meaningfully within the MCDM.

Based on the multiplicative form of HQ programming [6, 17], problem (6.34) can be rewritten as

$$\min_{s,w}\sum_{j}\alpha_j s_j^2+\psi(\alpha_j) \tag{6.35}$$

where $\alpha_j>0$ is the HQ auxiliary variable, and $\psi(.)$ is the convex conjugate of g defined as [27],

$$\psi(p)=\max_{e}ep-g(e). \tag{6.36}$$

To solve minimization (6.35), variables α and s must be updated iteratively until convergence is reached. Based on the HQ multiplicative theory [17], the update of variables is as follows:

$$\begin{aligned}&\alpha_j=\delta(s_j),\\ &s=\arg\min_{s}\sum_{j}\alpha_j s_j^2,\end{aligned} \tag{6.37}$$

where $\delta(.)$ is the minimizer function with respect to $g(.)$ (see Table 6.2).

In the next section, a new compromise method is developed based on the multiplicative HQ minimization, and it is shown that the auxiliary variable α plays the role of weights in the MCDM problem. Since the value of α is reliant on the type of HQ function $g(.)$, different HQ functions would result in different weights and different final aggregated rankings. We particularly consider the Welsch M-estimator, for two reasons. First, it has shown a promising performance in a variety of problems and it is known to be the most promising and outlier-robust estimator among the HQ functions [21, 25]. Second, we can calculate a consensus index and a trust level if the Welsch estimator is used.

6.5.2. AN HQ-BASED COMPROMISE METHOD

The proposed ensemble method can be used for any number of MCDM methods. In this regard, assume that there are M MCDM outranking methods which rank q alternatives (alignment systems) on the basis of n criteria (performance metrics).

A simple yet practical solution to estimate the overall rankings R^* is to minimize its Euclidean distance to rankings computed by each MCDM method. The corresponding minimization is,

$$\min_{R^*} \frac{1}{2} \sum_{m=1}^{M} \|R_m - R^*\|_2^2, \tag{6.38}$$

where M is the number of MCDM methods and R_m is the ranking of the m^{th} MCDM method. Minimization (6.38) has the following closed-form solution

$$R^* = \frac{1}{M} \sum_{m=1}^{M} R_m, \tag{6.39}$$

which is indeed the average of the rankings produced by different methods. However, averages are not reliable estimators, since they are sensitive to outliers [28], like other methods using the Euclidean norm as their basic loss function. In ranking aggregation, it means that, if one MCDM method has distinct rankings from the other methods, it can significantly influence the overall ranking. Instead, we utilize the HQ functions, which are potentially insensitive to outliers [29], as well as allowing us to compute a consensus index and trust level for the final aggregated rankings.

The proposed optimization problem to estimate R^* is,

$$\min_{R^*} \frac{1}{2} \sum_{m=1}^{M} g(\|R_m - R^*\|_2), \tag{6.40}$$

where $g(.)$ is an HQ function. Although minimization (6.40) is not convex, it can be solved efficiently using HQ programming [6, 17]. Using the HQ multiplicative form as in (6.35), minimization (6.40) can be restated as,

$$\min_{R^*,\alpha} J(R^*, \alpha) = \sum_{m=1}^{M} \alpha_m \|R_m - R_*\|_2^2 + \psi(\alpha_m), \tag{6.41}$$

where $\alpha \in R^M$ is the half-quadratic auxiliary variable. According to the HQ programming, the following steps must be iterated until convergence for the two variables is reached:

$$\alpha_m = \delta\Big(\|R_m - R^*\|_2\Big), \quad m = 1,...,M,$$
$$R^* = \arg\min_{R^*} \sum_{m=1}^{M} \alpha_m \|R_m - R^*\|_2^2. \tag{6.42}$$

The solution to the first step is obtained by the minimizer function tabulated in Table 6.2, and the optimum for the second step is obtained by setting the derivative of the objective function equal to zero, i.e.,

$$\frac{dJ}{dR^*} = 0 \Rightarrow \sum_{m=1}^{M} \alpha_m \|R^* - R_m\| = 0$$
$$\Rightarrow \sum_{m=1}^{M} \alpha_m R^* = \sum_{m=1}^{M} \alpha_m R_m$$
$$\Rightarrow R^* = \sum_{m=1}^{M} \lambda_m R_m, \quad \text{where } \lambda_m = \frac{\alpha_m}{\sum_{j=1}^{M} \alpha_j}. \tag{6.43}$$

Thus, the final aggregated rankings are computed as the weighted sum of all the MCDM rankings, with the weights $\lambda \in R^n$ being computed by the minimizer function. Interestingly, the weights of MCDM rankings in (6.43) are non-zero and fulfill the unit-sum property, which are the requirements for the MCDM methods. Note that the optimization problem is unconstrained and these properties are satisfied, thanks to the use of the HQ functions. Algorithm 5 summarizes the overall procedure of the proposed ensembling the MCDM methods.

Algorithm 5 Ensemble Ranking

Input: Rankings R_m, $m = 1,2,...,M$.
while *NotCongverged* **do**
  $\alpha_m = \delta(\|R_m - R^*\|), \quad m = 1,2,...,M$
  $\lambda_m = \alpha_m / \sum_j \alpha_j \quad m = 1,2,...,M$
  $R^* = \sum_m \lambda_m R_m$
end while
Output Final Ranking R^*, α

The following lemma guarantees the convergence of this algorithm.

Lemma 11 *The sequence* $\left\{\left(\alpha^k, R^{*k}\right), k = 1,2,...\right\}$ *generated by Algorithm 5, where k is the value at the* k^{th} *iteration, converges.*

Proof. The function $\delta(.)$ has the following property [6]

$$J(\alpha^{k+1}, R^{*k+1}) \leq J(\alpha^k, R^{*k+1}). \tag{6.44}$$

where R^* is assumed to be fixed. Similarly, the sequence of R^* is decreasing since J is convex, e.g.,

$$J(\alpha^{k+1}, R^{*k+1}) \leq J(\alpha^{k+1}, R^{*k}). \tag{6.45}$$

Thus, the sequence

$$\{...., J(\alpha^k, R^{*k}), J(\alpha^{k+1}, R^{*k}), J(\alpha^{k+1}, R^{*k+1}), ...\}$$

converges as $k \to \infty$ since J is bounded. □

Remark 12 *The proposed ensemble method is predicated on the fact that proper ranking methods are used, since the final aggregated rankings are naturally dependent on the ranking methods in question. If we add or remove a ranking method, the aggregated rankings are likely to change. However, in cases which include a significant number of methods, the proposed method is much less sensitive to adding or removing a ranking method. As such, the proposed method can be particularly useful in voting systems which usually contain a considerable number of votes.*

6.5.3. CONSENSUS INDEX AND TRUST LEVEL

The weights of each MCDM method differs with respect to the HQ function in question, since $\delta(.)$ relies on the $g()$ function. Consequently, various HQ functions would result in different weights and different final aggregated rankings. Among the HQ functions, the Welsch estimator has shown a promising performance in a number of domains [23, 24]. Interestingly, it is possible to obtain a consensus index and trust level using this estimator, owing to its use of the Gaussian distribution in the formulation. Prior to obtaining the consensus index and trust level, we first need to discuss how to select the parameter σ in the Welsch estimator. As a recent study has indicated [23], the parameter of this estimator can be tuned recursively in each iteration as,

$$\sigma = \frac{\sum_{m=1}^{M} \|R_m - R^*\|^2}{2q^2}. \tag{6.46}$$

After computing σ in the optimization procedure, we now discuss the consensus index and the trust level of the final rankings obtained by Algorithm 5.

Definition 13 (Consensus Index) *A consensus index C is an indicator to show the extent to which all MCDM methods agree upon the final rankings.*

The key element in this definition is that the consensus index shows the agreement among all the ranking methods being used, allowing us to compute the similarity of each ranking with the final aggregated rankings, thanks to the Welsch estimator. As a result, the consensus index C of given final rankings R^* with respect to rankings R_m, $m = 1, 2, ..., M$ can be computed as

$$C(R^*) = \frac{1}{KM} \sum_{k=1}^{K} \sum_{m=1}^{M} \zeta_{km}, \quad \zeta_{km} = \frac{\mathcal{N}_\sigma(R_m^* - R_{km})}{\mathcal{N}_\sigma(0)}, \tag{6.47}$$

where R_{km} is the ranking of the k^{th} alternative (alignment system) by the m^{th} ranking method, $\mathcal{N}_\sigma(.)$ is the probability density function of the normal distribution with a mean of zero and a standard deviation of σ, and $\mathcal{N}_\sigma(0)$ is used to normalize the similarity computation, thus $\zeta_{km}, C(R^*) \in [0,1]$.

The consensus index is one if all the ranking methods are the same, and it decreases as the ranking methods deviate from each other. Thus, if there is one ranking method different from the rest, it can adversely affect the consensus index. At the same time, this distinct ranking method is treated as an outlier in the HQ functions being used. As a result, it will have less impact on the final rankings, while it can profoundly influence the consensus index. We now define another indicator for the ensemble of MCDM methods that is much less sensitive to the distinct rankings.

Definition 14 (Trust Level) *A trust level T for the ensemble methods such as Algorithm 5 is the degree to which one can accredit the final aggregated rankings.*

The trust level is an indicator of reliability of the final ranking. For instance, if there is an MCDM ranking method that deviates significantly from the majority of rankings, it must be assigned a lower weight in Algorithm 5, and consequently, has less of an impact on the final rankings. Since the weight of such a method is lower than that of the other methods, it should also have less impact on the trust level. Taking this into account, the trust level can be computed as

$$T(R^*) = \frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \lambda_m \zeta_{km}, \tag{6.48}$$

where $\sigma, \lambda_m, m = 1, ..., M$, and R^* are computed based on Algorithm 5. Thus, the trust level is distorted to a lesser extent by a ranking method that are different from the majority of rankings, and it is a measurement of the reliability of the aggregated ranking R^* computed by Algorithm 5. It is evident from equation (6.48) that the trust level is equivalent to the consensus index if the weights of MCDM methods, i.e., $\lambda_m, m = 1, 2, ..., M$, are identical.

6.6. EXPERIMENTS

In this section, the MCDM methods and the proposed ensemble methodology are applied to five OAEI tracks and the alignment systems are compared and ranked accordingly. For each track, we first evaluate the performance metrics based on experts' preferences and then obtain the rankings of alignment systems. For information about the tracks as well as performance metrics, see Section 2.2.6 on page 25.

ANATOMY TRACK

We first apply the methodology to the OAEI anatomy track, for which 10 experts filled its associated survey of evaluating the performance metrics that includes execution time, precision, recall, consistency, and recall+. One expert identified precision, recall, and recall+ as the most essential metrics by assigning one to the pairwise comparison associated to these metrics, one expert selected solely precision and another picked recall

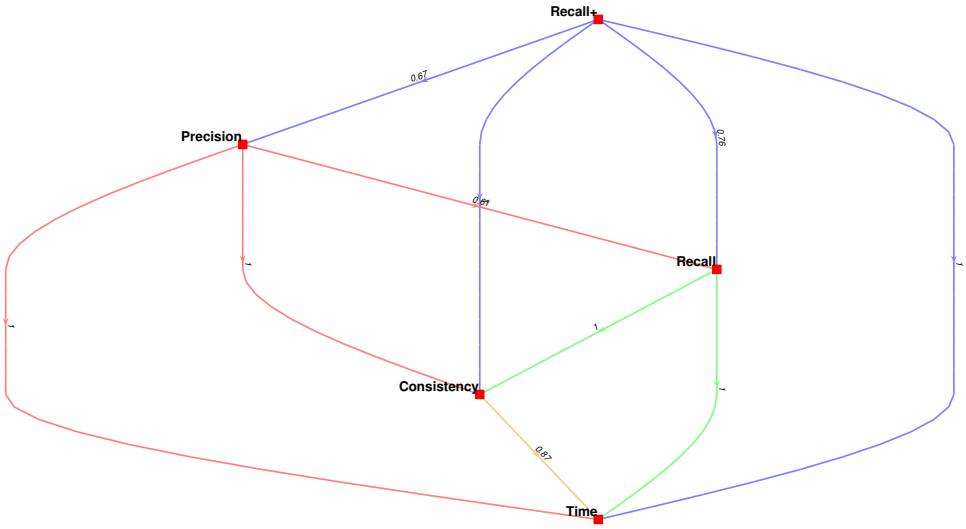


Figure 6.4: The credal ranking of performance metrics for the anatomy track.

alone. Among others, five experts identified only recall+ as the most important performance metric and the remaining two experts opted for consistency. In addition, five of the experts picked time as the least important, one selected consistency and time together, two recall+, and three experts opted for consistency as the least important performance metrics for this track. We applied the Bayesian BWM to the preferences of all experts to compute the aggregated priorities as well as the credal ranking of performance metrics. We summarize the outcome of credal ranking in a weighted, directed graph, where nodes are the performance metrics and each edge $A \overset{\nu}{\Rightarrow} B$ indicates that performance metric A is more important than B with confidence ν. Figure 6.4 shows the credal ranking of performance metrics for the anatomy track. Based on this figure that reflects the aggregated preferences of all 10 experts, recall+ is the most important metric, followed by precision and recall. Consistency and time are also the least important metrics according to the preferences of all experts.

Table 6.3 displays the rankings of the systems in the anatomy track computed by three MCDM methods, the final rankings are obtained by using the proposed ensemble method, as well as the weights of performance metrics and MCDM methods at the last row. The consensus index and trust level for this track are 0.95 and 0.97, respectively. Based on this table, AML, XMap, and LogMap are the top three systems in the anatomy track.

CONFERENCE TRACK

The performance metrics considered for the conference track are precision, recall, consistency, and conservativity. 11 experts filled the survey of this track, five of whom selected precision as well as recall as the most important performance metrics. In addition, two other experts picked only recall as the most important performance metric, two ex-

Table 6.3: Rankings of 14 systems participated in the OAEI anatomy track.

	Time (s)	Precision	Recall	Recall+	Consist.	TOPSIS	VIKOR	PROM	R^*	Final
LogMapBio	808	0.888	0.908	0.756	1	3	8	3	4.30	4
DOME	22	0.997	0.615	0.009	0	13	13	9	11.40	13
POMAP++	210	0.919	0.877	0.695	0	5	5	4	4.60	5
Holontology	265	0.976	0.294	0.005	0	14	14	14	14.00	14
ALIN	271	0.998	0.611	0.000	1	11	11	12	11.40	12
AML	42	0.950	0.936	0.832	1	1	1	1	1.00	1
XMap	37	0.929	0.865	0.647	1	2	2	2	2.00	2
LogMap	23	0.918	0.846	0.593	1	4	3	5	4.14	3
ALOD2Vec	75	0.996	0.648	0.086	0	12	10	11	11.08	11
FCAMapX	118	0.941	0.791	0.455	0	8	4	8	6.96	6
KEPLER	244	0.958	0.741	0.316	0	9	6	10	8.62	9
LogMapLite	18	0.962	0.728	0.288	0	10	7	6	7.62	8
SANOM	49	0.888	0.844	0.632	0	6	9	7	7.18	7
Lily	278	0.872	0.795	0.518	0	7	12	13	10.70	10
weight	0.0913	0.261	0.248	0.283	0.115	0.34	0.26	0.4		

* Consensus Index = 0.96
* Trust Level = 0.96

perts precision, and three experts consistency. Furthermore, one experts picked consistency as well as conservativity as the least important metrics, two opted for consistency alone, and the remaining eight experts selected conservativity as the least important performance metrics. Figure 6.5 displays the credal ranking of performance metrics for this track. According to this figure, precision and recall are the most important performance metrics and are significantly more important than consistency and conservativity.

Table 6.4 tabulates the rankings of the systems in the conference track with respect to multiple performance scores and their importance. Based on this table, AML, LogMap, and ALIN are the top systems, and KEPLER, Lily, and FCAMapX are the bottom three systems in this track. The last row of this table also shows the weights of performance metrics and MCDM methods. The weight of VIKOR is quite insiginificant, since its rankings are different from those of TOPSIS and PROMETHEE, whose difference in ranking the systems are not significant, e.g., TOPSIS and PROMETHEE rankings of KEPLER, ALIN, FCAMapX, LogMapLite, ALOD2Vec, and Lily are identical and rankings of SANOM, LogMap, DOME, and Holontology are more similar compared to the rankings of VIKOR. As a result, VIKOR is treated as an outlier and is assigned a much lower weight, making it have less contributions to the final aggregated rankings.

LARGE BIOMED TRACK

The performance metrics to rank the systems participated in this track are the execution time, precision, recall, and consistency. Out of 13 experts, 10 have filled the survey of this track. Four experts identified precision and recall together as the best by assigning one to the associated pairwise comparison, two experts solely precision, two solely recall, and two consistency. In addition, seven experts picked consistency as the least

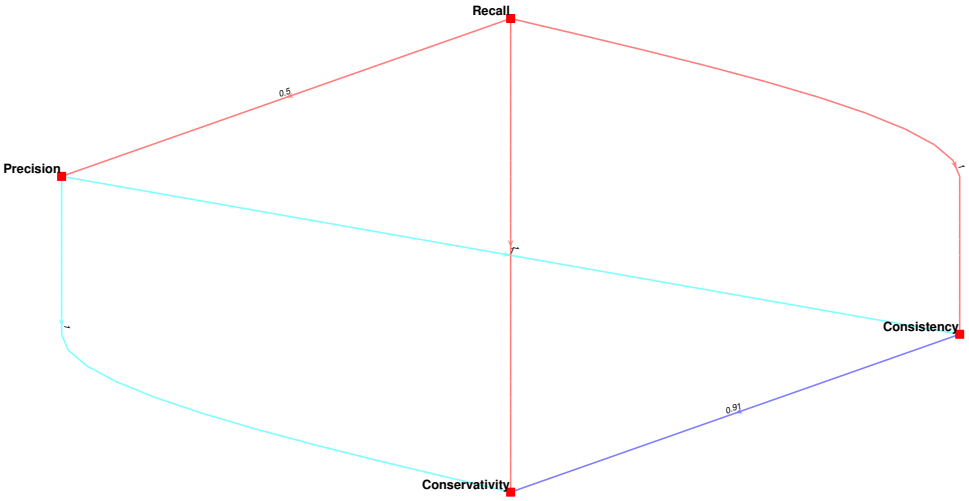


Figure 6.5: The credal ranking of performance metrics for the conference track.

Table 6.4: Rankings of the systems in the OAEI conference track.

	Precision	Recall	ConserViol	ConsisViol	TOPSIS	VIKOR	PROM	R^*	Final
SANOM	0.78	0.76	5.15	4.6	7	2	6	6.12	7
AML	0.83	0.7	1.86	0	1	1	2	1.45	1
LogMap	0.84	0.64	1.19	0	2	3	1	1.64	2
XMap	0.81	0.61	2.65	0.7	4	4	7	5.35	5
KEPLER	0.76	0.61	5.86	7.57	10	5	10	9.57	10
ALIN	0.88	0.54	0.1	0	3	6	3	3.26	3
DOME	0.88	0.54	5.05	0.48	6	8	5	5.73	6
Holontology	0.86	0.55	3.14	0.48	5	7	4	4.73	4
FCAMapX	0.71	0.61	5.9	13	12	9	12	11.74	12
LogMapLite	0.84	0.54	4.57	1.19	8	10	8	8.17	8
ALOD2Vec	0.85	0.54	5.9	1.29	9	11	9	9.17	9
Lily	0.59	0.63	7	6.2	11	12	11	11.09	11
weight	0.35	0.35	0.13	0.17	0.46	0.09	0.45		

* Consensus Index = 0.95
* Trust Level = 0.98

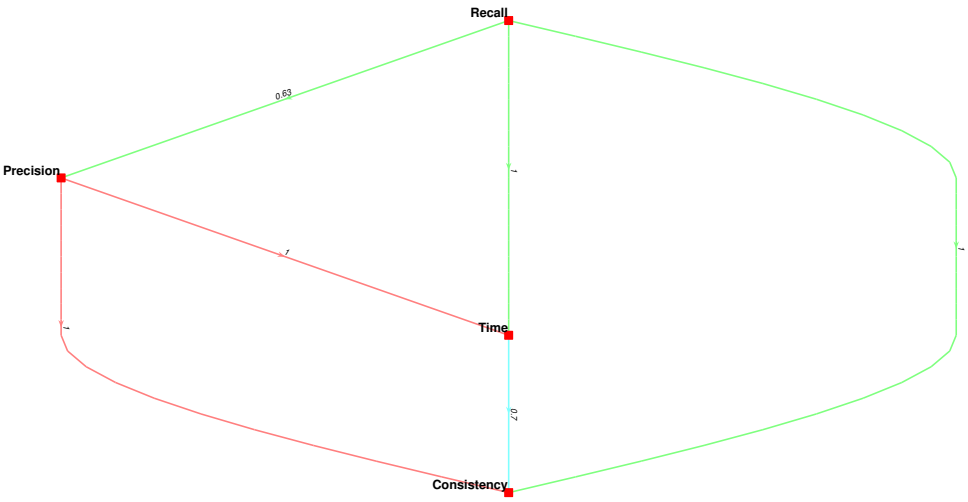


Figure 6.6: The credal ranking of four performance metrics for LargeBio track.

important performance metric and three opted for time. Figure 6.6 displays the credal ranking of four performance metrics of this track. According to this figure, recall is the most important performance metric for this track, followed by precision. Both precision and recall are significantly more important than time, that is itself more important than consistency.

We now apply the MCDM outranking methods to rank the alignment systems. Table 6.5 tabulates the rankings of eight systems that align FMA to NCI. This is an interesting case for the ensemble method, since the MCDM rankings are conflicting. In particular, the rankings of VIKOR and PROMETHEE are in line for LogMapBio as well as FCAMAPX, and are both different from the ranking of TOPSIS, while rankings of TOPSIS and VIKOR agree for LogMapLite and XMap and are distinct from that of PROMETHEE. By considering the weights of MCDM methods, it is interesting to consider that the weight of VIKOR

Table 6.5: Rankings of systems participated in the Large BioMed track for mapping FMA to NCI.

	Time(s)	Precision	Recall	TOPSIS	VIKOR	PROM	R^*	Final Rank
AML	55	0.84	0.87	1	1	1	1	1
LogMap	51	0.86	0.81	2	2	2	2	2
LogMapBio	1072	0.83	0.83	7	6	6	6.0002	6
XMap	65	0.88	0.74	3	3	4	3.0002	3
FCAMapX	881	0.67	0.84	6	7	7	6.9999	7
LogMapLt	6	0.68	0.82	4	4	3	3.9999	4
DOME	12	0.8	0.67	5	5	5	5	5
weights	0.141	0.356	0.378	0.00	1.00	0.00		

* Consensus Index = 0.81
* Trust Level = 1.00

Table 6.6: Rankings of systems participated in the Large BioMed track for mapping FMA to SNOMED.

	Time (s)	Precision	Recall	TOPSIS	VIKOR	PROM	R^*	Final Rank
FCAMapX	1736	0.82	0.76	6	5	5	5.00	5
AML	94	0.88	0.69	1	1	1	1.00	1
LogMapBio	1840	0.83	0.65	7	7	6	6.95	7
LogMap	287	0.84	0.64	2	2	4	2.08	2
XMap	299	0.72	0.61	3	6	7	6.02	6
LogMapLt	9	0.85	0.21	5	4	3	3.96	4
DOME	20	0.94	0.20	4	3	2	2.96	3
weights	0.141	0.356	0.378	0.0056	0.9502	0.0442		

* Consensus Index = 0.80
* Trust Level = 0.98

Table 6.7: Rankings of systems participated in the Large BioMed track for mapping SNOMED to NCI.

	Time (s)	Precision	Recall	TOPSIS	VIKOR	PROM	R^*	Final Rank
AML	168	0.90	0.67	1	1	1	1	1
FCAMapX	2377	0.80	0.68	6	4	5	4.07	4
LogMapBi	2942	0.85	0.63	7	6	6	6.02	6
LogMap	475	0.87	0.60	3	2	3	2.05	2
LogMapLt	11	0.80	0.57	2	3	4	3.00	3
DOME	24	0.91	0.48	4	5	2	4.90	5
XMap	427	0.64	0.58	5	7	7	6.95	7
weights	0.141	0.356	0.378	0.0255	0.9490	0.0255		

* Consistency Index = 0.95
* Trust Level = 0.95

is quite high and is close to one, while the weights of the other two methods are small and close to zero. This means that the proposed ensemble method favors the middle ground ranking among these three MCDM methods. Since the two methods have different rankings compared to the aggregated final rankings, the consensus index is not high and is around 0.80. At the same time, the trust level is 1.00 because the weights of the two other MCDM methods are nearly zero so that they cannot affect this indicator. This table shows that AML, LogMap, and XMap are ranked as the top three systems in this task.

In addition, Table 6.6 shows the rankings of participants for the alignment of FMA and SNOMED. This table is also similar to Table 6.5, since VIKOR takes a higher weight compared to the other methods as its rankings are in between the other rankings. The consensus index for the final rankings is 0.80, while the trust level hits 0.98. Similarly, Table 6.7 shows the rankings of the eight systems participated in matching SNOMED to NCI. According to this table, VIKOR again takes a bigger weight, and consequently, the final consensus index is 0.80, while it has the trust level of 0.98. According to Tables 6.6 and 6.7, AML and LogMap are the top two systems in both aligning FMA to SNOMED and SNOMED to NCI.

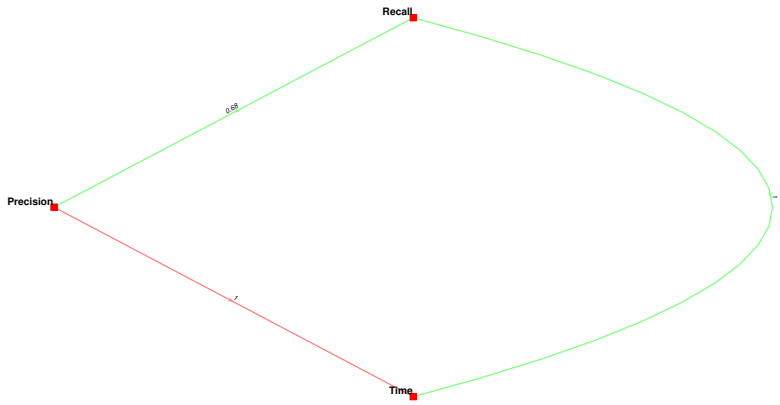


Figure 6.7: The credal ranking of performance metrics for the disease and phenotype track.

Table 6.8: Rankings of eight systems in the OAEI disease and phenotype track. The task involves the alignment of HP to MP.

	Time (s)	Precision	Recall	TOPSIS	VIKOR	PROM	R^*	Final Rank
LogMap	31	0.875	0.835	1	2	1	1.89	2
LogMapBio	821	0.862	0.841	3	3	4	3.00	3
AML	70	0.889	0.801	2	1	2	1.11	1
LogMapLite	7	0.993	0.609	4	4	3	4.00	4
POMAP++	1668	0.855	0.575	5	5	8	5.01	5
Lily	4749	0.682	0.647	6	8	7	7.78	8
XMap	20	0.994	0.314	7	6	5	6.10	6
DOME	46	0.997	0.308	8	7	6	7.10	7
weight	0.12	0.42	0.46	0.11	0.89	0.00		

* Consensus Index = 0.85
* Trust Level = 0.95

DISEASE AND PHENOTYPE TRACK

The performance metrics considered for this track are execution time, precision, and recall. Nine experts participated in evaluating the metrics for this track, six of whom selected precision as well as recall, and the remaining three experts identified recall alone as the most important performance metric. In addition, all experts identified the execution time as the least important metric for this track. Figure 6.7 plots the credal ranking of performance metrics for the disease and phenotype track. According to this figure, recall is more important than precision, and both are significantly more important than time based on experts' preferences.

Table 6.8 illustrates the rankings of the systems in the OAEI disease and phenotype track for mapping HP and MP ontologies. According to this table, the weight of PROMETHEE is nearly zero and that of TOPSIS is insignificant compared to VIKOR. This is because the rankings obtained by PROMETHEE deviate more from the other two methods. For instance, PROMETHEE ranked LogMapBio as four, while the other two

Table 6.9: Rankings of systems participated in the 2018 OAEI disease and phenotype track. The task is about the alignment of DOID and ORDO.

	Time (s)	Precision	Recall	TOPSIS	VIKOR	PROM	R^*	Final Rank
LogMap	25	0.937	0.775	1	1	3	1.05	1
LogMapBio	1891	0.898	0.799	2	2	2	2.00	2
POMAP++	2264	0.874	0.798	3	3	6	3.07	3
LogMapLite	7	0.988	0.615	4	4	1	3.93	4
XMap	15	0.969	0.548	5	5	7	5.05	5
KEPLER	2746	0.883	0.573	9	7	9	8.05	9
Lily	2847	0.589	0.783	8	8	8	8.00	8
AML	135	0.514	0.87	6	6	5	5.98	6
DOME	10	0.996	0.437	7	9	4	7.88	7
weight	0.12	0.42	0.46	0.50	0.48	0.02		

* Consensus Index = 0.90
* Trust Level = 0.98

consider it as the third alignment systems. As a result of such differences, the weight of PROMETHEE has become nearly zero. In addition, the rankings of VIKOR is an intermediate of those of TOPSIS and PROMETHEE, e.g., rankings of DOME and XMap, making it being assigned a higher weight. The consensus index for this ranking is 0.85 and its trust level is 0.95. Also, this table indicates that AML, LogMap, and LogMapBio are the top systems in this mapping task.

Another matching task in this track is the alignment of DOID and ORDO ontologies. Table 6.9 tabulates the rankings of the participating systems for this task. According to this table, TOPSIS and VIKOR are assigned significant weights, since their ranking are almost identical to each other. PROMETHEE has distinct rankings from VIKOR and TOPSIS, and is thus treated like an outlier and is assigned an insignificant weight. The consensus index and trust level of the rankings are 0.90 and 0.98, respectively. Accordingly, LogMap, LogMapBio, and POMAP++ are the top systems on this task regarding all the performance scores.

SPIMBENCH TRACK

The performance metrics for this track are precision, recall, and execution time. Six experts filled the survey of this track, four of whom selected both precision and recall as the most important performance metrics, while the remaining two picked only recall. In addition, all experts unanimously opted for execution time as the least important metric for this track. Figure 6.8 plots the credal ranking of performance metrics for the SPIMBENCH track. According to this figure, recall is the most important metric, followed by precision and time.

Tables 6.10 and 6.11 tabulates the rankings of the systems for the Sandbox and Mainbox tasks, respectively. The rankings of systems for these matching tasks is interesting, since two MCDM methods have identical rankings, while the other, i.e., TOPSIS, differs in rankings of two systems. As a result of this difference, the weight of that MCDM method takes an insignificant weight, and the weights of the other two rankings are

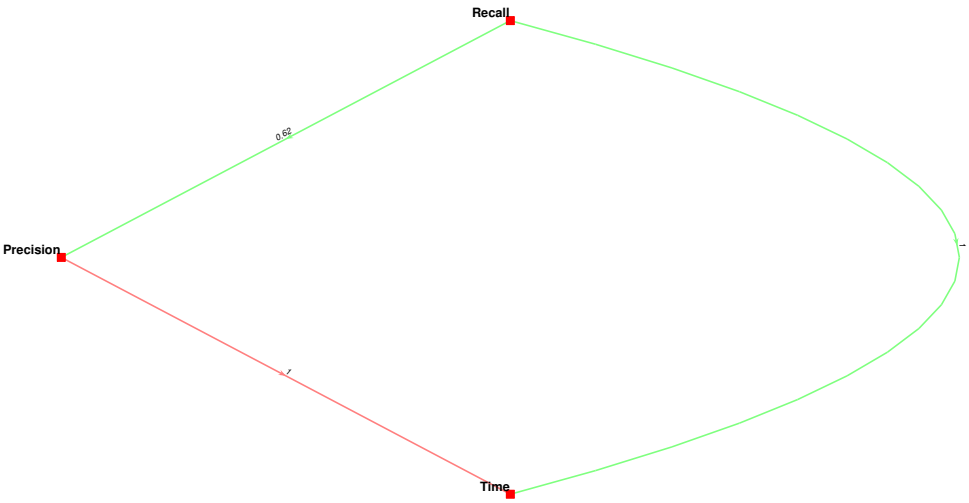


Figure 6.8: The credal ranking of performance metrics for the SPIMBENCH track.

Table 6.10: Rankings of systems participated in the 2018 OAEI SPIMBENCH track. The task is Sandbox.

	Time (s)	Precision	Recall	TOPSIS	VIKOR	PROM	R^*	Final Rank
AML	6220	0.8348	0.8963	2	3	3	3	3
Lily	1960	0.8494	1	1	1	1	1	1
LogMap	5887	0.9382	0.7625	3	2	2	2	2
weight	0.12	0.42	0.46	0.00	0.50	0.50		

* Consensus Index = 0.78
* Trust Level = 1.00

around 0.50. The consensus index for the final rankings is 0.77, but its trust level is 1.00, since the final rankings are precisely the rankings (or the average) of the other two MCDM methods. According to these tables, Lily is the top systems in both tasks, followed by LogMap and AML.

Remark 15 *We discussed the rankings of TOPSIS, VIKOR, and PROMETHEE for different OAEI tracks. Each of these methods had a higher weight in some of the tracks and had a lower weight in some others. However, the aim of this study is not the comparison of MCDM methods or discussion about their suitability. These method can take on higher or lower weights in different decision-making problems, and their weights are entirely dependent on the computed rankings based on the performance matrix of that particular decision-making problem.*

DISCUSSION

As we discussed before, the consensus index and the trust level indicate two different aspects of the final aggregated rankings. Generally speaking, the higher values of both indicators are desired. The consensus index is an indicator of the agreement among all

Table 6.11: Rankings of systems participated in the 2018 OAEI SPIMBENCH track. The task is Mainbox.

	Time(s)	Precision	Recall	TOPSIS	VIKOR	PROM	R^*	Final Rank
AML	37190	0.8385	0.8835	2	3	3	3	3
Lily	3103	0.8546	1	1	1	1	1	1
LogMap	23494	0.8925	0.7094	3	2	2	2	2
weight	0.12	0.42	0.46	0.00	0.50	0.50		

* Consensus Index = 0.78
* Trust Level = 1.00

the employed MCDM methods, while the trust level shows the reliability on the final aggregated weights. Below, based on the main properties of the proposed approach and the findings of the experiments, we elaborate on some general possible outcomes of the proposed ensemble method.

- **Consensus index high, trust level high:** If all the employed MCDM methods have identical rankings, their weights are analogous and is equivalent to $1/M$, where M is the number of ranking methods. In this case, the final aggregated rankings are precisely the average of the other rankings. As a result, the proposed ensemble method boils down to the averaging, or equivalently, the HQ functions operate as the Euclidean norm. This is indeed acceptable since no outlier exists when all rankings are identical. In this case, as we have the full agreement among all the employed MCDM methods, both consensus index and trust level equal to one.

- **Consensus index low, trust level high:** A low consensus index and a high trust level are the consequences of either one of the two cases. First, if a small fraction of the employed MCDM methods deliver distinct rankings from the majority of the rankings, then the proposed ensemble method treats them as outliers so that they take on lower weights. Consequently, those MCDM methods have zero or little impact on the final aggregated rankings. The existence of such methods can be detected by inspecting the weights obtained by the proposed ensemble method. The method with a lower weight is deemed as a deviation from the majority of MCDM rankings and also from the final rankings, and is thus treated as an outlier. The second case is when the number of methods with lower weights is significant compared to the number of all the employed MCDM methods. The MCDM rankings with higher weights are the ones that are the intermediate of all the methods. As a result, the intermediate rankings take on higher weights and more profoundly impact the final aggregated rankings. In both of these cases, the agreement among the employed MCDM methods is low, while the final rankings are fully captured by a fraction of the employed MCDM methods, which is why the consensus index is insignificant and the trust level is high.

- **Consensus index low, trust level low:** If all the employed MCDM rankings deviate significantly from each other, then the consensus index will be low. In this case, there is no fraction of the employed MCDM methods with significantly higher weights. Thus, the trust level is also low.

- **Consensus index high, trust level low:** This case does not happen since the trust level is high if there is a consensus among the employed MCDM methods.

This is a general framework for discussion, and we think that the levels could be defined by the decision-makers for a particular problem.

6.7. CONCLUSION

This chapter studied the use of multi-criteria decision-making (MCDM) methods for ranking the ontology alignment systems with respect to multiple performance metrics and experts' preferences. In this regard, a survey was designed based on the best-worst method (BWM) to elicit the preferences of experts for different OAEI tracks.

In order to obtain an overall aggregated priorities of performance metrics (or, criteria in general) for a group of experts, a Bayesian model is developed for the group decision-making within the BWM. The proposed method models the inputs of the BWM using the multinomial distribution and it is demonstrated that such a distribution would preserve the underlying idea of the original BWM. Further, the weight vector is modeled using the Dirichlet distribution. The proposed Bayesian model is able to compute the weight distribution of each individual in the group decision-making, and an aggregated final distribution representing the overall preferences of all experts / decision-makers (DMs). The credal ranking for performance metrics or criteria is introduced based on which each pair of criteria are assigned a relation and a confidence. The proposed Bayesian BWM is a promising method in the context of group decision-making where one is interested in the collective opinions of a group, but at the same time, one could check the ranking of the metrics or criteria in a probabilistic sense. The group will be more certain about the relation of two criteria if it is associated with a high confidence level, while the relations with low confidence level should be interpreted more carefully

After computing the aggregated priorities or weights of all experts, another class of MCDM methods, outranking methods, were used to rank the alignment systems. Since the MCDM outranking methods may generate distinct and conflicting rankings, we introduced an ensemble method to aggregate the ranking and compute a final rankings for each alignment system. The proposed ensembling utilizes the half-quadratic (HQ) theory and is able to compute a final aggregated ranking, which was obtained as the weighted sum of the MCDM outranking methods. The weight in the proposed method was computed using the minimizer functions inspired in the HQ theory, but it satisfied the basic properties of weights in MCDM. We also introduced two indicators, namely, consensus index and trust level, where the former denotes the level of agreement among MCDM ranking methods and the latter shows the reliability of the ranking schemes. We observed for the cases that a ranking method is different from the rest, it has a low consensus index and but high trust level. As a result, these two indicators are able to delineate different properties of the final aggregated ranking.

There are several avenues to extend current research. It is possible to apply the Bayesian modeling to other important MCDM methods, and study the properties of the Bayesian MCDM models, such as consistency, experts clustering, and detecting DMs with distinct preferences.

Another avenue that is probably particular to ontology alignment is the use of distri-

butions introduced in the previous chapter instead of the score. To the best of our knowledge, there are a few MCDM outranking methods that can handle distributional inputs in the performance matrix. Since the aim is to compare and rank alignment systems, one possible way is to first apply the Bayesian test introduced in the previous chapter to compute the extent to which one alignment system is preferred to another. Then, a new MCDM method is required to work with the probability of differences among alignment systems, rather than working with their performance scores. Devising such a method can be used to rank the alignment systems over multiple benchmark as well, in contrast to the current approach that is applicable to one benchmark only.

REFERENCES

[1] J. Euzenat, P. Shvaiko, *et al., Ontology matching*, Vol. 18 (Springer, 2007).

[2] J. Rezaei, *Best-worst multi-criteria decision-making method,* Omega **53**, 49 (2015).

[3] J. Rezaei, *Best-worst multi-criteria decision-making method: Some properties and a linear model,* Omega **64**, 126 (2016).

[4] G.-H. Tzeng and J.-J. Huang, *Multiple attribute decision making: methods and applications* (Chapman and Hall/CRC, 2011).

[5] S. Opricovic, *Multicriteria optimization of civil engineering systems,* Faculty of Civil Engineering, Belgrade **2**, 5 (1998).

[6] M. Nikolova and M. K. Ng, *Analysis of half-quadratic minimization methods for signal and image recovery,* SIAM Journal on Scientific computing **27**, 937 (2005).

[7] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical distributions* (John Wiley & Sons, 2011).

[8] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice* (CRC press, 1995).

[9] M. Plummer, *Jags: Just another gibbs sampler,* (2004).

[10] M. A. Abo-Sinna and A. H. Amer, *Extensions of topsis for multi-objective large-scale nonlinear programming problems,* Applied Mathematics and Computation **162**, 243 (2005).

[11] Y. Cha and M. Jung, *Satisfaction assessment of multi-objective schedules using neural fuzzy methodology,* International Journal of Production Research **41**, 1831 (2003).

[12] T.-C. Chu, *Facility location selection using fuzzy topsis under group decisions,* International journal of uncertainty, fuzziness and knowledge-based systems **10**, 687 (2002).

[13] S. Opricovic and G.-H. Tzeng, *Compromise solution by mcdm methods: A comparative analysis of vikor and topsis,* European journal of operational research **156**, 445 (2004).

[14] S. Opricovic and G.-H. Tzeng, *Multicriteria planning of post-earthquake sustainable reconstruction,* Computer-Aided Civil and Infrastructure Engineering **17**, 211 (2002).

[15] J. Brans, *L'ingenierie de la decision, l'laboration d'instruments d'aidea la decision. colloque sur l'aidea la decision,* Faculte des Sciences de l'Administration, Universite Laval (1982).

[16] B. Soylu, *Integrating prometheeii with the tchebycheff function for multi criteria decision making,* International Journal of Information Technology & Decision Making **9**, 525 (2010).

[17] D. Geman and G. Reynolds, *Constrained restoration and the recovery of discontinuities,* IEEE Transactions on Pattern Analysis & Machine Intelligence , 367 (1992).

[18] D. Geman and C. Yang, *Nonlinear image recovery with half-quadratic regularization,* IEEE transactions on Image Processing **4**, 932 (1995).

[19] H. Wang, H. Li, W. Zhang, J. Zuo, and H. Wang, *Maximum correntropy derivative-free robust kalman filter and smoother,* IEEE Access **6**, 70794 (2018).

[20] M. E. Mann and J. M. Lees, *Robust estimation of background noise and signal detection in climatic time series,* Climatic change **33**, 409 (1996).

[21] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, *Two-stage nonnegative sparse representation for large-scale face recognition,* IEEE transactions on neural networks and learning systems **24**, 35 (2013).

[22] R. He, T. Tan, and L. Wang, *Robust recovery of corrupted low-rankmatrix by implicit regularizers,* IEEE transactions on pattern analysis and machine intelligence **36**, 770 (2014).

[23] R. He, W.-S. Zheng, T. Tan, and Z. Sun, *Half-quadratic-based iterative minimization for robust sparse representation,* IEEE transactions on pattern analysis and machine intelligence **36**, 261 (2014).

[24] R. He, Y. Zhang, Z. Sun, and Q. Yin, *Robust subspace clustering with complex noise,* IEEE Transactions on Image Processing **24**, 4001 (2015).

[25] M. Mohammadi and G. A. Hodtani, *A robust acgh data recovery framework based on half quadratic minimization,* Computers in biology and medicine **70**, 58 (2016).

[26] M. Mohammadi, G. A. Hodtani, and M. Yassi, *A robust correntropy-based method for analyzing multisample acgh data,* Genomics **106**, 257 (2015).

[27] S. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).

[28] J. Demšar, *Statistical comparisons of classifiers over multiple data sets,* Journal of Machine learning research **7**, 1 (2006).

[29] P. J. Huber, *Robust statistics* (Springer, 2011).

7

INTEROPERABILITY IN LOGISTICS: AN ONTOLOGY ALIGNMENT APPROACH

The e-business infrastructure is built upon a variety of applications running on different platforms across different networking protocols – and they all need to share data. Universal access to data is a quantum leap towards the industry's goal of interoperability.

Linda Sanford

This chapter is dedicated to applying ontology alignment systems to the heterogeneity problem in logistics. The primary motivation for doing so is to enable interoperability among different IT systems in logistics that potentially use distinct standards or information systems. We first study different logistics standards that are implemented by XML schema definition (XSD) so that they do not represent the semantic relations between different entities. Accordingly, we create two ontologies from two well-known standards that are rich in terms of semantic relations between entities. Afterward, the ontologies are subjected to the state-of-the-art alignment systems to verify the applicability of ontology alignment to logistics interoperability. We specifically use direct matching and indirect matching with background knowledge from logistics, and realize that the alignment with an annotated background knowledge has more reliable and acceptable results and is thus applicable to real-world situations.

7.1. INTRODUCTION

This chapter studies the heterogeneity in logistics and verifies the applicability of ontology alignment to address it. The supply and logistics sector consists of millions of large and Small and Medium-sized Enterprises (SMEs). Its size in the EU only is estimated 878 billion Euro in 2012[1] with over 1.2 millions of enterprises [1]. Besides proprietary developed software, these enterprises can choose to use Commercial-Off-The-Shelve (COTS) software from over 200 different suppliers[2], each of which has its proprietary database scheme that can even be customized by the users.

The challenge is and has been to integrate the business processes of enterprises and their supporting IT solutions. Development of open standards addresses this challenge. Although these standards were developed, they did not solve the heterogeneity problem. Different implementation guides of (different versions of) open standards have been developed [2], leading to implementations that are only interoperable with additional efforts and costs. These implementation guides support process interoperability [3], which implies they will support a particular function for interconnecting business processes. There are also different open standards providing the same business functionality, e.g., a transport order developed by UN/CEFACT or one of the Uniform Business Language (UBL). To address differences in implementation guides of different open standards with identical or similar functionality, commercial organizations provide transformation services between various data standards.

Applying open standards and commercial transformation services reduces the transformation challenge, but the development and implementation of implementation guides of these open standards still take too much time for implementing supply chain innovations like agility and resilience [4], and synchromodal planning [5]. In other words, the use of different standards or implementing the same standards differently would impose another heterogeneity, though not more severe than having no standards, that still needs to be addressed properly. To reduce the development and implementation time for interoperability between any two organizations, this chapter explores the application of ontology alignment. The holy grail of ontology alignment in supply and logistics is to create semi-autonomous alignments between database schemes of different organizations, thus enabling what one could call 'plug and play'[3]: Plug a database scheme into an open data sharing infrastructure and be able to share data with relevant business partners. Plug and play requires an open data sharing infrastructure providing standardized services [6].

There are several issues in applying ontology alignment to enable interoperability between different logistics stakeholders. First, the number of enterprises is significant, making the pairwise alignment between every two parties very time-consuming. For instance, if the aim is to enable interoperability among 1.2 millions logistics parties only in the EU, then we need to execute an alignment system 1.44×10^{12} times. Another challenge is that standards are usually modeled in XML schema definition (XSD) that conveys no or limited semantics of entities in the associated standards. As a result, it is required to create at least two ontologies from these XSD models in order to be able to verify the

[1] https://ec.europa.eu/transport/themes/logistics-and-multimodal-transport/logistics_en
[2] https://www.capterra.com/logistics-software/
[3] The Digital Transport and Logistics Forum (DTLF)

applicability of ontology alignment. In addition, the standards have different level of granularities: Some are mode specific, e.g., air or rail, and some are more general and encompass different modes from a higher view.

In this chapter, we first review the standards used in logistics that aim at solving the interoperability problem. We then investigate two specific standards and create ontologies based on their XSD models. We finally use ontology alignment to find the shared entities of ontologies, and experiment the usefulness of such an approach to address the interoperability in logistics.

7.2. INTEROPERABILITY BY OPEN LOGISTICS STANDARDS

Logistics interoperability has to be considered in the context of collaborating business processes of stakeholders, applicable standards to support interoperability between these business processes, and the technical paradigm for data sharing applied. Any choices made in these different areas may affect the applicability of ontology alignment in logistics. Therefore, we will briefly present an overview of these aspects.

7.2.1. LOGISTICS BUSINESS PROCESSES

International trade and logistics are characterized by moving cargo from one location to another with one or more means of transport, (temporary) storage of this cargo, and authorities governing these cargo flows from different legal perspectives like safety, security, and VAT compliance. Each modality and each cargo type have their specific characteristics. For instance, bulk cargo considers weights and volumes and containerized cargo a container with its size and type and container identification. Furthermore, sea containers have other characteristics than containers used for air transport. The latter are called Uniform Load Devices and exactly fit into an airplane. Different transport modalities also use different infrastructures with their hubs, have different transport documents, etc. There are also different enterprises and authorities involved, like a food and drug inspection agency for transport of agricultural cargo, coastal police for vessel movements, and air traffic control for air transport.

Modalities may have standardized the structure of data sets they share for digitization of their business processes (see next section). Each individual organization collaborating in logistics chains will have its own IT system with its own internal data structure. Interoperability is about integrating these heterogeneous IT systems. It implies that each organization will have its implementation of for instance a transport order. These internal data structures have to be matched with the ones for modalities or any de facto structures used by their customers or major service providers. Since the number of logistics enterprises is large, business process integration of all collaborating organizations is a challenge.

Two collaborating organizations will share multiple data sets for business process digitization, like a booking, a transport order – and plan, and an event for reporting the progress. Thus, integration complexity increases. It can also increase in solutions for all variants of logistics chains are developed. Furthermore, integration complexity increases by the number of logistics enterprises, which runs into millions globally, all (inter)national legislation with their governing authorities, and changes in trade agree-

ments that have impact on data requirements of authorities.

Ontology alignment might provide a solution to reduce integration complexity. Another approach for complexity reduction that we will investigate together with ontology alignment, is to abstract from these chains and specify bilateral data sharing [7]. Each chain is constructed by its links of any two collaborating organizations and the outsourcing rules applied by each individual organization. This reduces complexity and still enables (dynamic) chain composition. Any interactions between these collaborating organizations can be modelled as a business process choreography [8], where a semantic model specifies all data that can be shared. The semantic model, which can be represented by an ontology, is called a Canonical Information Model [9]. The choreography supports business functionality as developed by the DTLF:

- Publish, search, and find logistics capacity. By posting a particular goal, a customer can find available transport -, storage – or other type of logistics capacity. The capacity might be offered by timetables of for instance vessels (called: voyage) or trains.
- Booking and ordering. Whenever capacity or a logistics service provider has been found, booking or a request for quotation can be made. This can result in a framework contract followed by orders for individual shipments or a booking can directly be confirmed as an order.
- Supply chain visibility. This is about sharing relevant milestones of the progress of a logistics service, e.g. loading, departure and (estimated) arrival of a truck at its destination.

Each functionality is supported by interactions for data sharing, where these interactions are of a type. For instance, booking and ordering is supported by a booking, a booking confirmation, a transport order and a plan. These interaction types specify the minimal data that needs to be shared and maximal that can be shared in the context of the choreography. For instance, a transport order should at least contain one cargo item and two locations (acceptance and delivery) with their respective time windows for acceptance and delivery. This minimal – and maximal data set is formulated in terms of the semantic model. A first version of a semantic model has been constructed in various EU funded projects. This will be elaborated at a later stage in this paper.

7.2.2. OPEN STANDARDS AND THEIR IMPLEMENTATION

There are different logistics standards to enable the interoperability in the domain. There are different classifications for standards that categorize them into different levels. For instance, the European interoperability framework (EIF) [10] has four levels: technical, semantic, organizational, and legal. A more thorough model is the levels of conceptual interoperability model (LCIM) [3] that identifies six different levels, to which the levels of EIF can be mapped. Figure 7.1 the figure plots the standards and their pragmatic application in logistics to the LCIM. The figure shows that semantic models generate syntactical standards and can be used to create implementation guides (IGs). The technical interoperability is based on communication protocols so that the heterogeneity is

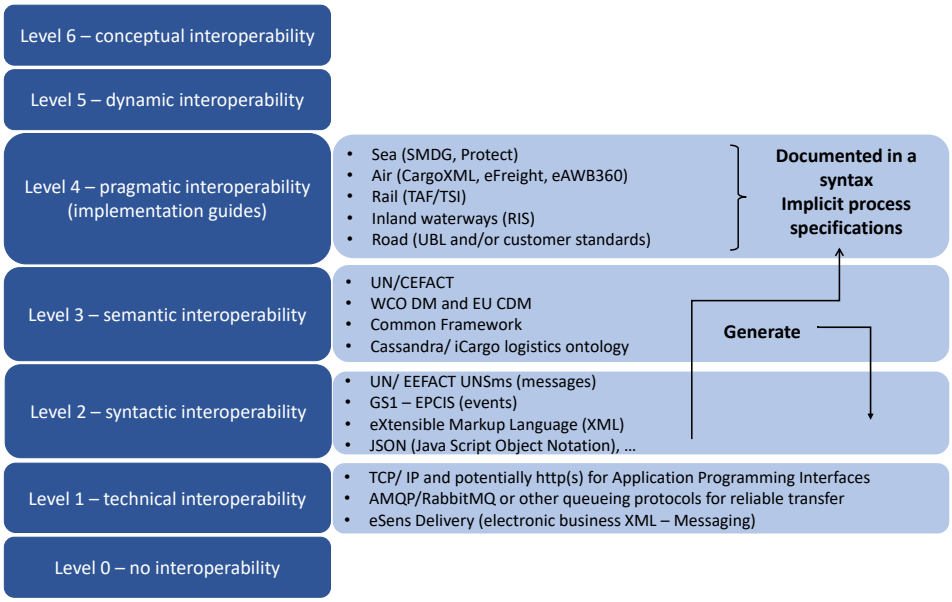


Figure 7.1: Overview of logistics standards [2].

trivial. To the best of our knowledge, there is no interoperability model for the top two levels, dynamic and conceptual interoperability. Therefore, three remaining LCIM levels are only explained in the following.

- **Syntactic Interoperability:** Syntactic Interoperability: This class refers to the structure of data during exchange. The syntax of a message governs that the use of protocols in technical interoperability level. XML and JSON are shown as potential syntaxes for syntactical interoperability. Another option is UN/CEFACT United Nations Standard Messages (UNSMs) as technical data structures representing specific interaction types, e.g. there is a UNSM for a transport order. These UNSMs use the EDIfact (Electronic Data Interchange for administration, commerce, and trade) as syntax [11]. UNSMs have lots of configuration options; they are generic.

- **Semantic Interoperability:** This class refers to the semantic representation of the data modelled, for instance, by unified modelling language (UML) or ontology web language (OWL). The models are mostly used to develop message structures for sharing data at syntactic level, for instance to automatically generate UNSMs or XSDs (XML Schema Definitions).

- **Pragmatic Interoperability:** it means that two or more stakeholders integrate their business processes. They start from existing processes and most often replace current procedures, which are mainly paper-based, with electronic versions. Business process integration can be based on level 2 standards; most often these

Figure 7.2: An example of the physical activities for a shipment from a consignor to a consignee.

collaborating business partners don't have access to data models (if they are available). In case the use existing open standards like UNSMs or XSDs, they configure these to their requirements resulting in so-called implementation guides of these open standards. In case these implementation guides are constructed by more than two organizations, they can be called community implementation guides, for example the ones specified for rail or sea. These community implementation guides can also be a basis for any two organizations constructing their bilateral implementation guides.

Thus, analysis of (open) standards and their implementation learns that 'closed' data sharing solutions are constructed. Open standards do not necessary lead to open ways of data sharing, meaning that it is easy to on-board organizations. On the one hand, implementation guides will be constructed and on the other hand they have to be implemented by matching to heterogeneous IT systems of organizations. Thus, having open standards only does not necessarily reduce enterprise interoperability complexity in logistics.

7.3. SEMANTIC LOGISTICS MODELS

In this section, two ontologies based on two logistics data models, eCMR and shipping instruction (SI), are created that also include the semantics among their different entities. eCMR is a standard that is used in the road transportation, while SI is for a transportation by sea. To show the importance of matching these two data structures, Figure 7.2 shows a typical transportation trip of a cargo from an origin to a destination. First, the consignor ships a cargo to a port, named here as the port of export. In this example, the transportation to the port of export is by road so that their associated shipping information is modeled by eCMR. Then, the cargo is transshipped to a port of import by sea, whose transshipping information is modeled by SI. Then, from the port of import, the cargo will be sent to the consignee by means of road, that again necessitates to store the information by the eCMR model. In this simple example, we need to transform the data first from eCMR to SI and then from SI back to eCMR. As a result, the alignment of these models are very essential in logistics. In the following, we first create two ontologies and then explore the alignment of the two ontologies together.

7.3.1. ELECTRONIC CMR ONTOLOGY

The CMR[4] is a United Nations convention that concerns with various legal issues concerning the transportation of a cargo by road. As of 2017, CMR has been ratified by most of the European states. The International Road Transport Union (IRU) developed a standard CMR waybill according to the CMR. As of 2008, it is also feasible to use an updated

[4]It stands for Convention on the Contract for the International Carriage of Goods by Road.

electronic consignment note, called eCMR. We create an ontology based on a subset of eCMR by using the terminology applied to the eCMR standard of UN-CEFACT, which is more than a data carrier and is able to contain semantic relations between entities.

The classes of the eCMR as well as their related properties in the created ontology are as follows (class names are identified by italic and bold characters and property names only by italic characters):

- ***LogisticsLocation*** includes the locations in a logistics trip and contains basic information for a location, such as name and country. The following two classes inherit from this class:
  - ***CarrierAcceptanceLogisticsLocation*** is the location where a cargo is picked up by a carrier.
  - ***ConsigneeReceiptLogisticsLocation*** is where a cargo will be delivered to a consignee.
- ***TradeParty*** represents different parties in a transaction. It has three main subclasses associated to different parties:
  - ***Consignor*** is the one that orders a consignment and must determine locations of acceptance and delivery.
  - ***Consignee*** is the one to whom a cargo must be delivered.
  - ***Carrier*** is the party that conducts the shipment by picking up the cargo from an acceptance place and delivering to a consignee receipt location.
- ***SupplyChainConsignmentItem*** includes the items in the shipment. This class has several relationships with other classes as:
  - The items are inside a transport cargo. Hence, we use object property *isInsideCargo* that relates this class to ***TransportCargo***.
  - The items are placed in a logistics package. Thus, we use object property *isPackagedIn* to relate this class to ***LogisticsPackage***. In addition, the packages must have shipping marks that makes ***LogisticsPackage*** have another object property *isMarkedAs* to link it to ***LogisticsShippingMarks***.
- ***SupplyChainConsignment*** is the main class in eCMR that represent the shipment process. It has multiple object properties that relates it to different classes and provide the necessary information for shipping a cargo. These object properties are as follows:
  - *includes* that relates it to ***SupplyChainConsignmentItem***.
  - *isDeliveredTo*, *isCarriedBy*, and *isOrderedBy* are with respect to the three different trade parties, ***Consignee***, ***Carrier***, and ***Consignor***, respectively.
  - *isPickedUpAt* and *isDroppedOffAt* relate the class to ***CarrierAcceptanceLogisticsLocation*** and ***ConsigneeReceiptLogisticsLocation***, respectively.

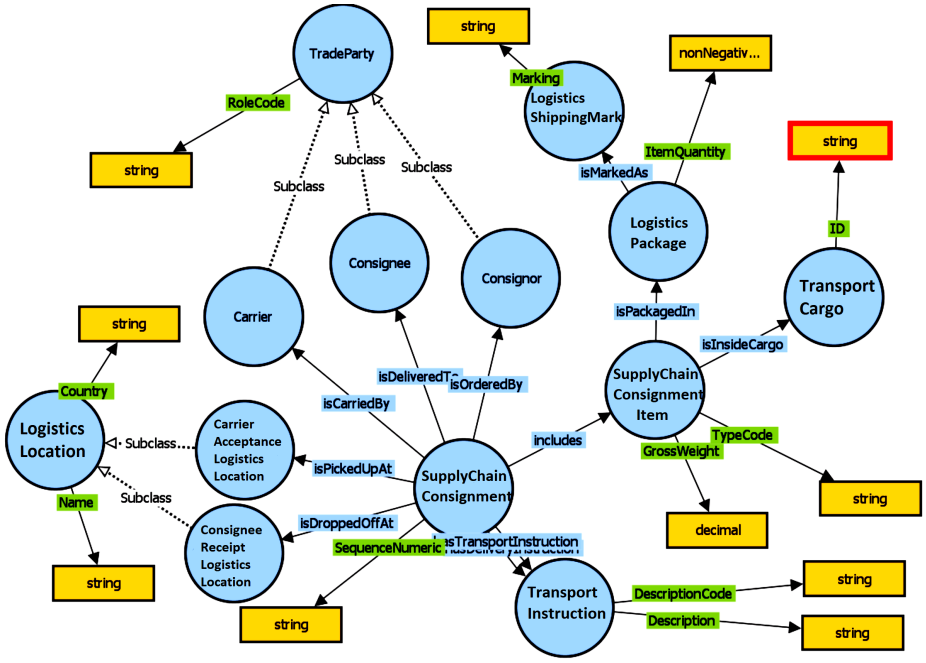


Figure 7.3: The eCMR ontology.

7

– *hasTransportInstruction* and *hasDeliveryInstruction* relate this class to transport instructions for the shipment and delivery, respectively. These instructions are necessary according to eCMR.

Figure 7.3 plots the eCMR ontology visualized by VOWL [12]. The classes are shown by circles and object properties are the labels of arrows of the two corresponding classes.

7.3.2. SHIPPING INSTRUCTION ONTOLOGY

A shipping instruction (SI) is a document, provided by a customer to a carrier, containing the details of a cargo to be shipped by sea and the requirements for its physical transportation. The fields in the document are the building blocks for creating the associated ontology that is based on the terminology used in SI. We particularly utilize the interface of a booking for shipment of containers by sea. The created ontology has a main class, ***ShippingInstruction***, that associates with other classes with proper object properties. These classes and properties are discussed in the following:

- It contains the information of locations where a container is picked up or delivered to. Thus, there are two object properties, *isPickedUpFrom* and *isDroppedOffAt* that associate it to ***PlaceOfReceipt*** and ***PlaceOfDelivery***. Each of these classes inherits from ***place*** that contains basic information of a place such as the name of the city and the United Nation location code (*CityUNLocationCode*). In addition, it contains the ports where a container is loaded and discharged, represented by classes

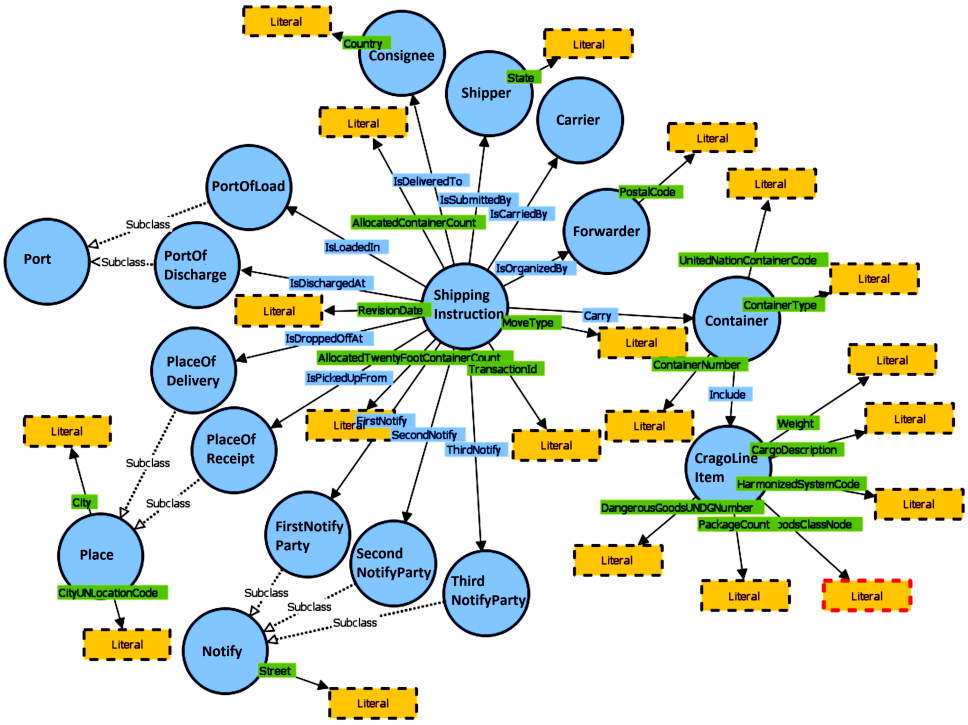


Figure 7.4: The SI ontology.

PortOfLoad and ***PortOfDischarge*** that are related to this class by object properties *isLoadedIn* and *isDischargedAt*, respectively. These classes also inherit from another class, named ***Port***, that stores the basic information of a port.

- Each ***ShippingInstruction*** contains three different parties that are modeled as classes with names ***Forwarder***, ***Shipper***, and ***Consignee*** that is related to with *isOrganizedBy*, *isSubmittedBy*, and *isDeliveredTo*, respectively.

- It also specifies the ***container*** that must be shipped. Thus, a class with the same name is created that is related to the SI class with object property *carry*. Besides, each container itself *includes* several ***CargoLineItem***, each of which represents a particular item in the container.

- According to SI, three different notification to three parties must be sent. The notification to parties is modeled with class ***NotifyParty***, that has three subclasses, ***FirstNotifyParty***, ***SecondNotifyParty***, and ***ThirdNotifyParty***, related to ***SI*** with object properties *FirstNotify*, *SecondNotify*, and *ThirdNotify*.

Figure 7.4 plots the SI ontology visualized by VOWL [12].

7.4. EXPERIMENTS

This section introduces the experiment where ontology alignment systems are applied for aligning two logistics ontologies. There are still choices to be made with respect to the experiments that will be discussed first.

7.4.1. CHOICES FOR EXPERIMENTS

Applying ontology alignment systems to logistics ontologies is different from that in the OAEI for some reasons. First of all, ontologies are not common in supply and logistics. Open standards, their implementation guides, and database schemes have to be transformed into ontologies to enable alignment. Second, the following alignment choices need to be considered:

- Database scheme alignment – one could consider the alignment between database schemes of different organizations. This option is not considered feasible, since databases provide more functionality than interoperability between two organizations; they support an organization in its business.
- Functional view alignment – this option considers creating a functional view of for instance a transport order on two database schemes that will be aligned. If ontology alignment would provide optimal results, this would be an ideal situation, since it does not require any formulation of open standards. It is however also complex, while it requires to align many structures, all using potentially different terminology.
- Open standard alignment – alignment of two open standards. This could be a first start which does not require any involvement of organizations (yet). Open standards are publicly available. However, the development of an ontology from an open standard might be complex, depending on the supported functionality. An open standard for a transport order may for instance cover all transport modalities and all types of cargo.
- Implementation guide alignment – alignment of implementation guides of an open standard. For this purpose, organizations will have to provide their implementation guides.
- Alignment with a Canonical Information Model –integration of IT applications of one organization can be via an upper ontology. This approach can also be applied for external integration, i.e., between IT applications of different organizations. It requires time for constructing an upper ontology for logistics, but in case the upper ontology can be used for automatic alignment between functional views of database schemes, it will support the 'plug and play'. There are different options using an upper ontology, like:
  - Alignment of a functional view with the upper ontology;
  - Alignment of an open standard with the upper ontology;
  - Alignment of an implementation guide of an open standard with the upper ontology.

For ontology alignment, the upper ontology acts as background knowledge that can boost the alignment outcome. We can conduct the experiment by aligning implementation guides of open standards with an ontology that has been developed in EU funded projects. This experiment can be completely controlled. The ontology that acts as an upper ontology is called LogiCO[5] [13], and an implementation guide of an existing open standard will be produced that is expected to contain concepts represented by LogiCO.

As a result, for the experiment, we use the notion of *indirect matching* discussed in Chapter 1 on page 8.

7.4.2. SETTING OF EXPERIMENTS

We conduct two different experiments. First, the eCMR and SI ontologies are directly aligned together without using background knowledge. Second, the experiment is conducted by the alignment of ontologies derived from the implementation guides of open standards by using LogiCO as an upper ontology in indirect matching. These choices will be further elaborated.

Using LogiCO will have some risks with respect to the experiment; it might not support the functionality of an implementation guide. To reduce this risk, an implementation guide of an open standard needs to be aligned as much as possible with LogiCO. Therefore, it is worthwhile to list the foundational concepts of LogiCO [13]:

- ***Activity*** denotes some action and is relevant for the purpose of logistics, such as, for example, the activities of transport, storage, transshipment, loading, and unloading. Some activities are atomic and can be used to compose more complex activities.

- ***Event*** represents an occurrence of interest for the execution of a certain activity. In contrast to an activity, which denotes an action that is continuous in time, an event denotes an occurrence at a specific moment in time. For example, the departure of transport means from a location of origin and its arrival to the destination can be regarded, respectively, as starting and ending events for the transport activity.

- ***Actor*** represents organizations, authorities or individuals that offer or require activities and operate on resources related to these activities. An actor can have a role, for example, customer and service provider, or shipper, consignee, forwarder, and carrier.

- ***Entity*** represents something that is used or exchanged during an activity. We specialize an Entity in a Spatial Entity, which represents tangible objects, such as an equipment or a person, and an Intangible Entity, which represents intangible objects, such as a modality, a characteristics or a dimension. We also define a Temporal Entity, which represents the start time, end time or time interval associated to activities and events. To this regard, since time is a basic (foundational) concept relevant for logistics, but common to other domains, we re-use the time on-

[5] LogiCO stands for **Logi**stics **C**ore **O**ntolgoy and is publicly available at http://ontology.tno.nl/logico/.

tology proposed by W3C (http://www.w3.org/TR/owl-time), instead of specifying our time ontology from scratch.

- ***Location*** represents the geographical area or geographical point used to define the place of origin and destination for entities and activities. Location can have different levels of granularity. Location can be coarse-grained for scheduling, since in long term planning it is sufficient to specify approximately the place of origin and destination, such as, for example, the Netherlands or the port of Rotterdam.

- ***MoveableResources*** are characterized by the capability of moving on their own or being contained in another entity for the purpose of movement, and Static Resources are used to host and/or handle moveable resources. An implementation guide has been constructed for an open standard representing document data for road transport. The open standard has been developed by UN/CEFACT for electronic waybills, with a specialization to the eCMR for road transport. The eCMR assigns one specific document type, the CMR, to a generic representation of data that can be stored by all types of transport documents. Thus, the core structure should as well be applicable for documents shared in other modalities. To conduct the experiments, the eCMR ontology discussed in previous section is used.

7.4.3. EXPERIMENTAL RESULTS

The first experiment was the alignment of the two ontologies representing implementation guides of open standards, eCMR and SI. The alignment is not satisfactory. Only concepts representing common roles of organizations in the two ontologies can be aligned with each other, but not other concepts. In particular, the outcome of direct matching for SANOM was two correspondences containing the mapping of ***Consignee*** and ***Carrier*** that are identical in the two ontologies, AML detected one extra false positive by mapping ***ShippingInstruction*** to ***TransportInstruction***, LogMap mapped two identical object properties associated the two mapped classes, *isDeliveredTo* and *isCarriedBy*.

This mismatch of alignments is due to naming conventions that differ between the two ontologies in question. For instance, we used in the eCMR concept names derived from XML element names, like 'SupplyChainConsignment' and 'LogisticsPackage'. These concepts are not present in the other ontology. The concept 'SupplyChainConsignment' is also not expected to be part of a shipping instruction ontology, since the latter represents a consignment. In general, it is not common in supply and logistics standards to use a type of prefix 'SupplyChain-' for naming concepts, which makes alignment only possible to those open standards that use the same prefix for naming. The same applies to 'LogisticsPackage'.

Furthermore, shipping instruction has additional roles, due to delivery conditions. Besides a consignor (which is equal to a shipper) and consignee, notifies will also be mentioned. A notify has to be informed when containers arrive at a port of discharge. Besides these differences in the naming of concepts, which will require a common data dictionary like the United Nations Trade Data Elements Directory (UNTDED), this naming difference might be solved by annotating LogiCO with terms used by other ontologies.

Another difference is that these open standards represent the transport services of

Table 7.1: The annotations made in LogiCO by using SI and eCMR terminologies.

SI	LogiCO	eCMR
Shipper	Consignor	Consignor
ShippingInstruction	Activity	SupplyChainConsignment

enterprises like carriers. For instance, a shipping line is able to transport a container between the hinterland and a port and position a container for stuffing at the location of a shipper or only transport a container between two ports. The difference is known as carrier - and merchant haulage respectively and is, in fact, a combined service. This combined service is however not represented by an eCMR. However, the eCMR contains another modeling issue, namely that of modeling a shipment that can consist of more than one consignment. The shipment concept is used for data sharing between a customer and his carrier; the consignment concept for data sharing between a shipper and a forwarder.

A third difference is the representation of cargo. There are two different concepts used by these three ontologies, namely *LogisticsPackage* and *container*. One could argue that a container represents the actual cargo, but it also has packages stuffed inside. What is required besides agreement on the naming of concepts is the associations between those concepts, package and container.

We also use indirect matching to match eCMR and SI ontologies by the aid of LogiCO. Since LogiCO has also different naming from SI and eCMR, it is required to add some annotations to this ontology. In particular, we added the two extra annotations from the SI and eCMR ontology to LogiCO: we added *Shipper* from the SI ontology as synonym to *Consignor* and added *ShippingInstruction* from SI and *SupplyChainConsignment* from eCMR as synonyms to *Activity* in LogiCO. Table 7.1 shows the annotation of LogiCO concepts with those of SI and eCMR ontologies.

We experimented with three top alignment systems from the OAEI: SANOM, LogMap, and AML. These alignment systems had very promising outcome and were identical to each other. Figure 7.5 displays the mapping of eCMR concepts to those in SI that are obtained by indirect matching using the annotated LogiCO.

One of the reason of having such outcome is that the object properties in two created ontologies are similar to each other. The identity of the object properties, such as *IsDroppedOffAT*, *IsDeliveredTo*, *IsPickedUpFrom*, as well as the annotations added to LogiCO increase the quality of the alignment by mapping the classes in the domains and ranges of the corresponding object properties.

7.5. CONCLUSION

This chapter addressed the main question of the applicability of ontology alignment to interoperability in logistics by means of an experiment. The experiment was the alignment of ontologies representing implementation guides of two open standards, eCMR and shipping instruction (SI), with and without the use of an upper ontology, called LogiCO. The experiments of direct alignment did not give satisfactory results due to differences in naming convention and systems modeled by the ontologies. A second ex-

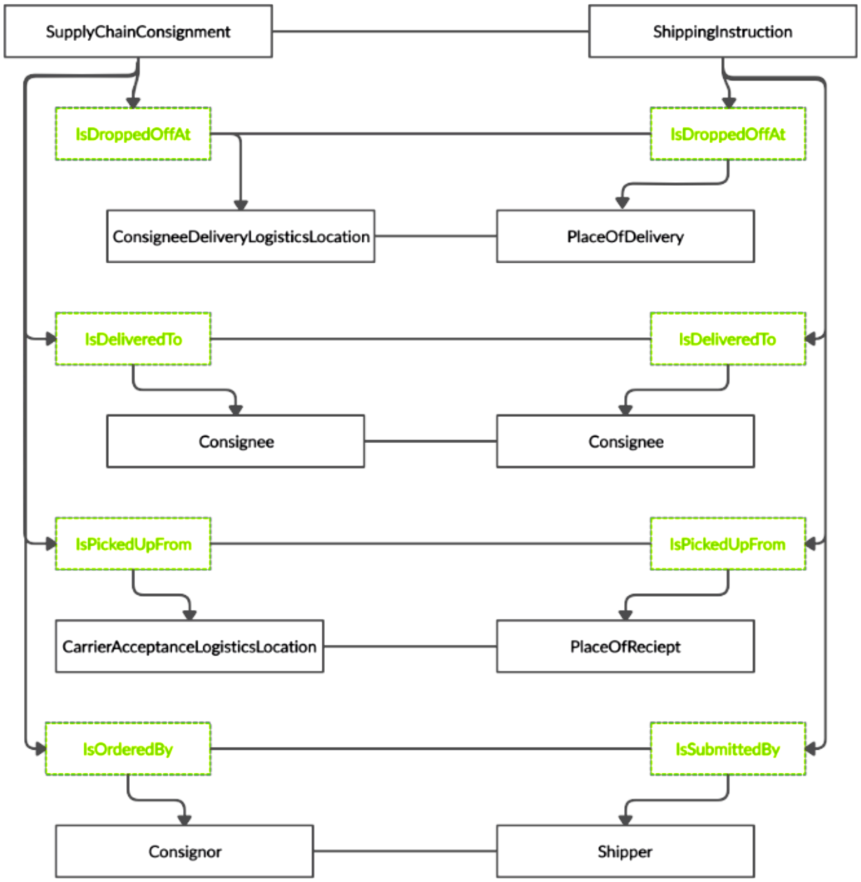


Figure 7.5: The alignment of eCMR to SI.

periment was performed by using indirect matching with the annotated LogiCO that resulted in a more acceptable alignment. The use of indirect matching is particularly useful in logistics since there are millions of enterprises. The indirect matching reduces the number of alignment from quadratic to linear with respect to the number of enterprises (or ontologies).

In view of the challenges encountered for alignment of implementation guides of open standards, it is safe to assume that alignment of (functional views of) database schemes represented as ontologies will even be more difficult. We cannot expect that ontologies derived from database schema use the same naming conventions and they will also have different structures, making the outcome of best ontology alignment systems, e.g., AML, SANOM, and LogMap, not acceptable and inappropriate for enabling interoperability in logistics. Note that these systems take advantages of several complex similarity metrics including string, linguistic, and structural, but were not able to detect enough correspondences. It is our expectation that ontology alignment will only im-

prove if there is a common understanding of what needs to be represented by an upper ontology. The use of the upper ontology can enhance the outcome of matching systems if the upper ontology is properly annotated. However, the upper ontology has to be extended with knowledge of business service composition to address all possible standards. It means that we need to have a shared background knowledge for both standards and alignment development. In the latter case, the alignment systems may also have to be extended. A more complete upper ontology reduces the experts efforts in annotating it and increases the performance of matching systems.

A thorough background knowledge can be created by using some existing ontologies in logistics like LogiCO and annotate it with proper names from different logistics models or standards. Such annotations require logistics expertise who is able to identify the related concepts in ontologies. An alternative approach is to create an integrated ontology by using ontology integration techniques based on the alignment of eCMR and SI, and then align new logistics models with the integrated ontology. Based on the alignment of the integrated ontology and the new logistics model, we can again use the ontology integration techniques to come up with a new integrated ontology. The alignment of each new model to the integrated ontology must be approved by a user so that the integrated ontology is reliable and does not contain redundant concepts. The creation of the integrated ontology is an iterative and augmentative approach that can finally result in a comprehensive upper ontology for logistics. While such an upper ontology can be exploited to enhance alignment outcome, it cannot be used for developing logistics IT systems since the integrated ontology gets more and more complex as the number of alignments and integration increases. The complexity of such an integrated ontology makes it difficult for even logistics experts to comprehend or annotate it. The alternative solution is to create a comprehensive upper ontology for logistics that contain all the concepts and names in different standards. Such an ontology can be used by companies as well as it can help improve the alignment outcome.

REFERENCES

[1] G. Satta, F. Parola, and S.-w. Lee, *The eu-27 logistics industry: structure and trends of major subsectors and national markets,* Kmi International Journal Of Maritime Affairs and Fisheries **3**, 1 (2011).

[2] W. Hofman, *Toward large-scale logistics interoperability based on an analysis of available open standards,* in *Enterprise Interoperability VIII* (Springer, 2019) pp. 249–261.

[3] W. Wang, A. Tolk, and W. Wang, *The levels of conceptual interoperability model: applying systems engineering principles to m&s,* arXiv preprint arXiv:0908.0191 (2009).

[4] A. Wieland and C. Marcus Wallenburg, *The influence of relational competencies on supply chain resilience: a relational view,* International Journal of Physical Distribution & Logistics Management **43**, 300 (2013).

[5] B. Behdani, Y. Fan, B. Wiegmans, and R. Zuidwijk, *Multimodal schedule design for synchromodal freight transport systems,* Behdani, B., Fan, Y., Wiegmans, B., & Zuidwijk , 424 (2014).

[6] W. Hofman and S. Dalmolen, *Data sharing in supply and logistics networks–development and implementation of extendable, standardized platform services for the physical internet in an open dynamic ecosystem of organizations,* in *International Physical Internet Conference. London* (2019).

[7] A. Schönberger, C. Wilms, and G. Wirtz, *A requirements analysis of business-to-business integration,* (2010).

[8] O. M. GROUP *et al.*, *Business process model and notation specification version 2.0,* (2011).

[9] T. Erl, *Service-oriented architecture: concepts, technology, and design* (Pearson Education India, 2005).

[10] E. Union, *new european interoperability framework promoting seamless services and data flows for european public administration,* Publication Office of the European Union, Luxembourg (2017).

[11] ISO, *Electronic data interchange for administration, commerce and transport (edifact)—application level syntax rules,* (1988).

[12] S. Lohmann, S. Negru, F. Haag, and T. Ertl, *Visualizing ontologies with vowl,* Semantic Web **7**, 399 (2016).

[13] L. Daniele and L. F. Pires, *An ontological approach to logistics,* Enterprise Interoperability, Research and Applications in the Service-oriented Ecosystem, IWEI'13 Proceedings , 199 (2013).

8

CONCLUSION

When ideas fail, words come in very handy.

Johann Wolfgang von Goethe

8.1. CONLUSION

This dissertation had the following research objective:

> *To address interoperability between heterogeneous IT systems in logistics by using ontology alignment.*

We put forward the following two research questions along with several sub-questions to accomplish the objective:

RQ1 What is the evolution and progress of ontology alignment and what enhancement needs to be made in view of the research objective?

- RQ1.1 Is ontology alignment problem solved in an efficient way?
- RQ1.2 Can we favor one ontology alignment system over another?
- RQ1.3 Can we give more meaning to an individual performance metric?
- RQ1.4 How to compare alignment systems with respect to multiple performance scores?

RQ2 To what extent does ontology alignment address interoperability between IT-systems in logistics in practice??

- RQ2.1 What is the state-of-the-art advances in enabling interoperability in logistics?
- RQ2.2 What would be the result of applying ontology alignment systems to logistics?

The answer to these research questions can help accomplish the research objective. For research question RQ1, we explored the ontology alignment literature by using both qualitative and quantitative approaches. We realized that the ontology alignment problem can be modeled, among others, as zero-one non-convex programming that is non-deterministic polynomial-time (NP) and is difficult to solve in a reasonable time. Although the use of indirect matching by an upper ontology decreases the number of alignments from quadratic to linear with respect to the number of ontologies, the execution time of matching a new logistics model to the upper ontology is particularly important for enabling interoperability between IT systems in logistics since there are often restricted time and resources for computations. As a promising alternative approach, the evolutionary algorithms (EAs) had already been applied to the ontology alignment problem. However, the population-based EAs have been used in the literature, which generally require a considerable time and massive memory for solving the problems. On top of that, population-based EAs often suffer from premature convergence, allowing to converge to only a local optimum of the given optimization problem. We also realized that EA-based ontology alignment systems have other issues, such as a lack of pre-processing the ontologies, which significantly affect their outcome. That was the motivation of RQ1.1 that was addressed in Chapter 3 by using simulated annealing for solving ontology alignment problem. Simulated annealing operates over one state that makes it more optimal in terms of memory consumption and execution time for computing the fitness. In addition, simulated annealing escapes the local optimum and converges to the global optimal solution. As a result, the ontology alignment problem

was first revised as the minimization of a fitness function that is optimized by using simulated annealing. We also developed a compound fitness function by using several similarity metrics. We showed that SANOM is superior to other EA-based systems in terms of execution time, precision, and recall, and is also competitive with the best systems in the ontology alignment evaluation initiative (OAEI). The content of Chapter 3 was in line with the objective, since we needed to have an efficient alignment system for enabling the interoperability in logistics.

Another lesson learned from the literature was that there was no methodology for comparison of alignment systems together. The only criterion was several performance scores or their averages, based on which the alignment systems were compared. Having a methodology for comparing alignment systems was the incentive of RQ1.2 that was studied in Chapter 4 by using frequentist statistics. More in detail, we first distinguished the comparison of alignment systems over one or multiple benchmarks and compare different statistics in the literature for comparing ontology alignment systems. The outcome of the comparison showed that we can favor one alignment system over another for a particular domain supported by statistical evidence. The content of this chapter conformed to the second objective, which involves the selection of an alignment system for logistics.

We also learned from the ontology alignment literature that the typical way for evaluating ontology alignment systems was to compute several performance scores that are basically a ratio directly related to true positives and true negatives. The statistical tests studied in Chapter 4 are also applied to the performance scores for comparison over multiple benchmarks. However, summarizing the accomplishment of an alignment system by a score is not informative since scores are only a ratio directly true positives or false positives. That was the stimulus of RQ1.3 that includes providing more meaning by representing distributions ratios. A solution based on Bayesian statistics was recommended in Chapter 5, where a distribution was estimated with respect to a performance metric, or better to say *alignment risk* that is a function of false positives and false negatives. Accordingly, we showed that the evaluation of alignment systems is a statistical inference problem and studied two primary schools of thought in statistics, frequentist and Bayesian, for estimation. We perceived that the estimation based on the frequentist approach is tantamount to the performance scores computed by ratios. As a more informative alternative, we developed a novel Bayesian model that was able to estimate a distribution with respect to the defined alignment risk. According to this distribution, a Bayesian test was also established that can compute the extent to which one alignment system is superior to another. The content of Chapter 5 was in line with the the objective, since we need to select the most appropriate alignment system for logistics. The difference of the Bayesian model with frequentist tests studied in Chapter 4 is that the former also concerns the evaluation of alignment systems, while the latter only involves the comparison. In addition, the comparison of the Bayesian model does not suffer from the drawbacks of decisions based on p-values, which makes it even more appropriate. However, if the alignments generated by different systems are not available, or the performance metric for comparison is not based on true positive or true negatives (e.g., execution time), then the frequentist tests such as Wilcoxon Singed-rank or Friedman tests and their Bayesian counterparts must be used.

Despite the effectiveness of the Bayesian model for evaluation and comparison of alignment systems, it can only take one performance score into account and make the comparison accordingly. However, a comprehensive comparison must include different performance metrics. The consideration of multiple performance metrics simultaneously was the incentive of RQ1.4 and studied in Chapter 6, where we proposed the use of multi-criteria decision-making (MCDM) methods for comparing and ranking alignment systems based on multiple performance metrics. Since each of the metrics has different importance in different ontology alignment applications or tasks, we particularly studied and elicited the preferences of ontology alignment experts for different OAEI tracks. The content of Chapter 6 showed that we can compare the alignment systems with respect to multiple performance metrics as well as experts' preferences, and find a ranking for alignment systems on different benchmarks.

For research question RQ2.1, we first looked into the literature of interoperability in logistics. Interestingly, there are many standards that are currently used in different logistics sectors for enabling interoperability, each with a specific data model, that create heterogeneity among different standards. On top of that, the standards and data models do not have semantics and are basically used as data carriers. Therefore, ontologies have to be extracted manually from existing open standards and models, where our study did not focus on analyzing existing tools to support this operation. As the first step, we developed two ontologies based on two well-known data models in logistics, shipping instruction (SI) and eCMR, where we named the concepts in ontologies identical to those in SI and eCMR. We expected that the direct alignment of such ontologies would bear useful outcome, but the drastic discrepancy among the names of concepts in these standards made the outcome of the direct alignment unsatisfactory. We therefore applied the indirect matching by using an upper ontology in logistics, *LogiCO*. We slightly annotated and used LogiCO as the upper ontology for matching SI to eCMR. The outcome of indirect matching could detect most of the shared entities in the ontologies. The study on the applicability of ontology alignment to logistics was presented in Chapter 7 that covered research question RQ2 and the first objective.

Overall, the use of direct matching, or even indirect matching with incomplete upper ontology, does not bear acceptable outcome and cannot address the heterogeneity in logistics. The annotation of ontologies must be done by a logistics expert who is sufficiently familiar with the domain. However, the efforts required for annotating is much less than that for aligning ontologies. In addition, as the number of logistics experiment as well as annotations increase, the need for an expert for annotation decreases. In addition, an expert must also inspect the alignment generated by a system to verify the veracity for data transformation since false positives and false negatives in the alignment can significantly influence the outcome of the transformation. The verification of alignment is not specific to logistics since it is an essential step for many other tasks such as ontology integration that use ontology alignment as a prerequisite. Furthermore, using indirect matching reduces considerably the number of alignments, which is particularly important in logistics that has millions of stakeholders in Europe only. Thus, in response to the research objective, we conclude that ontology alignment can address interoperability in logistics and reduce the human efforts provided that one of the following conditions holds:

- A proper upper ontology for logistics is developed for alignment in logistics interoperability.
- An existing upper ontology is used and manually annotated by an expert with the terminologies of the given ontologies.
- Many alignment experiments are conducted by different logistics standards and data models, and an ontology is annotated based on the outcome of alignment systems.

The answers to research questions RQ1.2, RQ1.3, and RQ1.4 allowed us to conclude that we can meaningfully compare different alignment systems and select the most appropriate one for a particular domain. The methodologies can be used in any domain with some standard benchmarks with known reference alignment. In addition, we realized that the direct alignment of ontologies does not produce an acceptable alignment, while the indirect matching with properly-annotated background knowledge can result in a satisfying alignment, regardless of the matching system being used. Another shortcoming of logistics is that the standards and data models are not developed as ontologies, making the number of logistics benchmarks restricted. However, when a number of ontologies representing the logistics standards are created, the methodologies for evaluating and comparing alignment systems can be used for selecting the most appropriate one.

8.2. A SUMMARY OF CONTRIBUTIONS

We now reiterate the contributions of this dissertation first in science and then in practice.

8.2.1. CONTRIBUTIONS TO SCIENCE

We first look into the contributions of this dissertation to ontology alignment. In this dissertation, we developed an efficient ontology alignment system by using simulated annealing as well as several methodologies based on statistics and MCDM for evaluating and comparing alignment systems. To organize the contributions to ontology alignment, we use the framework of the classification of ontology alignment publications that is plotted in Figure 8.1. Based on this figure, the ontology alignment contributions of this dissertation are categorized as (i) Review in Chapter 2; (ii) Matching technique and matching system in Chapter 3; (iii) Evaluation in Chapters 5 and 6; (iv) Comparison in Chapters 4-6; (v) Practical applications in Chapter 7.

In Chapter 2, we first put forward a revisited framework of ontology alignment publications and explain different classes in this framework. The new framework is more fine-grained that highlighted the lack of methodologies for comparison and evaluation of ontology alignment systems. In addition, a quantitative approach based on bibliometric analysis was used to provide a broad overview of the ontology alignment current status, progress, and possible ways for its future research. To the best of our knowledge, it was the first attempt of reviewing the ontology alignment field by using a quantitative approach, though there are several well-studied review papers with a qualitative approach. Based on the analyses discussed in Chapter 2, we realized that two segregated

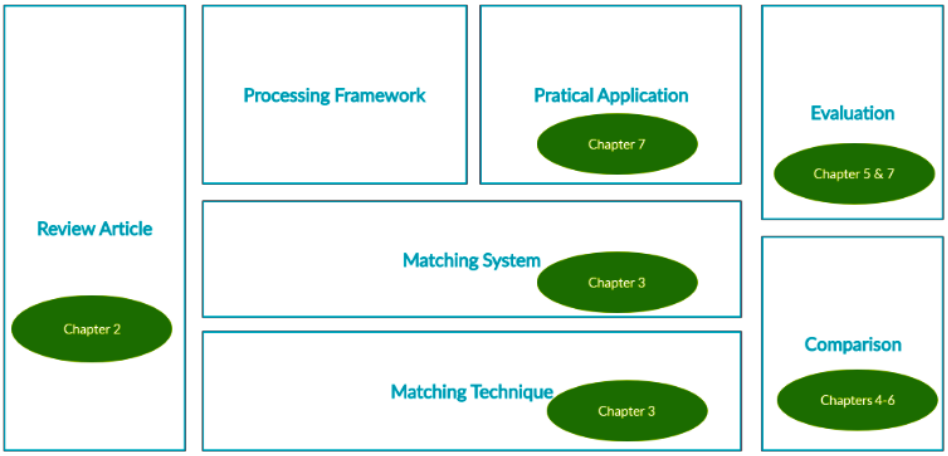


Figure 8.1: Ontology alignment article classification discussed in Chapter 2, as well as the relation of each chapter of this dissertation to a class of ontology alignment contributions.

communities, OAEI and Chinese, with the least collaborations form the ontology alignment domain. In addition, the academia-corporate collaborations need to be enforced, since the current status of such collaborations is trivial, i.e., less than 3%.

For the matching technique as referred to in Figure 8.1, we extended the Soft TF-IDF (term frequency-inverse document frequency) with two basic similarity metrics that is regarded as a new *similarity metric*, and we used simulated annealing as the *matching strategy*. The overall use of similarity metrics and matching strategy along with other pre- and post-processing formed the system, SANOM, as a matching system. It was investigated that SANOM is notably faster than other systems that employ evolutionary algorithms and is also efficient in terms of the amount of memory being used. SANOM has participated in the OAEI since 2017 and has been one of the top alignment systems in the conference track.

In Chapter 4-6, the evaluation and comparison of alignment systems were studied. We compared the frequentist statistical tests for comparing the alignment systems in Chapter 4 and recommended the appropriate test based on the number of benchmarks and the number of alignment systems. In Chapter 5, the evaluation of alignment systems was translated into a statistical inference problem, and a Bayesian model was developed that is able to compute a distribution with respect to a performance metric, or generally speaking, with respect to a function of false positives and false negatives. A Bayesian test was also developed that could compute the extent to which one alignment system is preferred over another. In Chapter 6, the comparison of alignment systems with respect to multiple performance scores and experts' preferences was studied and a proper research methodology for computing a ranking of alignment systems for different benchmarks was put forward and discussed in detail.

As a practical application, we studied the heterogeneity in logistics and the ways to deal with this problem to enable interoperability in this domain. As the analysis of literature suggested in Chapter 2, introducing new practical problems that can be addressed

by ontology alignment helps boost the field and can put forward new lines of research. In addition, it can strengthen and expand the corporate-academia collaborations as well. Hence, introducing such an application was of critical importance for ontology alignment.

Aside from contributions to ontology alignment, we developed two MCDM methods in Chapter 6. The first one was the Bayesian best-worst method (BWM) that used for computing the priorities based on the preferences of a group of experts (or decision-makers). The method is a new model that can address group decision-making in MCDM and is a promising alternative to the most common way for aggregating the preferences of multiple experts, i.e., computing the average of priorities of all experts. The Bayesian model applies to the preferences of all experts (instead of their priorities) that are expected to estimate the weights more accurately. In addition, it provides more information about the interrelation among different criteria by computing the extent to which one criterion is more important than another. Another contribution related to MCDM was the ensemble method that was able to combine the rankings of different MCDM outranking methods and compute a final rankings. In addition, we introduced two scores, consensus index and trust level, that can provide further information for the final rankings.

8.2.2. CONTRIBUTIONS TO PRACTICE

We verified the applicability of ontology alignment to the logistics domain. We showed that heterogeneity is prevalent in the logistics domain and the current use of standards does not address this problem properly. We put forward ontology alignment as a viable solution for resolving the heterogeneity and enabling interoperability. One big challenge was that most of the standards and data models in logistics do not have any semantical models. Therefore, we developed two ontologies based on the terminology and structure of the existing models in logistics. The alignment experiments on these ontologies showed that the direct matching bears unacceptable results since the terminology used in the logistics ontologies are sharply distinguished. We further used another experiment by the aid of an annotated upper ontology that resulted in an acceptable alignment. Therefore, the conclusion was drawn that ontology alignment can address the heterogeneity in logistics domain and reduce the human efforts to a minimum. However, this is basically achieved by using a proper upper ontology, otherwise, the results are not satisfactory. In addition, the use of indirect matching, aside from increasing the likelihood of matching with a proper upper ontology, can significantly decrease the number of alignment from quadratic to linear with respect to the number of ontologies. This is essential in logistics, where there are millions of enterprises in Europe only. The content of Chapter 7 indicated that ontology alignment can address the heterogeneity in logistics and enable interoperability in the domain provided that a proper logistics upper ontology will be used.

8.3. REFLECTION AND FUTURE RESEARCH

Besides the contributions that this dissertation puts forward, there are several critical challenges as well as multiple lines for future research. One of the crucial points dis-

cussed in Chapter 2 is that two research communities, OAEI and Chinese, form ontology alignment, each of which with tens of publications per year. More in detail, the collaborations of the Chinese community is not sufficient that reflected in the number of citations they received. As a result, finding common research agendas for Chinese and OAEI researchers can make the communities become closer and can significantly help the progress of ontology alignment to reach its full potential.

Finding new applications and domains to which ontology alignment can be applied is critical since it will help the its faster growth and progress. The logistics use case presented in this dissertation exemplifies such applications that existed for years but was not known to ontology alignment community, given that interoperability is one of the primary motivations of ontology alignment. Identification of new problems and applications, and consequently, creating use cases and ontologies for alignment makes them recognizable to both ontology alignment community as well as the community that the applications or problems lie.

The comparison of alignment systems is also distinguished from evaluation in this dissertation, although a proper method for evaluation, like the proposed Bayesian model in Chapter 5, could lead to a more meaningful comparison among alignment systems. In general, the evaluation and comparison of the alignment system is a statistical inference problem. One potential line for future research is to develop another specific Bayesian model for comparison only since the Bayesian test based on *risk* overestimates the difference between alignment systems if they are applied to the same benchmarks. In other words, the proposed Bayesian test is suitable for comparison of alignment systems when they are applied to different benchmarks. Another Bayesian test should be developed, potentially by using the idea of contingency table put forward in Chapter 4, that is suitable for comparison of alignment systems when they are applied to the same benchmarks.

For comparison based on multiple performance metrics, a methodology is proposed based on different MCDM metrics. While MCDM methods can be applied to performance scores to compare and rank alignment systems, they cannot simply accommodate the distributional data, like the distributions that are estimated by the proposed Bayesian model in Chapter 5. Another crucial drawback is that MCDM methods can be applied only to one benchmark at a time. Thus, finding a ranking of systems over multiple benchmarks is not a straightforward task. There are two venues that can further progress the evaluation and comparison of alignment systems. First, an MCDM method must be created that can accommodate distributional data. Thus, using MCDM methods can provide us with a distribution representing the overall performance of an alignment system that considers multiple performance metrics. In addition, an MCDM outranking method should be developed that, instead of using the performance scores of alignment systems, work with the probabilities that one alignment system is better than one another, where these probabilities are computed based on Bayesian models for comparison. In other words, the combination of Bayesian statistics and MCDM methods provides a comprehensive and meaningful comparison and evaluation of alignment systems.

SANOM can also be extended in several ways. One important step is to enable SANOM to be applied to large-scale ontologies, i.e., ontologies with tens of thousands of con-

cepts. Although SANOM is memory- and time-efficient, but handling large-scale ontologies is another matter that needs to be considered, especially in terms of implementation. In addition, SANOM can be equipped with the biomedical background knowledge that makes its outcome be comparable with state-of-the-art systems that typically use such background knowledge. Since the primary motivation of this dissertation was a use case in logistics, where there are many medium-sized ontologies, the capability of handling large-scale and biomedical ontologies was not a priority of the project, but are certainly important ways to further enhance the performance of SANOM. In addition, a mapping repair is seemingly necessary for SANOM, since the simulated annealing randomly generates and transitions to successors that potentially contain trivial false correspondences, while the overall fitness of the successor might be improved. This necessitates the existence of a mapping repair for SANOM.

We also investigated the applicability of ontology alignment to interoperability in logistics. This domain is distinct from other domains like biomedical, where we have some large-scale ontologies. In addition, the logistics models do not carry semantics and are not developed as ontologies. Thus, creating more logistics ontologies based on the available standards is an important contribution to this domain. Besides, we realized that the discrepancy in the terminology of logistics standards and data models is more severe, making the outcome of alignment systems unacceptable when ontologies are directly aligned. On the other hand, the use of an upper ontology can pave the way and enhance the outcome of alignment systems significantly. However, there is not a proper logistics dictionary or a comprehensive upper ontology. Therefore, creating a large upper ontology is a very crucial step in enabling interoperability in logistics by using ontology alignment. Such an upper ontology can be further enhanced by conducting many alignment experiments by logistics ontologies and annotate the upper ontology by the new terminologies encountered in new logistics standards or data models.

After discovering the alignment, a logistics expert needs to verify the correctness of each correspondence for transforming the data in ontologies. Correspondences have also a confidence level showing the extent to which they are reliable. When conducting logistics experiments, we can devise a learning method to learn from the confidence of each correspondence and the expert's opinions. Such a learning methodology can be used for *mapping repair* and makes the data transformation fully automated. Such a mapping repair can be applied to the outcome of any alignment system that is used for matching logistics ontologies.

LIST OF PUBLICATIONS

The followings list presents my publications during PhD, some of which formed the major parts of this dissertation.

JOURNAL ARTICLES

20. **M. Mohammadi**, *A Projection Neural Network for the Generalized Lasso*, IEEE Transactions on Neural Network and Learning Systems, 2019.

19. **M. Mohammadi**, *A Compact Neural Network for Fused Lasso Signal Approximator*, IEEE Transactions on Cybernetics, 2019.

18. **M. Mohammadi**, *Bayesian Evaluation and Comparison of Ontology Alignment Systems*, IEEE Access, 2019.

17. **M. Mohammadi**, *A New Discrete-time Neural Network for Quadratic Programming with General Linear Constraints*, Neurocomputing, 2019.

16. **M. Mohammadi** , H. Mousavi, S. Effati, *Generalized Variant Support Vector Machine*, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019.

15. **M. Mohammadi** , Wout Hofman, Yao-Hua Tan, *Simulated Annealing-based Ontology Matching*, ACM Transactions on Management Information Systems, Volume 10, Issue 1, May 2019.

14. **M. Mohammadi** , Wout Hofman, Yao-Hua Tan, *A Comparative Study of Ontology Matching Systems via Inferential Statistics*, IEEE Transactions on Knowledge and Data Engineering, Volume 31, Issue 4, April 1 2019.

13. **M. Mohammadi** , J. Rezaei, *Bayesian best-worst method: A probabilistic group decision making model*, Omega, 2019.

12. **M. Mohammadi** , Wout Hofman, Yao-Hua Tan, *Comparison of Ontology Alignment Systems Across Single Matching Task Via the McNemar's Test*, ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 12, Issue 4, July 2018.

11. **M. Mohammadi** , F. Farahi, *An Entropy-Regularized Framework for Detecting Copy Number Variants*, IEEE Transactions on Biomedical Engineering, Volume 66, Issue 3, March 2019.

10. **M. Mohammadi**, Wout Hofman, Yao-Hua Tan, *SANOM-HOBBIT: Simulated Annealing-based Ontology Matching on the Hobbit Platform*, Knowledge Engineering Review, 2019.

9. A. Ebrahimi Fard, **M. Mohammadi** , Y. Chen, B. van de Walle, *Computational Rumor Detection Without Non-Rumor: A One-Class Classification Approach*, IEEE Transactions on Computational Social Systems, 2019.

8. **M. Mohammadi** , A. Mansouri, *A Projection Neural Network for Identifying Copy Number Variants*, IEEE Journal of Biomedical and Health Informatics, Volume 23, Issue 5, Sept. 2019.

7. **M. Mohammadi** , Yao-Hua Tan, Wout Hofman, H. Mousavi, *A novel one-layer recurrent neural network for the l1-regularized least square problem*, Neurocomputing, Volume 315, 13 November 2018, Pages 135-144.

6. **M. Mohammadi** , H. Sharifi, Yao-Hua Tan, *Robust group fused lasso for multi-sample copy number variation detection under uncertainty*, IET System Biology, Volume 10, Issue 6, December 2016, p. 229 – 236.

5. **M. Mohammadi**, J. Rezaei, *Ensembling multi-criteria decision-making methods based on half-quadratic theory* , submitted to *Omega*, Second Revision.

4. **M. Mohammadi**, J. Rezaei, *Evaluating and Comparing Ontology Alignment Systems: An MCDM Approach*, submitted to *Journal of Web Semantics*, Second Revision.

3. **M. Mohammadi** , A. Ebrahimi Fard, *Ontology Alignment Revisited: A Bibliometric Narrative*, submitted to *Semantic Web*.

2. **M. Mohammadi**, *Solving l_1-regularized least square problem via a one-layer neural network*, submitted to *Applied Mathematics and Computation*.

1. A. Ebrahimi Fard, **M. Mohammadi**, Scott Cunningham, B. van de Walle, *Managing the Crisis within the Crisis: Tackling Rumours in Disasters*, submitted to *Plos One*.

CONFERENCE PAPERS

8. A. Ebrahimi Fard, **M. Mohammadi**, Scott Cunningham, B. van de Walle, *Rumour As an Anomaly: Rumour Detection with One-Class Classification*, *IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*.

7. **M. Mohammadi**, Wout Hofman, Yao-Hua Tan, *SANOM Results for OAEI 2019*, *International Semantic Web Conference, 2019*.

6. E. Jiménez-Ruiz, T. Saveta, O. Zamazal, S. Hertling, M. Röder, I. Fundulaki, A. Ngonga Ngomo, M. A. Sherif, A. Annane, Z. Bellahsene, S. Ben Yahia, G. Diallo, D. Faria, M. Kachroudi, A. Khiat, P. Lambrix, H. Li, M. Mackeprang, **M. Mohammadi**, M. Rybinski, B. S. Balasubramani and C. Trojahn, *Introducing the HOBBIT platform into the ontology alignment evaluation campaign*, *International Semantic Web Conference, 2018*.

5. **M. Mohammadi**, Wout Hofman, Yao-Hua Tan, *SANOM Results for OAEI 2018*, *International Semantic Web Conference, 2018*.

4. M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, E. Jimenez-Ruiz, K. Kolthoff, E. Kuss, P. Lambrix, H. Leopold, H. Li, C. Meilicke, **M. Mohammadi**, S. Montanelli, C. Pesquita, T. Saveta, P Shvaiko, A. Splendiani, H. Stuckenschmidt, E. Thieblin, K. Todorov, C. Trojahn, O. Zamazal=, *Results of the ontology alignment evaluation initiative 2017*, *International Semantic Web Conference, 2017*.

3. **M. Mohammadi**, Wout Hofman, Yao-Hua Tan, *SANOM Results for OAEI 2017*, *International Semantic Web Conference, 2017*.

2. **M. Mohammadi**, Wout Hofman, Yao-Hua Tan, *Ontology matching evaluation: A statistical perspective*, *International Semantic Web Conference, 2016*.

1. S. Abbassi, **M. Mohammadi**, E. Shams, *Robust crowdsourcing-based linear regression*, *6th International Conference on Computer and Knowledge Engineering (ICCKE), 2016*.