Evaluating and comparing ontology alignment systems

An MCDM approach

Mohammadi, Majid; Rezaei, Jafar

DOI
10.1016/j.websem.2020.100592

Publication date
2020

Document Version
Final published version

Published in
Journal of Web Semantics

Web Semantics: Science, Services and Agents on the World Wide Web 64 (2020) 100592

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Evaluating and comparing ontology alignment systems: An MCDM approach

Majid Mohammadi [a,b,*], Jafar Rezaei [a]

[a] Faculty of Technology, Policy, and Management, Delft University of Technology, The Netherlands
[b] The Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands

ARTICLE INFO

Article history:
Received 1 September 2019
Received in revised form 20 June 2020
Accepted 3 July 2020
Available online 12 July 2020

Keywords:
Ontology alignment
Ranking
Evaluation
MCDM
Bayesian BWM

ABSTRACT

Ontology alignment is vital in Semantic Web technologies with numerous applications in diverse disciplines. Due to diversity and abundance of ontology alignment systems, a proper evaluation can portray the evolution of ontology alignment and depicts the efficiency of a system for a particular domain. Evaluation can help system designers recognize the strength and shortcomings of their systems, and aid application developers to select a proper alignment system. This article presents a new evaluation and comparison methodology based on multiple performance metrics that accommodates experts' preferences via a multi-criteria decision-making (MCDM) method, i.e., Bayesian best–worst method (BWM). First, the importance of a performance metric for a specific task/application is determined according to experts' preferences. The alignment systems are then evaluated based on proposed expert-based collective performance (ECP) that takes into account multiple metrics as well as their calibrated importance. For comparison, the alignment systems are ranked based on a probabilistic scheme, where it includes the extent to which one alignment system is preferred over another. The proposed methodology is applied to six tracks from ontology alignment evaluation initiative (OAEI), where the importance of performance metrics are calibrated by designing a survey and eliciting the preferences of ontology alignment experts. Accordingly, the participating alignment systems in the OAEI 2018 are evaluated and ranked. While the proposed methodology is applied to six OAEI tracks to demonstrate its applicability, it can also be applied to any benchmark or application of ontology alignment.

1. Introduction

Ontology alignment is the process of finding similar entities in two different ontologies stating similar pieces of information in distinct ways. Since ontology alignment has extensive applications and can address many real-world problems, there is numerous research on this problem in the literature as well as a significant number of alignment systems. Ontology alignment evaluation initiative (OAEI) has been taken place for more than a decade whose objectives are to monitor the advancement of this field and compare systematically various alignment systems on several standard benchmarks with known reference alignment. Despite the tremendous progress for developing alignment systems for different challenges such as complex and large-scale ontologies [1,2], little efforts have been taken for developing a reliable means for evaluation and comparison of the systems [3].

There are different problems which have been solved by using ontology alignment. In semantic web service discovery, ontology alignment is used to help better discovery of services, or respond to a request by a composition of services [4,5]. In agent-based modeling, the communication between agents with different syntaxes is possible by using ontology alignment [6]. Or, ontology alignment is used to enable interoperability in logistics by aligning different logistics standards and transforming the associated instances [7]. Other applications of ontology alignment include, but not limited to, ontology integration [8,9], linked data [10], and peer-to-peer information sharing [11].

The existence of tools for evaluation and comparison is of the essence for several reasons. First, the evaluation of ontology alignment systems helps system designers to estimate the strengths and weaknesses of their systems that can be further used to enhance them. In addition, it guides the developers to select a proper alignment system for their matching tasks. It is particularly essential because there are numerous ontology alignment systems in the literature. Moreover, it is the primary aim of the OAEI to evaluate and compare the participating alignment systems on various matching tasks.

* Corresponding author at: The Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands.
E-mail address: majid.mohammadi690@gmail.com (M. Mohammadi).

https://doi.org/10.1016/j.websem.2020.100592

Multi-criteria decision-making (MCDM) is a sub-discipline of Operation Research that concerns with decision-making with respect to multiple conflicting criteria. A typical MCDM problem includes a set of alternatives and a set of criteria, and the aim is to rank, sort, or select the best alternative(s). This paper models the evaluation and comparison of ontology alignment systems as an MCDM problem, where the performance metrics and ontology alignment systems served as criteria and alternatives, respectively. Then, MCDM techniques can be used to evaluate the performance metrics as well as evaluating and comparing the alignment systems.

1.1. Related work

The typical way of ontology alignment evaluation is to use several performance metrics. Two widely-used performance metrics for ontology alignment are precision and recall. Given an alignment A and a reference A^*, precision is the ratio of true positives to the total correspondences in an alignment generated by a system and can be written as:

$$Pr(A, A^*) = \frac{|A \cap A^*|}{|A|}, \tag{1}$$

where Pr is precision and $|.|$ is the cardinality operator. Recall is another popular metric that is computed as the ratio of the true positives to the total number of correspondences in the reference. Thus, it can be calculated as:

$$Re(A, A^*) = \frac{|A \cap A^*|}{|A^*|}, \tag{2}$$

where Re is recall.

While precision and recall are arguably the most popular performance metrics in ontology alignment, they have several generalizations that address some of their shortcomings. One extension is confidence-based precision and recall that take into account the confidences of each identified correspondence by a system. The underlying idea is that if a system identifies an incorrect (correct) correspondence with a low confidence, it should be penalized (rewarded) by the confidence of the related correspondence. Accordingly, the weighted precision Pr_W of an alignment A is defined as:

$$Pr_W(A, R) = \frac{\sum_{c \in A} 1 - |\eta_A(c) - \eta_R(c)|}{|A|}, \tag{3}$$

and weighted recall is given by

$$Re_W(A, R) = \frac{\sum_{c \in A} 1 - |\eta_A(c) - \eta_R(c)|}{|R|}, \tag{4}$$

where $\eta_A(c)$ and $\eta_R(c)$ give the confidences of correspondence c in alignment A and reference R, respectively.

Semantic precision and recall are another extensions of precision and recall that consider the proximity of an alignment to the reference, instead of the strict size of their overlaps. The problem of precision and recall is that the unidentified correspondences are not considered, making them not differentiate between an alignment that is close to the reference and the one remote from it. Semantic precision and recall regard the proximity of the alignment with the reference, which is particularly a more proper indicator for the required efforts from a user to scrutinize the alignment. Given a proximity function ω, semantic precision and recall are defined as follows:

$$Pr_\omega(A, R) = \frac{\omega(A, R)}{|A|}, \quad Re_\omega(A, R) = \frac{\omega(A, R)}{|R|}, \tag{5}$$

where ω is a proximity function that must satisfy positiveness, maximality, and boundedness [12]. In some cases where the reference alignment is not available, mainly due to the sizes of the input ontologies, sample-based precision and recall are introduced as well [13].

While precision and recall (and their extensions) are both essential in many applications [14] as well as all OAEI tracks based on the literature, F-measure is another popular metric and is computed as the harmonic mean of precision and recall:

$$F(A, R) = \frac{2Pr(A, R)Re(A, R)}{Pr(A, R) + Re(A, R)}. \tag{6}$$

F-measure for other extensions of precision and recall is also computed accordingly.

Aside from these popular performance metrics, experts identified two important principles for a given alignment. The first is *conservativity* [15,16], which states that the alignment being identified by a system must not impose any new semantic relationship between the concepts of each of ontologies involved. The second is *consistency*, which states that the discovered correspondences should not lead to unsatisfiable classes in the merged ontology [16]. Furthermore, *Recall+* shows the portion of correspondences that a system cannot readily detect, and its higher values indicate that the associated system is able to identify the non-trivial correspondences. In addition, the execution time is another critical performance metric that must also be included.

Each of the performance metrics introduced above is important in some application domains and unessential in some others. Euzenat and Shvaiko [14] studied the importance of four performance metrics, i.e., precision, recall, speed, and an automation measure (i.e., a measure to show if the alignment system is automatic), for different applications of ontology alignment such as ontology evolution, Web service composition, and data integration. The importance of each performance metric in each domain is measured by three levels: *low*, *medium*, and *high*. Based on these levels, a weight is elicited by assigning a value to each level (i.e., low=1, medium=3, high=5) and then normalizing the values for each task. Then, the performance of an alignment A with respect to several performance metrics $M_i, i \in I$ is defined as the weighted harmonic mean of all the performance scores, e.g.,

$$Agg(A, R) = \frac{\sum w_i}{\sum \frac{w_i}{M_i(A,R)}}, \tag{7}$$

where Agg is the aggregated performance metric, and w_i is the normalized value of M_i. Notice that F-measure is not considered here, since Eq. (7) contains F-measure that is the combination of precision and recall. In other words, F-measure is the weighted harmonic mean of precision and recall, where the weights of precision and recall are set to *0.5*. This is due to the fact that there is no available information to favor precision over recall, or vice versa, so they are assumed equally important. In Eq. (7), on the other hand, the weights are computed, not only for precision and recall, but for other performance metrics based on the expert's preferences, therefore, F-measure is not correctly considered.

For comparing ontology alignment systems, the dominant strategy is to first select a performance metric and then compare the systems based on the averages or micro-averages of the same metrics over multiple benchmarks. Since averaging is not statistically safe and appropriate, several statistical methods have been recently put forward for evaluation and comparison of alignment systems [17–19]. In particular, the Bayesian model in [19] estimates the performance of an alignment system by a distribution instead of a ratio. For instance, the model outputs a precision distribution rather than a precision score. Such distributions provide more information about the overall performance of alignment systems that can be further used for comparison.

As a result, the Bayesian model is able to calculate the extent to which one alignment system is superior to one another based on the computed distributions. Although the statistical methods are a reliable means for evaluation and comparison, especially as opposed to averaging, they can only encompass one performance metric for comparison and evaluation. However, a more thorough evaluation or comparison should include multiple performance metrics, each represents an aspect of the alignment system accomplishment.

One way to taking two performance metrics into account is to use different curves such as precision–recall and receiver operating characteristic (ROC) [14]. These models, though efficient in several situations, have several pitfalls. First and foremost, they can only take two performance metrics into account. In addition, while they present a broad picture of the performance of systems, they cannot systematically compare or rank ontology alignment systems.

1.2. Contribution

All the performance metrics introduced so far are essential in ontology alignment, each of which has different importance in different ontology alignment applications/tasks. For instance, the execution time is much more important for the on-the-fly matching, while recall is much more preferred over recall for the semi-automatic matching, since removing false mappings is much simpler than detecting the missed ones. In this paper, we put forward the use of MCDM methods for evaluating and comparing alignment systems. To this end, we model the evaluation and comparison of alignment systems as an MCDM problem, where different performance metrics are served as criteria, different systems are considered to be the alternatives, and the decision-makers (DMs) are the ontology alignment experts (or users). Based on this formulation, we propose a methodology for evaluating and comparing different alignment systems with respect to multiple performance metrics. The methodology presented in this article can consider the importance of a performance metric for a particular ontology matching application/task, according to which it amalgamates multiple performance metrics.

In particular, we take six OAEI tracks as the case study, for each of which proper performance metrics are identified by inspecting the literature of the OAEI and asking the experts in the domain. Then, the preferences of domain experts over the identified performance metrics are elicited by designing a survey based on the Bayesian best–worst method (BWM) [20], which is a pairwise comparison-based MCDM method for aggregating the preferences of a group of experts or DMs. Accordingly, the importance of different performance metrics for six OAEI tracks is calibrated based on experts' preferences, as well as the extent to which one performance metric is more important than another.

Another contribution is the evaluation and comparison of alignment systems for each OAEI track based on multiple performance metrics and their identified importance. For evaluation, expert-based collective performance (ECP) is proposed that is the weighted mean of all scores, where weights are calculated based on experts' preferences. The ECP is used to compare two different alignment systems and rank them accordingly. Since Bayesian BWM is a stochastic model, ranking of alignment systems are also probabilistic, which means that we can find the extent to which one alignment system is superior to one another with respect to multiple performance metrics and experts' preferences. We visualize the comparison of performance metrics and alignment systems using a weighted directed graph.

In summary, the contributions of this article can be itemized as follows:

- A methodology for evaluation and ranking of ontology alignment systems is presented based on multiple performance metrics and experts' preferences.
- For evaluation of alternatives, we introduce ECP which takes into account multiple performance metrics and their importance.
- A probabilistic ranking of alignment systems is presented.
- As a case study, six OAEI tracks are used, for each of which the importance of different performance metrics is calibrated based on the preferences of multiple ontology alignment experts.

Some of the ontology alignment tasks are general and can be representative of some ontology alignment applications. We ask the ontology alignment experts to express their preferences over different performance metrics to calibrate the importance of metrics in general. However, the importance of metrics can be different in view of some specific applications. In that regard, the proposed methodology can be used, but it is required to elicit the preferences of experts again and the evaluation and comparison need to be made according to the new preferences. In addition, the proposed methodology can be used for any other ontology alignment task or application, where the goal is to evaluate, compare, and rank alignment systems with respect to multiple performance metrics as well as the preferences of single or multiple experts or users. Further, note that ECP is identical to any other composite score, such as F-measure, which are very common in situations where a set of systems are characterized by a set of performance metrics. Although the individual scores per performance metric could be insightful, a composite score could be useful to provide a holistic evaluation of a system and can be used to compare and rank the alignment with respect to multiple metrics and experts' or users' preferences. This would allow, among others, application developers to select an alignment system based on the need of the application they develop. The implementation of the proposed method in this article as well as the preferences of the ontology matching experts on six OAEI tracks are publicly.[1]

1.3. Organization

Section 2 contains the methodology used for this article, including the selection of performance scores for different OAEI track, Bayesian BWM for eliciting the preferences of ontology alignment experts, and the definition of ECP as well as an outranking method for evaluating and comparing alignment systems. We apply the overall methodology to the outcome of the systems participated at the OAEI 2018, and the results are presented in Section 3. The discussion and important lessons learned from using MCDM method for evaluation and comparison are presented in Section 4, and the paper is concluded in Section 5.

2. Ontology alignment evaluation and comparison based on MCDM

In this section, we discuss the steps required to use the MCDM methods for evaluating and comparing ontology alignment systems that are displayed in Fig. 1. In the following, the steps in the figure are explained in more details.

[1] https://github.com/Majeed7/OM_MCDM.

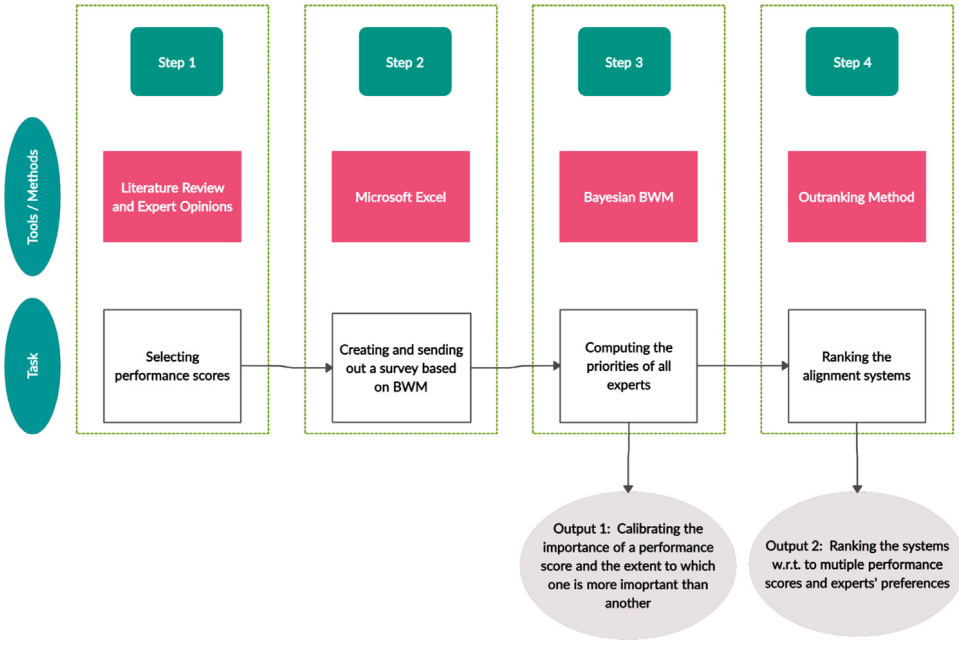


Fig. 1. The workflow of applying MCDM methods for comparing ontology alignment systems. The outputs of the methodology are twofold; (1) The importance of different performance metrics for a task/application are calibrated based on the experts' preferences, as well as the extent to which one performance metric is more important than one another in the given task/application; (2) The alignment systems are ranked based on experts' preferences and multiple performance metrics.

Table 1
The selected performance measures of five tracks of the OAEI.

OAEI track	Performance measures/indicators
Anatomy	Time, precision, recall, recall+, consistency
Conference	Precision, recall, conservativity, consistency
Multifarm	Time, precision, recall
LargeBioMed	Time, precision, recall
Disease and Phenotype	Time, precision, recall
SPIMBENCH	Time, precision, recall

Step 1: Selecting performance metrics

The first step is to specify the performance metrics for each OAEI track. To that end, we inspected the ontology alignment literature and OAEI website to accumulate the appropriate performance metrics for each track. Accordingly, we created a list of metrics for each track, and then ask the ontology alignment experts, who were mainly the OAEI organizers, for their suitability. After the solicitation, the list of performance metrics got completed, which is tabulated in Table 1 for six OAEI tracks.

Step 2: Creating and sending out a survey

After the determination of 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 the best–worst method (BWM) [20] so that experts can specify their preferences for different OAEI tracks. The devised survey contained instruction and an example describing the way the experts can correctly evaluate different performance metrics. Experts were asked to fill out only the survey of the tracks with which they are familiar. Overall, 13 experts participated in this study, each of whom expressed their preferences over at least one of the six tracks.

Step 3: Computing the priorities of all experts

Since the survey was created based on the BWM, we use it to calibrate the priorities of different performance metrics for each expert as well as the aggregated priorities. In this regard, we use the Bayesian BWM [20], which is able to take into account the preferences of multiple experts (or DMs) and provide a final aggregated priorities reflecting the group opinions. The priorities for each expert is a Dirichlet distribution computed by the Bayesian BWM. In addition, we can calibrate the extent to which a group of experts prefers one performance metric to one another. As a result, the first outcome of this study is the importance or priorities of different performance metrics for six OAEI tracks based on experts' preferences.

Assume that K experts evaluate n performance metrics $C = \{c_1, \ldots, c_n\}$ for an OAEI track. In order to apply the Bayesian BWM, the following steps must be taken for each OAEI track:

Step 1: Expert k first selects the best (c_B^k) and the worst (c_W^k) performance metrics from C.

In this step, each expert needs to select only the best and the worst performance metrics from the ones that previously identified for the corresponding track. The expert does not make any pairwise comparison between performance metrics at this stage. In addition, the best performance metric is the most important, while the worst is the least important metric only for the associated track.

Step 2: Expert k makes the pairwise comparison between the best (c_B^k) and the other performance metrics.

In this step, each expert expresses 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 of expert k in this stage leads to the "Best-to-Others" vector A_B^k as

$$A_B^k = \left(a_{B1}^k, a_{B2}^k, \ldots, a_{Bn}^k\right), \quad k = 1, 2, \ldots, K, \tag{8}$$

where a^k_{Bj} represents the preference of the best performance metric (c^k_B) over metric $c_j \in C$ for expert k.

Step 3: Expert k makes the pairwise comparison between the worst (c^k_W) performance metrics and the other metrics from C.

In this step, each expert needs to express his/her preferences of the other performance metrics over the worst metric by a number between one and nine. The outcome of this step for expert k is the "Others-to-Worst" vector A^k_W as

$$A^k_W = \left(a^k_{1W}, a^k_{2W}, \ldots, a^k_{nW}\right)^T \tag{9}$$

where a^k_{jW} represents the preference of metric $c_j \in C$ over the worst metric for expert k (c^k_W).

Step 4: Obtaining the aggregated weights $w^* = (w^*_1, w^*_2, \ldots, w^*_n)$ and the weight for each expert $w^k, k = 1, \ldots, K$ based on the following probabilistic model:

$$\begin{aligned}
&A^k_B \big| w^k \sim multinomial(1/w^k), \quad \forall k = 1, \ldots, K,\\
&A^k_W \big| w^k \sim multinomial(w^k), \quad \forall k = 1, \ldots, K,\\
&w^k \big| w^* \sim Dir(\gamma \times w^*), \quad \forall k = 1, \ldots, K,\\
&\gamma \sim gamma(0.1, 0.1),\\
&w^* \sim Dir(1),
\end{aligned} \tag{10}$$

where *multinomial* is the multinomial distribution, *Dir* is the Dirichlet distribution and *gamma*(0.1, 0.1) is the gamma distribution with shape parameters of 0.1. Since this model does not have a closed-form solution, Markov-chain Monte Carlo (MCMC) [21] methods like JAGS [22] must be used. As a result of sampling, S samples from the posterior distribution of w^* are then available and can be used for studying the preferences of experts. First, the credal ranking of performance metrics is derived based on these samples.

Definition 2.1 (*Credal Ordering [20]*)**.** For a pair of performance metrics c_i and c_j, a credal ordering O is defined as

$$O = (c_i, c_j, R, d) \tag{11}$$

where

- R is the relation between the performance metrics c_i and c_j, i.e., $<$, $>$, or $=$;
- $d \in [0, 1]$ is the confidences of the relation.

Definition 2.2 (*Credal Ranking [20]*)**.** For a set of performance metrics $C = (c_1, c_2, \ldots, c_n)$, the credal ranking is a set of credal orderings which includes all pairs (c_i, c_j), for all $c_i, c_j \in C$.

Credal ranking provides the extent to which one performance metric is more important than one another, which are obtained after computing w^* for each OAEI track. The degree in credal ordering for each pair of performance metrics c_i and c_j is computed as:

$$P(c_i > c_j) = \frac{1}{S}\sum_{s=1}^{S} I(w^{*s}_i > w^{*s}_j). \tag{12}$$

Step 4: alignment systems evaluation and comparison

After summarizing the priorities of all experts in w^*, we can evaluate and rank the alignment systems with respect to multiple performance metrics. For evaluation, we can aggregate different performance metrics into one based on experts' preferences. To this end, expert-based collective performance (ECP) for alignment A_i is defined as

$$\begin{aligned}
ECP(A_i) &= E_{w^*}\left(P^T_i w^*\right)\\
&= P^T_i E_{w^*}\left(w^*\right)
\end{aligned} \tag{13}$$

where $E_{w^*}()$ is the mathematical expectation with respect to w^*, and $P_i \in R^n$ is the vector of all performance scores for alignment system A_i. Since the expectation with respect to w^* can be estimated by the MCMC samples, ECP can be approximated as

$$ECP(A_i) = P^T_i \hat{w}^* = \sum_{j=1}^{n} P_{ij}\bar{w}^*_j, \qquad \text{where } \bar{w}^* = \frac{1}{S}\sum_{s=1}^{S} w^{s*}. \tag{14}$$

In fact, Eq. (14) is the weighted mean of all performance scores, where the weights are calculated based on experts' preferences. Similarly, one can compute the harmonic mean, similar to Eq. (7). Since the harmonic mean of zero with any number is zero and we have zero scores for some performance metrics in some domains (e.g., consistency in conference track), we use Eq. (13) for the evaluation. As a result, the overall performance of an alignment system with respect to multiple performance metrics can be summarized into one score given the experts' preferences. The use of such a score is to basically compare two alignment systems based on multiple metrics as well as users' preferences, so that it can be used to rank the alignment systems. Another important point here is the type of performance metrics. Some of the metrics, like time, are cost, a lower value of which is desired, while a higher value for the benefit metrics, like precision and recall, is desired. The types of criteria are also considered in computing equation (14).

The alignment systems can also be ranked based on experts' preferences. Since w^* is a distribution, the ranking will be probabilistic as well that shows to what degree one alignment system is preferred over another. For two alignment systems A_i and A_j and the aggregated priorities w^*, the probability that A_i being superior to A_j is computed as

$$P(A_i > A_j) = \int I\left(P^T_i w^* > P^T_j w^*\right) P\left(w^*\right) dw^*. \tag{15}$$

Since we have S MCMC samples of w^*, then equation (15) is estimated as

$$P(A_i > A_j) \approx \frac{1}{S}\sum_{s=1}^{S} I\left(P^T_i w^{s*} > P^T_j w^{s*}\right) \tag{16}$$

where w^{s*} is sample s of w^*. Eq. (16) needs to be computed for each pair of systems. To distinguish the credal ranking of alignment systems with the ranking of alignment systems, we call the latter *outranking*.

3. Experiments

In this section, we evaluate and compare the alignment systems with respect to multiple performance metrics based on experts' preferences. The alignments produced by various systems are available on the OAEI website. For each of the tracks, we first evaluate the performance scores by plotting a graph for their corresponding credal ranking, and we then evaluate and compare the alignment systems using ECP and credal outranking.

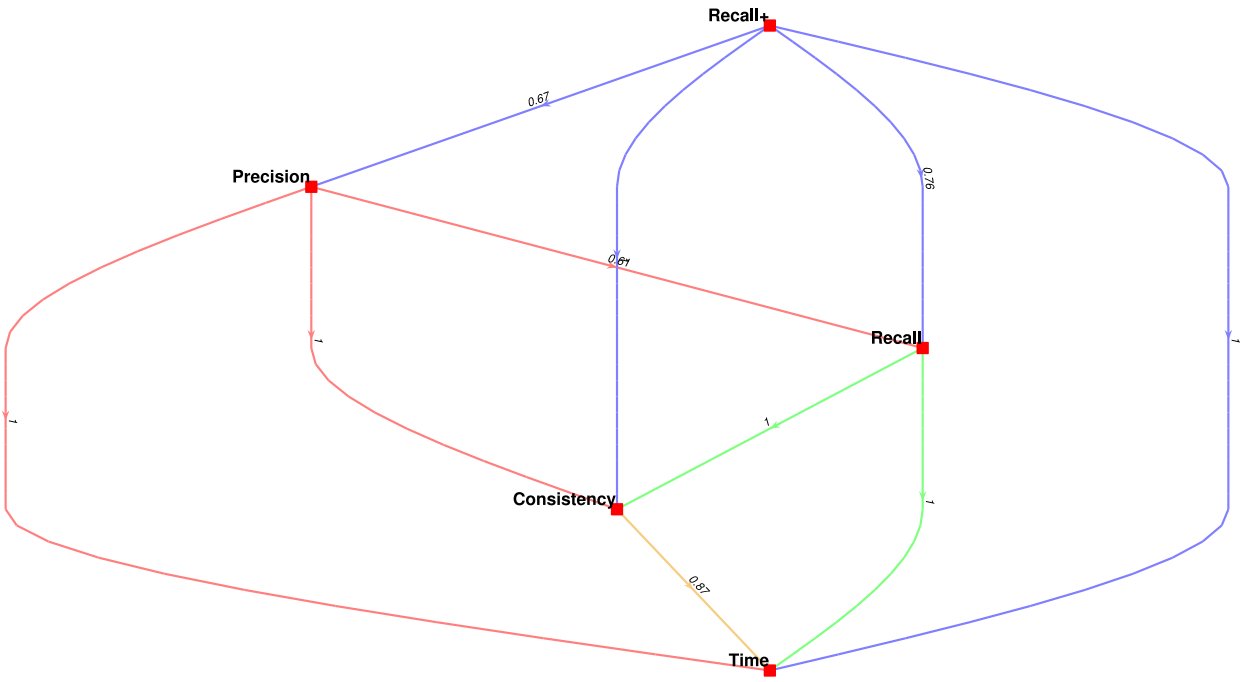


Fig. 2. The credal ranking of performance metrics for the anatomy track.

Table 2
Evaluation of alignment systems on the OAEI anatomy track.

	Time (s)	Precision	Recall	Recall+	Consist.	ECP	F-measure
LogMapBio	808	0.89	0.91	0.76	1	0.85	0.90
DOME	22	1.00	0.62	0.01	0	0.52	0.76
POMAP++	210	0.92	0.88	0.70	0	0.78	0.90
Holontology	265	0.98	0.29	0.01	0	0.40	0.45
ALIN	271	1.00	0.61	0.00	1	0.60	0.76
AML	42	0.95	0.94	0.83	1	0.98	0.94
XMap	37	0.93	0.87	0.65	1	0.90	0.90
LogMap	23	0.92	0.85	0.59	1	0.87	0.88
ALOD2Vec	75	1.00	0.65	0.09	0	0.54	0.79
FCAMapX	118	0.94	0.79	0.46	0	0.69	0.86
KEPLER	244	0.96	0.74	0.32	0	0.62	0.84
LogMapLite	18	0.96	0.73	0.29	0	0.63	0.83
SANOM	48.7	0.89	0.84	0.63	0	0.76	0.87
Lily	278	0.87	0.80	0.52	0	0.68	0.83
Weight	0.0913	0.261	0.248	0.283	0.115		

3.1. Anatomy track

The anatomy track involves the alignment of the adult mouse anatomy to a part of NCI thesaurus describing the human anatomy. In the OAEI 2018, 14 systems participated in this track that are compared based on the execution time, precision, recall, consistency, and recall+. 10 out of 13 experts filled the associated survey to this track and evaluated the corresponding performance metrics. One expert identified precision, recall, and recall+ as the most essential metrics by assigning one to the pairwise comparison associated with these metrics, one expert selected solely precision and another picked recall 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 $M \overset{v}{\Rightarrow} M'$ indicates that performance metric M is more important than M' with confidence v. Fig. 2 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 aggregate preferences of all experts.

We also evaluate the alignment systems with respect to multiple performance metrics and experts' preferences. Table 2 shows the performance scores of different alignment systems along with their ECP and F-measure, as well as the mean of the weight distribution in the last row. We particularly compare ECP and F-measure, since the latter is typically used for comparing alignment systems. Based on both ECP and F-measure, AML is the top system in this track with an ECP of 0.98 and a F-measure of 0.94. XMap is the second system based on ECP, while it shares it with LogMapBio and POMAP++ in terms of F-measure. The execution time of XMap is far lower than that of LogMap, and that is basically the reason that XMap is superior to LogMapBio in terms of ECP. In addition, recall and recall+ of XMap is significantly better than that of POMAP++ that makes it a better system in terms of ECP, though their F-measure is equivalent.

Table 3
Evaluation of alignment systems participated in the 2018 OAEI conference track.

	Precision	Recall	AvgConserViol	AvgConsisViol	ECP	F-measure
SANOM	0.78	0.76	5.15	4.60	0.81	0.77
AML	0.83	0.70	1.86	0.00	0.92	0.76
LogMap	0.84	0.64	1.19	0.00	0.91	0.73
XMap	0.81	0.61	2.65	0.70	0.84	0.70
KEPLER	0.76	0.61	5.86	7.57	0.68	0.68
ALIN	0.88	0.54	0.10	0.00	0.90	0.67
DOME	0.88	0.54	5.05	0.48	0.80	0.67
Holontology	0.86	0.55	3.14	0.48	0.83	0.67
FCAMapX	0.71	0.61	5.90	13.00	0.59	0.66
LogMapLite	0.84	0.54	4.57	1.19	0.78	0.66
ALOD2Vec	0.85	0.54	5.90	1.29	0.76	0.66
Lily	0.59	0.63	7.00	6.20	0.62	0.61
Weight	0.35	0.35	0.13	0.17		

Another crucial difference is the position of LogMap: It is the third system in terms of ECP and fifth in terms of F-measure. The reason for its position regarding ECP is its speed compared to LogMapBio and its consistency compared to POMAP++. Lily and KEPLER have also different ranks with respect to ECP and F-measure: Lily is the eighth in terms of ECP and 10th in terms of F-measure, while KEPLER is ranked as the 10th regarding ECP and the eighth concerning F-measure.

In addition, we compute the credal outranking of alignment systems and visualize the outcome in a weighted, directed graph. Fig. 3 plots the outranking of alignment systems in the anatomy track. In this figure, the nodes represent the alignment systems and each edge $A_1 \overset{v}{\Rightarrow} A_2$ says that system A_1 is superior to A_2 with the confidence v. The overall ranking of alignment systems are similar to that of ECP, but the outranking provides the extent to which one system is superior to one another. For instance, KEPLER is superior to Alin with 0.80 confidence based on multiple performance metrics.

3.2. Conference track

This track consists of matching seven ontologies from different conferences and includes 21 matching tasks with reference alignment. The performance metrics considered for this 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 experts precision, and three experts consistency. Furthermore, one expert 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 metric. Fig. 4 displays the credal ranking of performance metrics for this track. According to this figure, precision and recall are the most important performance metrics that are significantly more important than consistency and conservativity.

Table 3 tabulates the result of the analysis on 12 systems participated in this track at the OAEI 2018 along with their ECP and F-measure, as well as the mean of the weight distribution at the last row. Regarding F-measure, SANOM is the top system, while it is ranked as the sixth system based on ECP. This is basically due to the conservativity and consistency violation of SANOM, while the systems with better ranks in terms of ECP, e.g., AML, LogMap, and ALIN, have lower violations. Another significant difference is the rank of ALIN that is the third in terms of ECP and sixth (jointly with DOME and Holontology) regarding recall. In addition, KEPLER, that is the fifth system regarding F-measure, becomes the tenth with respect to ECP, since its conservativity and consistency violation is significantly high. In addition to ECP, Fig. 5 plots the outranking of alignment systems participated in the OAEI 2018 conference track, that provides to what degree an alignment system is superior to another.

Table 4
Evaluation of the alignment systems on the OAEI Multifarm track.

	Time	Precision	Recall	ECP	F-measure
AML	26	0.72	0.35	1.00	0.47
KEPLER	900	0.4	0.21	0.50	0.28
LogMap	39	0.72	0.25	0.87	0.37
XMap	22	0.02	0.07	0.23	0.03
	0.138	0.426	0.436		

Table 5
Ranking of systems participated in the 2018 OAEI disease and phenotype track. The task involves the alignment of HP and MP.

	Time	Precision	Recall	ECP	F-measure
LogMap	31	0.875	0.835	0.94	0.85
LogMapBio	821	0.862	0.841	0.92	0.85
AML	70	0.889	0.801	0.93	0.84
LogMapLite	7	0.993	0.609	0.87	0.75
POMAP++	1668	0.855	0.575	0.75	0.69
Lily	4749	0.682	0.647	0.64	0.66
XMap	20	0.994	0.314	0.71	0.48
DOME	46	0.997	0.308	0.71	0.47
Weight	0.12	0.42	0.46		

3.3. Multifarm track

This track involves the alignment of ontologies in different languages. The ontologies of this track are the ones in the conference track that are translated into eight different languages. For this track, execution time, precision, as well as recall are considered for evaluation and comparison. Out of 13 experts, 10 filled the survey regarding the performance metrics of this track, five of whom selected both precision and recall as the most important performance score for this track. Three experts opted for only metric and the remaining two selected precision alone as the most important metrics. On the other hand, all experts identified the execution time as the least important metric for this track. Fig. 6 plots the credal ranking of performance metrics of the Multifarm track. According to this figure, recall is slightly more important than precision based on all experts' preferences, and they are both significantly more important than time.

Table 4 tabulates the results of alignment systems participated in the OAEI 2018 Multifarm track, along with their precision, recall, and the average weight distribution of performance metrics at the last row. According to this table, AML is the best system, followed by LogMap, KEPLER, and XMap in terms of both ECP and F-measure. In addition, Fig. 7 plots the credal outranking of alignment systems participated in the Multifarm track.

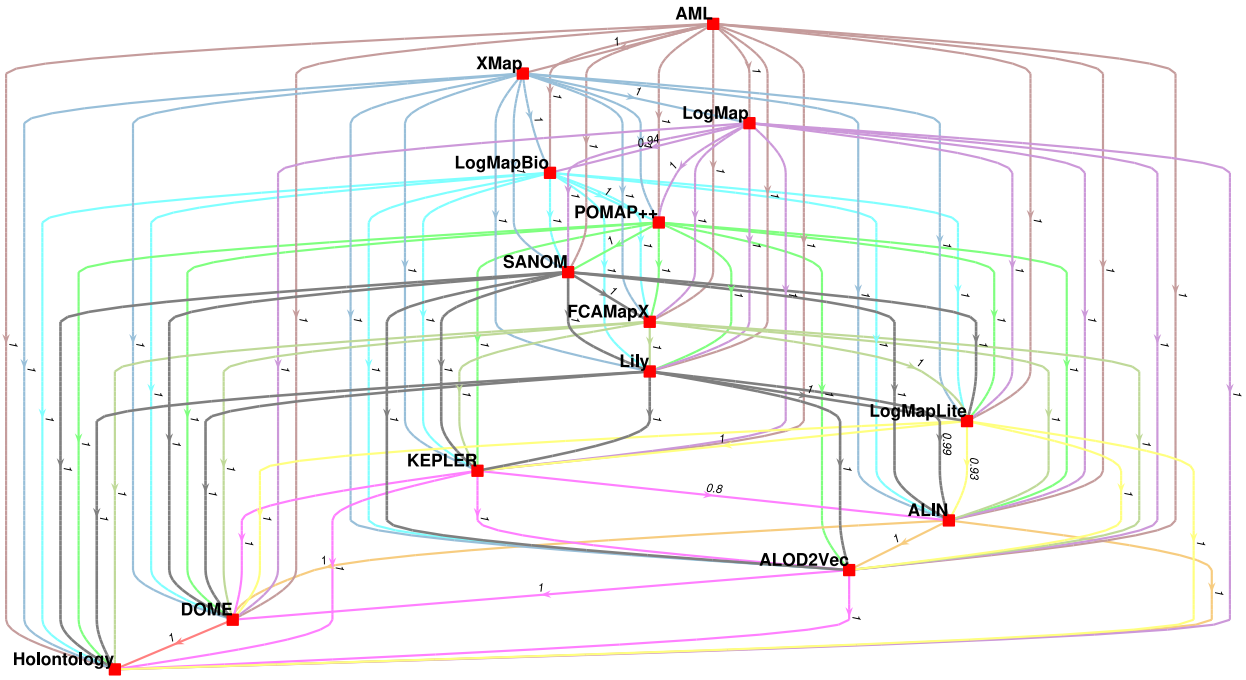


Fig. 3. The outranking of alignment systems in the OAEI 2018 anatomy track.

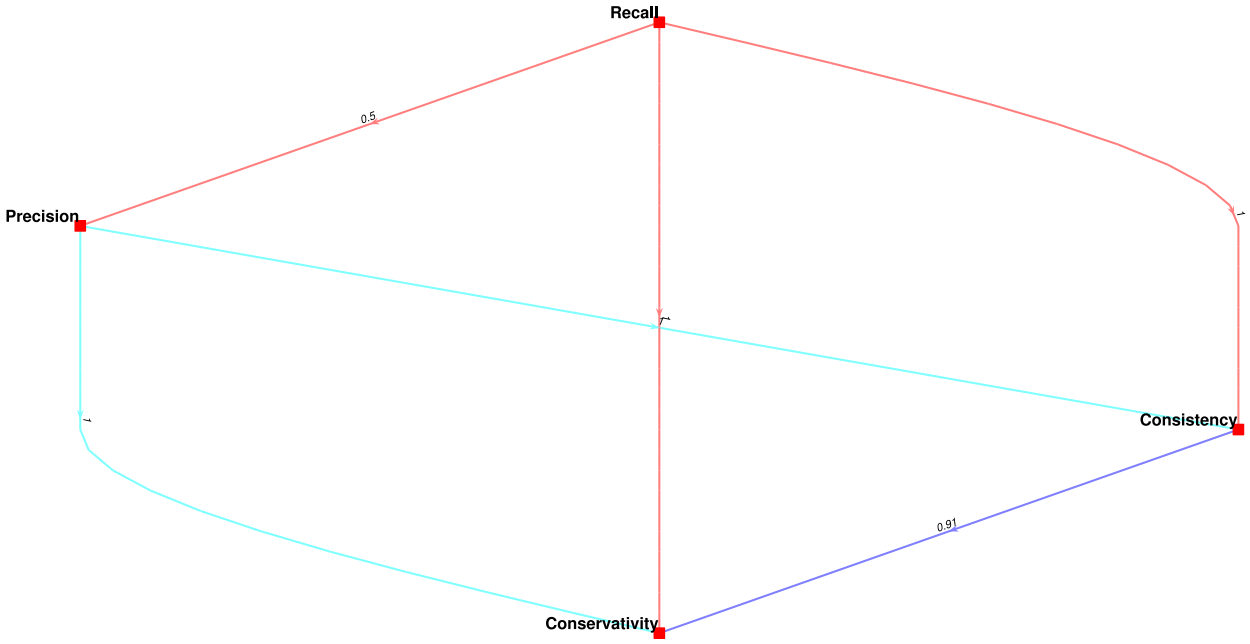


Fig. 4. The credal ranking of performance metrics for the conference track.

Disease and phenotype track

The OAEI disease and phenotype track comprises matching different disease and phenotype ontologies. The OAEI 2018 consisted of two tasks: The first one was to align the human phenotype (HP) ontology to the mammalian phenotype (MP), the second to align the human disease ontology (DOID) and the orphanet and rare diseases ontology (ORDO). The performance metrics used 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 metrics. In addition, all experts identified the execution time as the least important metric for this track. Fig. 8 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.

In the interest of avoiding duplication, we only consider the alignment of HP to MP, in which eight systems participated in the OAEI 2018. Table 5 illustrates the results of the systems participated in the OAEI 2018 disease and phenotype track for

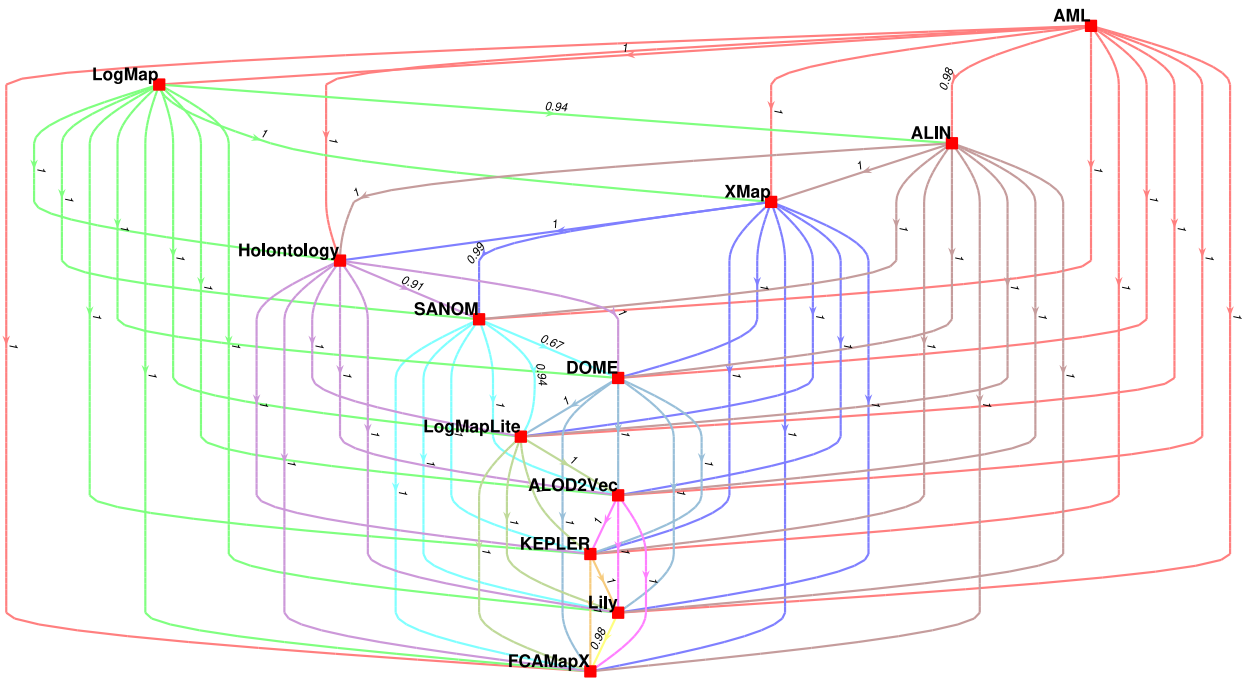


Fig. 5. The outranking of alignment systems in the OAEI 2018 conference track.

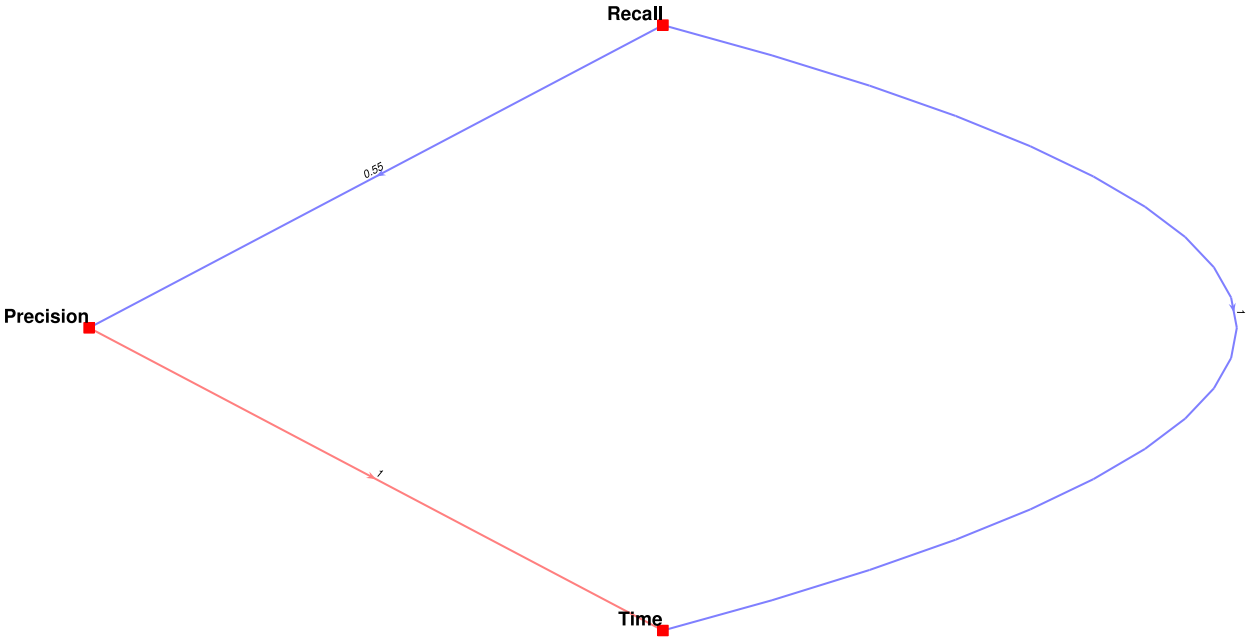


Fig. 6. The credal ranking of performance metrics for the OAEI Multifarm track.

mapping HP and MP ontologies as well as their associated ECP and F-measure. According to this table, LogMap is the best system in terms of ECP and is best jointly with LogMapBio in terms of F-measure. AML is the second-best system in terms of ECP, but it is the third regarding F-measure. Similarly, LogMapBio is the first system concerning F-measure, but it becomes the third with respect to ECP. That main reason for the change of positions is that AML is significantly faster than LogMapBio that compensates the better recall of LogMapBio. Another significant difference is the rank of Lily that is the eighth system in terms of ECP due to its execution time, but the sixth regarding F-measure. Fig. 9 plots the outranking of alignment systems on mapping HP to MP in the OAEI disease and phenotype track.

3.4. Large biomedical track

The track involves finding alignments between the Foundational Model of Anatomy (FMA), SNOMED CT, and the National Cancer Institute Thesaurus (NCI) ontologies. The ontologies are large and include tens of thousands of classes. The performance metrics considered in this track are the execution time, precision, recall, and consistency. Out of 13 experts, 10 have filled

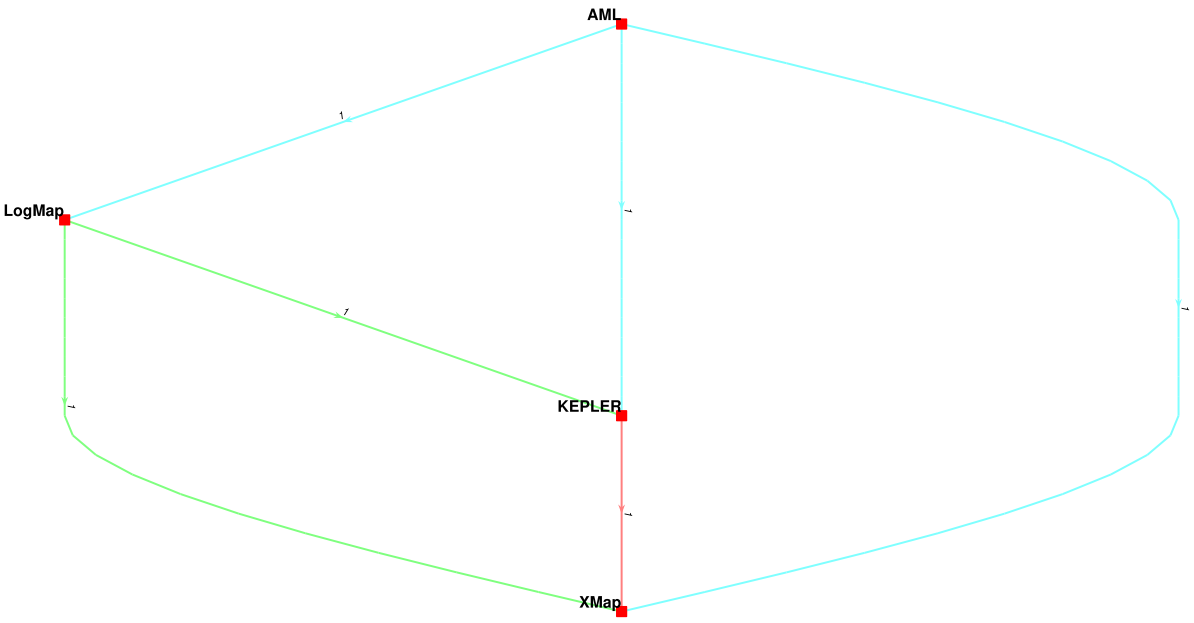


Fig. 7. The credal outranking of alignment systems for the OAEI Multifarm track.

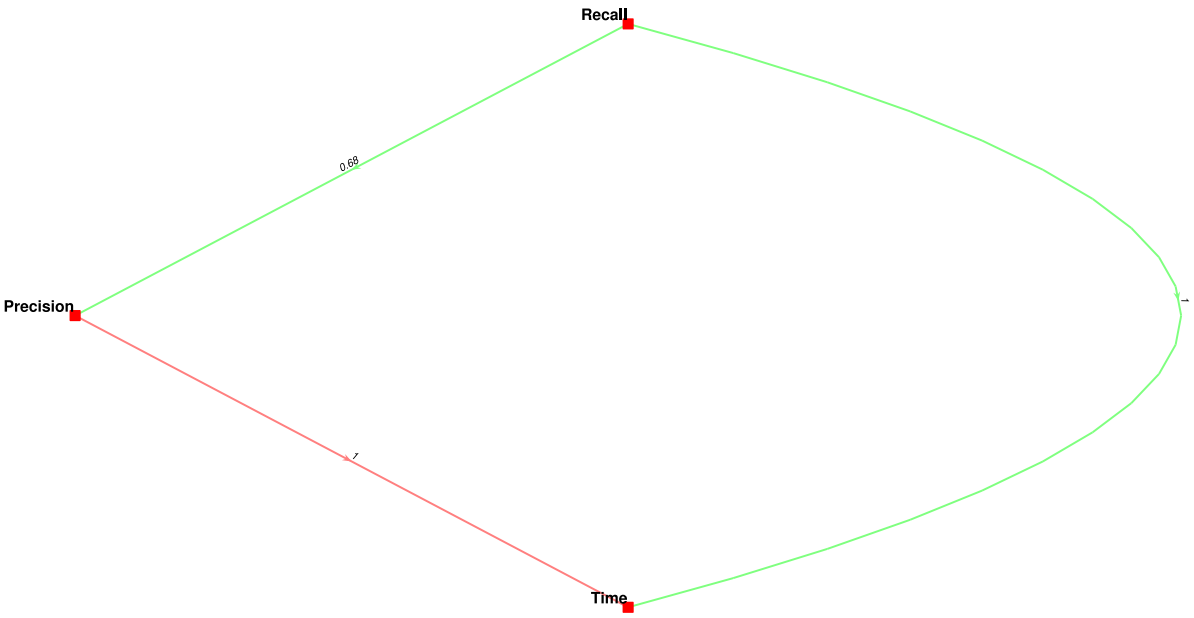


Fig. 8. The credal ranking of performance metrics for the disease and phenotype track.

Table 6
Evaluation of the alignment systems taking part in the large biomedical track for mapping FMA to NCI.

	Time (s)	Precision	Recall	Consistency	ECP	F-measure
AML	55	0.838	0.872	0.007	0.85	0.85
LogMap	51	0.856	0.808	0.006	0.83	0.83
LogMapBio	1072	0.83	0.831	0.006	0.70	0.83
XMap	65	0.878	0.742	0.006	0.81	0.80
FCAMapX	881	0.665	0.841	58.1	0.78	0.74
LogMapLite	6	0.676	0.819	18	0.81	0.74
DOME	12	0.803	0.668	2.5	0.76	0.73
Weights	0.140	0.356	0.378	0.125		

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 merely recall, and two consistency alone. In addition, seven of the experts picked consistency as the least important performance metric and the remaining three opted for time. Fig. 10 displays the credal ranking of four performance scores 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 also significantly more important than time, that itself is more important than consistency.

Table 6 shows the results of the alignment systems on mapping FMA to NCI, as well as their ECP, F-measure, and the mean of the weight distribution at the last row. According to this table, AML is the best system in terms of both F-measure and ECP. In addition, while LogMap and LogMapBio have the same F-measure, the former has better ECP due to its lower execution time and the latter is ranked as seventh opposed to its second rank regarding F-measure. Another significant difference is the

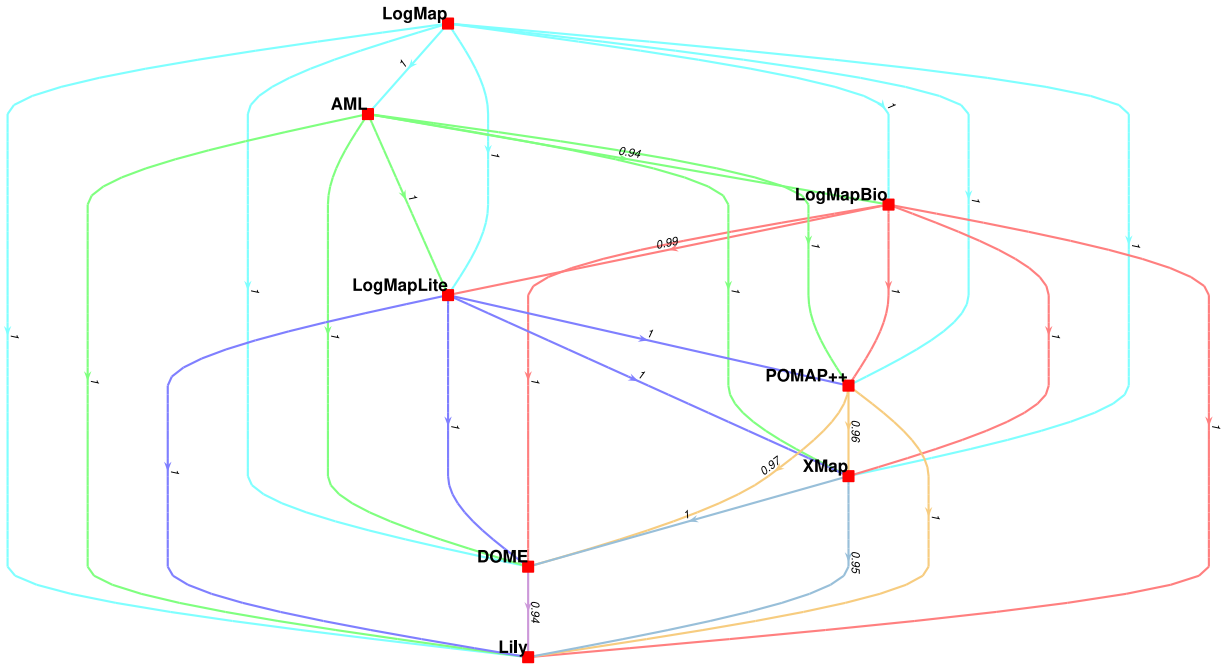


Fig. 9. The credal outranking of systems participated in the OAEI 2018 disease and phenotype track. The tasks involves matching HP to MP.

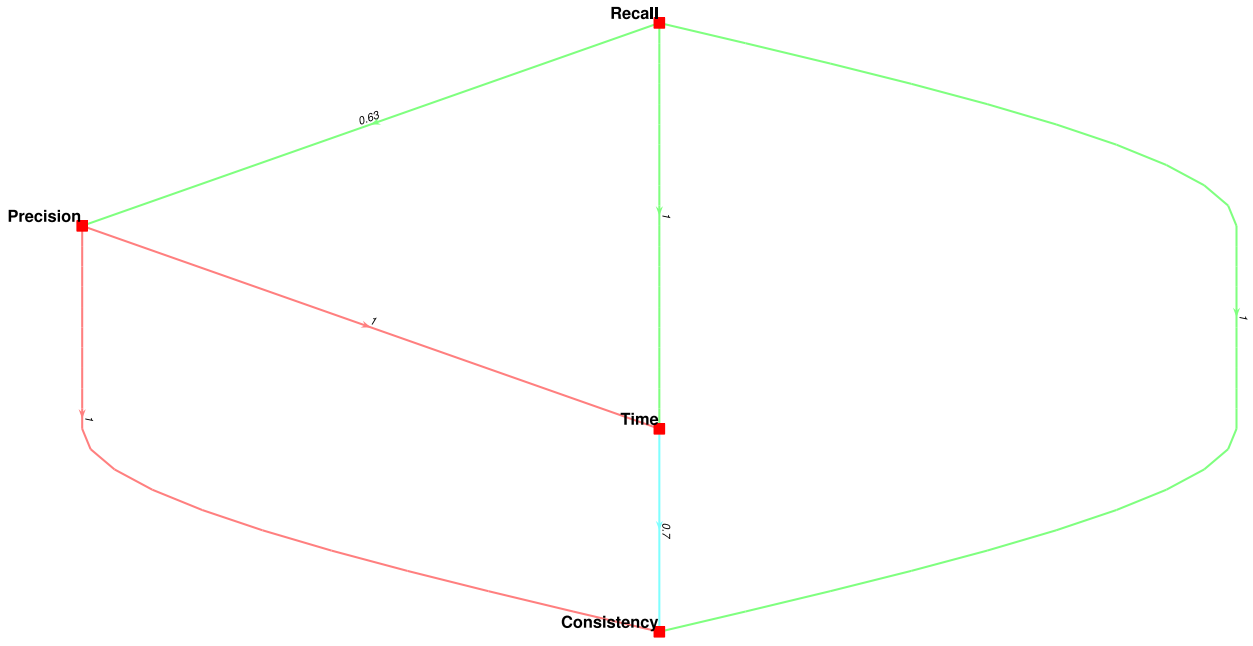


Fig. 10. The credal ranking of performance metrics for the OAEI large biomedical track.

rank of LogMapLite that is the fifth regarding F-measure, but the third with respect to ECP. DOME, which is the last system with regard to F-measure, outperforms LogMapBio due to solely having a better execution time. Fig. 11 plots the outranking of alignment systems for mapping FMA to NCI ontologies.

SPIMBENCH track

The SPIMBENCH track is another matching task which is deemed to determine when two OWL instances describe the same Creative Work. There are two datasets, called 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 is required to be aligned to Abox which contains instances only. The

Table 7
Results of alignment systems participated in the 2018 OAEI SPEMBENCH track. The task is Sandbox.

	Time	Precision	Recall	ECP	F-measure
AML	6220	0.83	0.90	0.79	0.86
Lily	1960	0.85	1.00	0.92	0.92
LogMap	5887	0.94	0.76	0.78	0.84
Weight	0.12	0.42	0.46		

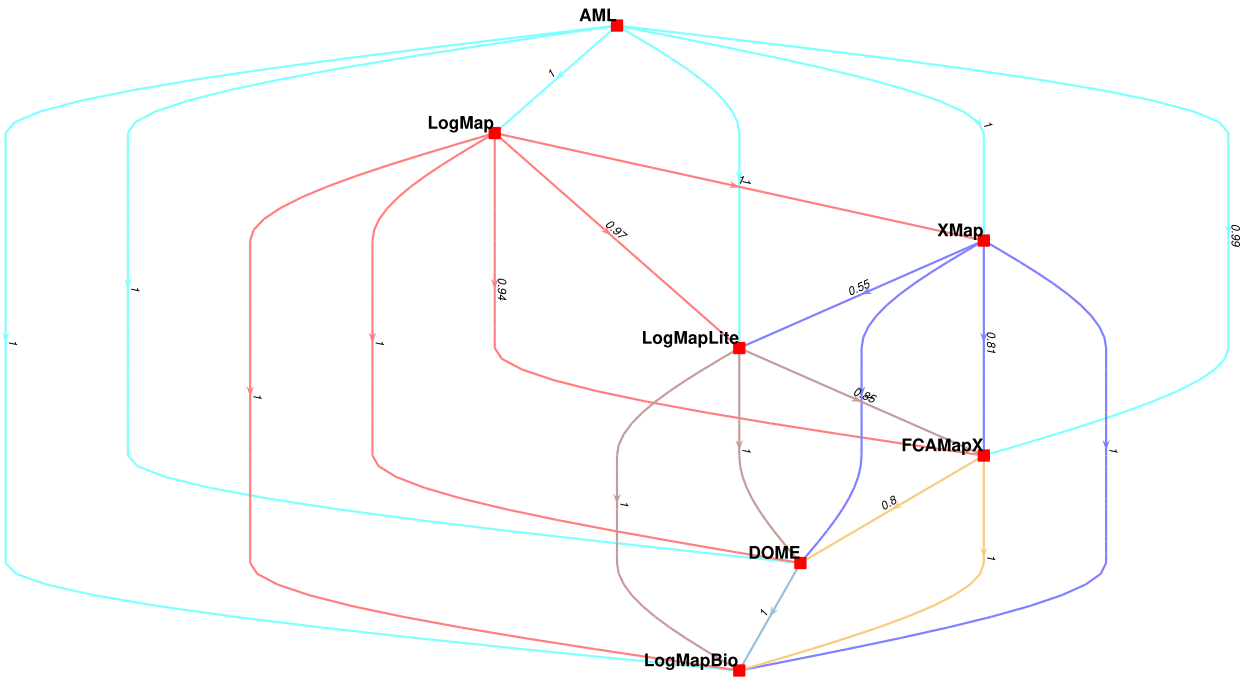


Fig. 11. The credal outranking of alignment systems for aligning FMA and NCI in the large biomedical track.

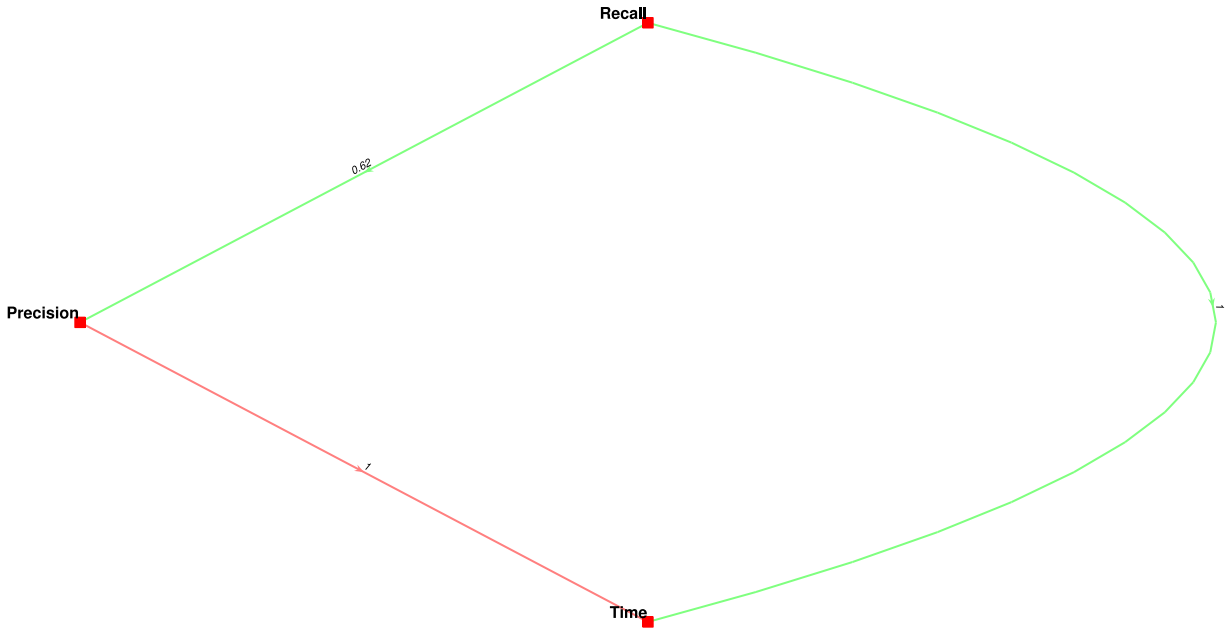


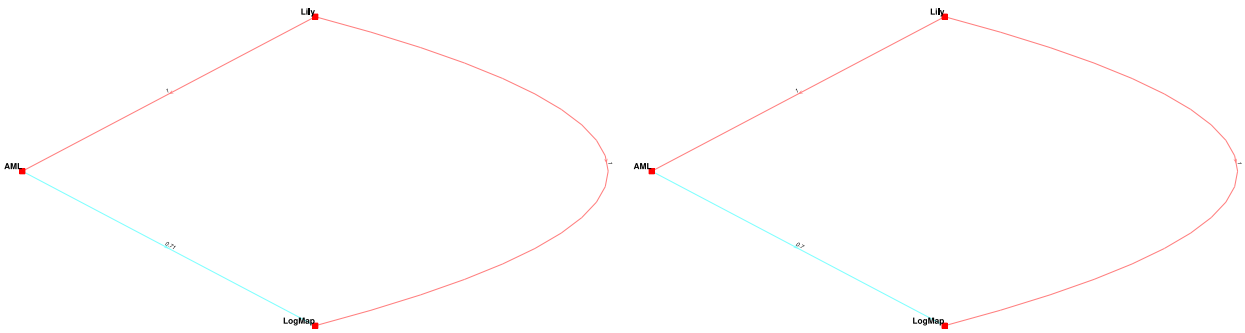
Fig. 12. The credal ranking of performance metrics for the SPIMBENCH track.

Table 8
Results of alignment participated in the 2018 OAEI SPEMBENCH track. The task is Mainbox.

	Time	Precision	Recall	ECP	F-measure
AML	37190	0.84	0.88	0.80	0.86
Lily	3103	0.85	1.00	0.97	0.92
LogMap	23494	0.89	0.71	0.79	0.79
Weight	0.12	0.42	0.46		

difference between Sandbox and Mainbox is that the reference of the former is available to the participant, while the latter is a blind matching task so that participants do not know the real alignment beforehand. 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. Fig. 12 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.

There are only three systems that participated in this track at the OAEI 2018. Tables 7 and 8 tabulates the results of the systems for the Sandbox and Mainbox tasks, respectively, as well as their ECP and F-measure. In addition, Figs. 13(a) and 13(b) plots the



(a) Sandbox task (b) Mainbox task

Fig. 13. The credal outranking of alignment systems participated in the OAEI 2018. There are two tasks in this track: (a) Sandbox; (b) Mainbox.

credal outranking of alignment systems participated at the OAEI 2018 for Sandbox and Mainbox, respectively. In both tasks, the ranking in order of F-measure is in line with that of ECP: Lily is the best system that is significantly superior to AML and LogMap, followed by AML and LogMap.

4. Discussion

The experiments on six OAEI tracks show that, as expected, precision and recall are the most important performance metrics in almost all tracks, except for anatomy where recall+ is identified as the most important metric based on experts' preferences. While such results seem trivial, the advantage of using MCDM is that it can also compute the extent to which one metric is more important than another, allowing us to aggregate the metrics accordingly. Only knowing that precision and recall are more important than the other metrics would not help aggregate the metrics together.

Another essential point about precision and recall is that, while these two metrics are typically the most important ones with identical weights, the importance of these two metrics varies for almost all the OAEI tracks, e.g., anatomy, multifarm, disease and phenotype, large biomedical, and SPIMBENCH. On the other hand, F-measure is a weighted harmonic mean of precision and recall, where the weights of precision and recall are set to 0.5. What can be done instead is to use the weights of precision and recall computed based on experts' preferences and calculate F-measure accordingly. Such an F-measure is more informative and can be adapted based on the needs of different ontology matching application and tasks.

It is important to note that, while the results obtained by the proposed ECP metric could be compared with those of F-measure, applying statistical methods to verify the difference between the rankings of ECP and F-measure do not corroborate the suitability of the proposed method nor its improperness, and the OAEI tracks are just to show the applicability of the proposed method.

The proposed MCDM-based approach has been used to evaluate and compare ontology alignment systems based on multiple performance metrics, where the importance of each performance metric is identified by using an MCDM method, i.e, Bayesian best–worst method. Similarly, application developers as well as other users can also use the proposed methodology to select the most appropriate ontology alignment system for their task. In this regard, instead of using experts' preferences, they should first identify the appropriate performance metrics for the task at hand, and then express their preferences over the selected performance metrics using Bayesian BWM. Then, Steps 3 and 4 of the proposed methodology shown in Fig. 1 are taken to evaluate, compare, and rank the alignment systems, or select the best alignment system for their task.

5. Conclusion

This paper modeled the evaluation and comparison of alignment systems with respect to multiple performance metrics as a multi-criteria decision-making (MCDM) problem, where performance metrics and alignment systems served as the criteria and alternatives, respectively. We elicited the preferences of ontology alignment experts on the performance metrics for different OAEI tracks to calibrate the importance of each metric as well as the extent to which one metric is preferred over another. Based on this calibration, we introduced the expert-based collective performance (ECP) metric that includes multiple performance scores for evaluation and comparison of alignment systems. We showed that the rankings of alignment systems in order of ECP is different from those of F-measure, which only includes precision and recall. In addition, the credal outranking of alignment systems obtained and visualized by a weighted, directed graph. While we focused the experiments on the OAEI tracks to demonstrate the applicability of the MCDM-based comparison and evaluation, the proposed methodology can be used to evaluate and compare ontology alignment systems for any tasks or application based on multiple performance metrics.

A possible way for future research is to use statistical methods such as Bayesian models proposed in [19] that provides more information about the performance of alignment systems. There are few studies in MCDM that can handle distributional inputs. As a result, a new method should be especially-tailored for alignment comparison for distributional inputs. In addition, in some cases, experts have different importance that can change their influence on the final aggregated weights of criteria. This is another area that can be studied in future research. An evaluation and comparison based on Bayesian statistics and MCDM are more comprehensive and reliable that can shed lights on the alignment systems performance to further advance them. Further, the same methodology can be applied to different ontology alignment applications by devising a survey based on the performance metrics being important to different applications, according to which the alignment systems can be compared and ranked.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Faria, C. Pesquita, I. Mott, C. Martins, F.M. Couto, I.F. Cruz, Tackling the challenges of matching biomedical ontologies, J. Biomed. Semant. 9 (1) (2018) 4.

[2] L. Zhou, M. Cheatham, A. Krisnadhi, P. Hitzler, A complex alignment benchmark: Geolink dataset, in: International Semantic Web Conference, Springer, 2018, pp. 273–288.
[3] L. Otero-Cerdeira, F.J. Rodríguez-Martínez, A. Gómez-Rodríguez, Ontology matching: A literature review, Expert Syst. Appl. 42 (2) (2015) 949–971.
[4] P. Jain, P. Hitzler, A.P. Sheth, K. Verma, P.Z. Yeh, Ontology alignment for linked open data, in: International Semantic Web Conference, Springer, 2010, pp. 402–417.
[5] Z. Duo, L. Juan-Zi, X. Bin, Web service annotation using ontology mapping, in: Service-Oriented System Engineering, 2005. SOSE 2005. IEEE International Workshop, IEEE, 2005, pp. 235–242.
[6] F. Wiesman, N. Roos, P. Vogt, Automatic ontology mapping for agent communication, in: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2, 2002, pp. 563–564.
[7] M. Mohammadi, Ontology Alignment: Simulated Annealing-Based System, Statistical Evaluation, and Application to Logistics Interoperability (Ph.D. thesis), Delft University of Technology, 2020.
[8] E. Jiménez-Ruiz, B.C. Grau, I. Horrocks, R. Berlanga, Ontology integration using mappings: Towards getting the right logical consequences, in: European Semantic Web Conference, Springer, 2009, pp. 173–187.
[9] N.F. Noy, Semantic integration: a survey of ontology-based approaches, ACM Sigmod Record 33 (4) (2004) 65–70.
[10] C. Pedrinaci, J. Domingue, et al., Toward the next wave of services: Linked services for the web of data, J. ucs 16 (13) (2010) 1694–1719.
[11] J.J. Jung, Reusing ontology mappings for query routing in semantic peer-to-peer environment, Inform. Sci. 180 (17) (2010) 3248–3257.
[12] J. Euzenat, Semantic precision and recall for ontology alignment evaluation, in: IJCAI, vol. 7, 2007, pp. 348–353.
[13] W.R. Van Hage, A. Isaac, Z. Aleksovski, Sample evaluation of ontology-matching systems, in: EON, vol. 2007, 2007, pp. 41–50.
[14] J. Euzenat, P. Shvaiko, et al., Ontology Matching, vol. 18, Springer, 2007.
[15] A. Solimando, E. Jiménez-Ruiz, G. Guerrini, Detecting and correcting conservativity principle violations in ontology-to-ontology mappings, in: International Semantic Web Conference, Springer, 2014, pp. 1–16.
[16] A. Solimando, E. Jiménez-Ruiz, G. Guerrini, A multi-strategy approach for detecting and correcting conservativity principle violations in ontology alignments, in: OWLED, 2014, pp. 13–24.
[17] M. Mohammadi, W. Hofman, Y.-H. Tan, A comparative study of ontology matching systems via inferential statistics, IEEE Trans. Knowl. Data Eng. 31 (4) (2018) 615–628.
[18] M. Mohammadi, A.A. Atashin, W. Hofman, Y. Tan, Comparison of ontology alignment systems across single matching task via the Mcnemar's test, ACM Trans. Knowl. Discov. Data (TKDD) 12 (4) (2018) 1–18.
[19] M. Mohammadi, Bayesian Evaluation and comparison of ontology alignment systems, IEEE Access 7 (2019) 55035–55049.
[20] M. Mohammadi, J. Rezaei, Bayesian Best-worst method: A probabilistic group decision making model, Omega (2019) 102075.
[21] W.R. Gilks, S. Richardson, D. Spiegelhalter, Markov Chain Monte carlo in Practice, CRC press, 1995.
[22] M. Plummer, JAGS: Just another gibbs sampler, 2004.