Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 31(9): 5732–5748.
DOI: 10.3934/era.2023291
Received: 12 July 2023
Revised: 05 August 2023
Accepted: 13 August 2023
Published: 21 August 2023

Interactive complex ontology matching with local and global similarity deviations

Xingsi Xue[1,2,*] **and Miao Ye**[3]

[1] Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, China

[2] Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin Universitiy of Electronic Technology, Guilin 541004, China

[3] School of Information and Communication, Guilin University of Electronic Technology, Guilin 540014, China

* **Correspondence:** E-mail: jack8375@gmail.com.

Abstract: Ontology serves as a central technique in the semantic web to elucidate domain knowledge. The challenge of dealing with the heterogeneity introduced by diverse domain ontologies necessitates ontology matching, a process designed to identify semantically interconnected entities within these ontologies. This task is inherently complex due to the broad, diverse entities and the rich semantics inherent in vocabularies. To tackle this challenge, we bring forth a new interactive ontology matching method with local and global similarity deviations (IOM-LGSD) for ontology matching, which consists of three novel components. First, a local and global similarity deviation (LGSD) metrics are presented to measure the consistency of similarity measures (SMs) and single out the less consistent SMs for user validation. Second, we present a genetic algorithm (GA) based SM selector to evolve the SM subsets. Lastly, a problem-specific induced ordered weighting aggregating (IOWA) operator based SM aggregator is proposed to assess the quality of selected SMs. The experiment evaluates IOM-LGSD with the ontology alignment evaluation initiative (OAEI) Benchmark and three real-world sensor ontologies. The evaluation underscores the effectiveness of IOM-LGSD in efficiently identifying high-quality ontology alignments, which consistently outperforms comparative methods in terms of effectiveness and efficiency.

Keywords: ontology matching; similarity measure; similarity deviation; induced ordered weighting aggregating; genetic algorithm

1. Introduction

Semantic web (SW) [1] represents an enhancement of the current World Wide Web, offering a standard way to interpret content and provide data and information with explicit meaning, thereby enabling machines and people to work in cooperation. Ontology [2], an integral aspect of the SW, refers to the formal representation of knowledge within a specific domain. Ontologies provide a structured and unified way to represent, share and reuse knowledge across various domains and platforms, which facilitate data interoperability [3], enhance information retrieval [4], enable knowledge sharing and reuse [5] and provide a foundation for intelligent information processing [6]. However, with the proliferation of ontologies across different domains, a new issue has emerged known as the ontology heterogeneity problem [7], which is due to differences in naming schemes, entity definitions, modeling methods or even domain coverage across multiple ontologies. The negative impacts of this problem include the inability to effectively share or integrate data, a reduction in the usefulness of data and the possibility of incorrect information retrieval or data analysis results. Ontology matching (OM) [8], an effective solution to the Ontology Heterogeneity problem, involves identifying similar entities from different ontologies. It enables the harmonization of diverse ontologies by finding correspondences between them, facilitating more effective communication and information exchange between different systems and platforms. Nevertheless, the challenge of OM grows exponentially as the scale and complexity of entities increase. Handling complex, diverse, large-scale entities and adapting to complicated heterogeneous situations remain open challenges in the field of OM [9].

Similarity measure (SM) [10] refers to a metric that quantifies the degree of similarity between two entities within different ontologies. SMs are of paramount importance in tackling the ontology heterogeneity problem since they facilitate the identification of matches between heterogeneous ontologie by quantifying how similar two entities are [11], enabling effective data interoperability and integration. However, it is crucial to select and combine appropriate similarity measures to ensure the quality of the matching results since not all SMs are effective in all situations [12]. The efficiency of a SM can vary depending on the specific characteristics of the ontologies being matched, such as their complexity, the domain they belong to or their level of detail. Therefore, a combination of multiple SMs might be required to achieve a more comprehensive and accurate match. Selecting and combining proper SMs presents a significant challenge due to the variety and complexity of ontologies and the lack of a one-size-fits-all solution. Factors like the size and diversity of ontologies, the degree of heterogeneity, the computational complexity and the inconsistency among different measures can pose difficulties [13]. Furthermore, the optimal combination of similarity measures may vary from one case to another, requiring the development of dynamic strategies that can adapt to the specifics of each matching task.

Genetic algorithms (GAs) [14] are a class of evolutionary algorithms inspired by the process of natural selection. For optimizing the selection of suitable SMs for OM, GAs are particularly apt due to their ability to explore a vast solution space and their adaptive nature [15, 16]. They can work towards optimal or near-optimal solutions by iteratively refining the population of similarity measures, providing a robust approach when dealing with the high dimensionality and complex interaction of measures in ontology matching. The use of GAs allows the system to learn from past matching tasks and incrementally improve the selection and combination of similarity measures, leading to better-quality results over time. Despite their efficacy and adaptability, GAs do exhibit certain constraints that make

user involvement indispensable in the selection of SMs. Primarily, by virtue of being heuristic methods, GAs cannot always guarantee the identification of the globally optimal solution. Particularly when dealing with an intricate array of candidate SMs and their interactions, GAs may fall prey to local optima, resulting in less-than-ideal matching outcomes. This is where user intervention can significantly mitigate these shortcomings, since with their domain-specific expertise, users can offer invaluable insights that guide the selection of SMs [17]. However, striking the right balance between leveraging user input and maintaining the advantages of automated processes is a nuanced challenge [18]. It demands deliberate planning and intelligent design to effectively integrate human expertise without undermining the benefits of automation, thus optimizing the OM process.

The ordered weighted aggregating (OWA) operator [19] is a popular mathematical tool for information aggregation. It combines the ideas of ordered set and weighted arithmetic mean to carry out an aggregation, which means it considers both the magnitude of each input and its relative rank. This characteristic enables the OWA operator to capture a wide range of aggregation behaviors, from pure 'and'-like (min) to pure 'or'-like (max) behaviors, and everything in between. Compared to traditional methods, the OWA operator's suitability for aggregating SMs lies in its flexibility and inclusiveness. It provides an approach that can handle a variety of situations and account for different interaction effects among similarity measures. Unlike methods that simply average or select the maximum or minimum similarity, the OWA operator considers the relative importance of each measure, thus providing a more comprehensive and nuanced aggregation. However, its fixed, predefined weights do not adapt to the varying characteristics or importance of the similarity measures, making it insensitive to the context of specific matching tasks [20]. Therefore, the results might not accurately reflect the true similarity if there are significant differences in the importance or reliability of the measures. This inflexibility makes it less responsive to the context of specific matching tasks, and as a result, its outputs may not accurately mirror the true similarity, especially when there are notable discrepancies in the importance or dependability of the measures.

Our main goal is to develop an interactive ontology matching method with local and global similarity deviations (IOM-LGSD) to improve the quality of matching results. To begin, we use the LGSD, which quantifies the consistency of various SMs. The LGSD identifies less consistent SMs and presents them for user validation. The user feedback refines the LGSD component and reduces the deviations in the SMs, consequently improving the quality of matching. The LGSD is a dynamic tool that adjusts in response to user inputs, thereby ensuring a more reliable selection of SMs for the next steps. Following this, the GA-based SM selector enters the process. This component utilizes the refined SMs and employs a genetic algorithm to evolve the optimal subset of these measures. The GA-based SM selector capitalizes on evolutionary computing principles to dynamically optimize the SM subsets. It adapts not only to the evolving set of SMs but also determines when to invoke user involvement, making the matching process both efficient and interactive. Finally, the IOWA-based SM aggregator takes the optimized SMs, as selected by the GA-based selector, and aggregates them. This aggregator uses an induced ordered weighting approach, providing a flexible and dynamic weighting scheme. It adapts to the specific reliabilities of the SMs, ensuring the final aggregation is a robust and comprehensive similarity measure that can effectively match ontologies. In essence, the LGSD lays the groundwork for a reliable set of SMs through user validation. The GA-based selector then optimizes this set and feeds it into the IOWA-based aggregator, which finally produces the comprehensive measure for ontology matching. This cyclic, adaptive and interactive mechanism ensures a constantly improving,

user-involved ontology matching process. We will make sure to include a clearer articulation of this process in the paper for the benefit of the readers. The major contributions are as follows:

- A LGSD is designed to measure the consistency of SMs and identify the less consistent SM for user validation;
- A GA-based SM selector is presented to evolve the SM subsets. This algorithm can automatically optimize the selected SMs and adaptively determine the timing of user involvement;
- A problem-specific IOWA based SM aggregator is proposed to evaluate the quality of selected SMs. It provides a more flexible and dynamic weighting scheme based on the approximate metrics on alignment' quality, capable of adapting to the specific reliabilities of the SMs being aggregated.

The remainder of this paper is structured as follows: Section 2 introduces the fundamental concepts of OM and reviews the related work. Section 3 illustrates the framework of IOM-LGSD. Section 4 details the IOWA-based similarity deviation metrics. In Section 5, GA-based SM selector is presented. Experimental results are provided in Section 6. Lastly, Section 7 summarizes the conclusions and outlines directions for future research.

2. Background

2.1. Ontology and ontology matching

Ontology is a formal representation of knowledge within a specific domain, usually presented as a set of entities, i.e., classes, properties and instances [21]. It serves as a common vocabulary for researchers and forms the basis for semantic interoperability among various systems. OM is the process of identifying semantically equivalent entities (such as classes or properties) from different ontologies [22]. The output of an OM process is the a set of correspondences, so-called the ontology alignment, where each correspondence is a triple $< e_1, e_2, r, conf >$ indicating that the relationship r holds between the entities e_1 and e_2 from the two different ontologies, the confidence of holding this relationship is $conf$.

Evaluation of ontology alignments is crucial to measure the effectiveness of ontology matching. Commonly used metrics include recall, precision and f-measure [23]. Recall is the ratio of correctly identified correspondences to all actual correspondences, measuring the ability of the method to find all relevant matches. Precision, on the other hand, is the ratio of correctly identified correspondences to all identified correspondences, gauging the accuracy of the matches found. The f-measure is the harmonic mean of recall and precision, offering a balanced measure of the performance of ontology matching. These metrics help assess the quality of ontology alignment, ensuring the integrity and usability of the matched data.

2.2. Related work

OM techniques generally fall into two main categories: machine learning-based and heuristic-based methods [24]. Machine learning-based techniques [9, 25] leverage learning algorithms to predict matching pairs based on training data, while heuristic-based methods [26] use expert knowledge and rules to determine the matches. Despite the increasing popularity of machine learning methods, heuristic-based OM techniques are often preferred due to their intuitive nature and lower computational requirements. They don't require large amounts of labeled data for training, making them more feasible

in situations where such data is scarce or unavailable. Moreover, they allow for human-readable rules and expert knowledge incorporation, providing more interpretability and control over the matching process. Among the heuristic methods, GA have gained considerable attention in the realm of OM [27]. This can be attributed to their unique search mechanism, inspired by the principles of natural evolution such as selection, mutation and crossover. GAs are particularly suited for optimizing the selection and combination of SMs in OM due to their ability to explore a large solution space effectively, and adjust dynamically to find the best set of measures.

Automatch [28] uses a GA to automatically learn the optimal weights of different SMs from training data. Its primary merit is its ability to adapt to different domains without requiring manual intervention. However, the quality of the alignment significantly depends on the availability and quality of the training data. OntoDNA [29] applies a GA to optimize the combination of multiple matching strategies. It provides a high level of flexibility and adaptability. However, the initial configuration and parameter tuning can be challenging and time-consuming. LogMap [30] utilize a GA for optimizing the mapping repair process, making it one of the most effective repair systems. Despite their strengths, LogMap may struggle with large and complex ontologies due to computational limitations. Optima [31] uses a GA to create an optimal linear combination of base matchers. It can adaptively tune the weights of different matchers based on their performance. However, its limitation lies in the fact that it assumes all base matchers are independent and does not take into account possible interactions between them. Hertuda [32] also uses a GA to find the optimal combination of similarity measures. It shows good results in detecting complex correspondences. However, the use of a GA makes the system computationally intensive and slower compared to other methods. AgreementMakerLight (AML) [33] employs a GA for weight optimization in the process of aligning large ontologies. It exhibits impressive performance on the OAEI's large biomedical track. However, it requires a significant amount of computational resources, which might not be feasible in all scenarios.

While each of the methods, such as AML, Optima and Hertuda, have advanced the field of ontology matching with their unique contributions, several shortcomings still remain. A pervasive limitation across these methods relates to their limited adaptability and the depth of solution space exploration. These methods primarily engage GAs in the optimization of existing parameters, rather than uncovering novel amalgamations of SMs. Consequently, this confines the discovery potential for inventive solutions that could lead to more accurate matching results. Furthermore, certain assumptions made by these methods regarding SMs introduce other limitations. For example, Optima, in its approach, presumes that the SMs operate independently from one another. This assumption tends to disregard the potential interplay among SMs, which can be a pivotal factor in devising effective OM strategies. Moreover, previous work tends to overlook the potential advantage of more interactive approaches that leverage user feedback in OM. This user-centric perspective can bring to light domain-specific knowledge and preferences that might not be readily apparent to algorithmic approaches, thereby improving the quality of ontology alignment. These highlighted limitations underscore the necessity for a more exploratory, interactive and integrative approach that not only enhances the exploration of the solution space but also appreciates the intricate interdependencies among SMs. Addressing these shortcomings, we put forth an interactive GA for SM selection in OM, which is tailored to draw insights from human involvement. This proposed method is designed to transcend the constraints of existing techniques, providing a more efficient navigation of the solution space and ultimately aiming for enhanced OM results. Our approach, thereby, signifies a crucial stride towards bridging the gap between purely algorithmic approaches and

those that effectively leverage user input for improved performance.

3. The framework of interactive ontology matching with local and global similarity deviation

IOM-LGSD seeks user input primarily in two ways. First, during the selection of SMs, user feedback is solicited to affirm the relevance and applicability of the proposed measures. The GA-based SM selector in the IOM-LGSD presents a set of potential SMs to the user, based on which the user can either approve or recommend alterations. This feedback serves to steer the SM selection process in a direction that aligns with the user's understanding of the domain knowledge, thereby enhancing the quality of matching results. Second, the LGSD component identifies less consistent SMs, and these instances are presented to the user for validation. By confirming or rejecting these instances, users contribute to improving the consistency of SMs in the ontology matching process.

Balancing user input and automation is crucial in our method to ensure the efficiency and effectiveness of ontology matching. The proposed method minimizes the need for constant user input by smartly determining the timing of user involvement. Rather than relying on users to make every minor decision, our method involves users at key decision points, primarily when selecting SMs and validating less consistent SMs, thereby ensuring a balance between user involvement and automation. The automated components of the system utilize user input as guidance to refine the selection and combination of SMs, and the timing of user validation is adaptively determined to reduce unnecessary user involvement. Moreover, the IOM-LGSD mechanism is designed to learn from user feedback over time. As users continue to interact with the system, the IOM-LGSD adapts its decision-making process in line with user preferences, further minimizing the need for user input and optimizing the balance between user involvement and automation.

4. Local and global similarity deviation

The LGSD is a new measure designed to gauge the consistency of SMs. It functions by considering the deviations in the similarity scores provided by the various SMs for a given pair of entities. The "Local" and "Global" aspects pertain to the levels at which these deviations are assessed. The Local Similarity Deviation (LSD) evaluates the variance of the similarity scores for each entity pair, considering the scores provided by different SMs. The LSD helps in identifying entity pairs where the selected SMs produce significantly divergent similarity scores, suggesting a potential inconsistency in the assessment of how similar these entities are. The global similarity deviation (GSD), on the other hand, operates at a higher, aggregate level. It assesses the overall variation in the similarity scores across all entity pairs for a given SM. The GSD is beneficial in identifying SMs that tend to produce divergent similarity scores across multiple entity pairs. Such SMs might be less reliable or less suitable for the specific ontology matching task at hand. In terms of application within our IOM-LGSD method, the LGSD serves a twofold purpose. First, it aids in identifying less consistent SMs or less reliable entity matches, which are then flagged for user validation. Involving the user at these critical points enhances the robustness of the ontology matching results while optimizing the use of the user's time and effort. Second, the LGSD is used as a feedback mechanism in the GA-based SM selector, contributing to the evolution of the SM subsets by emphasizing the selection of more consistent SMs. In summary, the LGSD adds an essential layer of quality control to the ontology matching process, enabling the detection and resolution

of potential inconsistencies and enhancing the overall accuracy and reliability of the matching results.

4.1. Approximate metrics on ontology alignment

In OM, the reliability of a SM is contingent upon the correctness of its corresponding alignment. However, the inherent challenge lies in the fact that the ground truth alignment is not available beforehand. Consequently, accurately determining a SM's confidence prior to the OM process poses a significant challenge. Addressing this concern requires an effective yet approximate metric that can anticipate the quality of an alignment, subsequently allowing us to calculate a SM's confidence. Given SM's corresponding alignment A, its quality can approximately measured in terms of the following completeness and correctness metrics:

$$completeness(A) = \frac{|E_{O_1}| + |E_{O_2}|}{|O_1| + |O_2|}, \tag{4.1}$$

$$correctness(A) = \frac{\sum_{i=1}^{|A|} sim_i}{|A|}, \tag{4.2}$$

where $|E_{O_1}|$ and $|E_{O_2}|$ are the number of matched entities in ontologies O_1 and O_2, respectively; and $|O_1|$ and $|O_2|$ are the entity scales of O_1 and O_1, respectively; sim_i is the similarity value of the i-th entity mapping, and $|A|$ is the number of correspondences in A. $completeness(\cdot)$ and $correctness(\cdot)$ aims to predict A's completeness and correctness, respectively, and we aggregate them according to the following formula:

$$Q(A) = \sqrt{completeness(A) \times correctness(A)}. \tag{4.3}$$

While these formulas serve as an approximation, they are expected to provide a useful prediction of the alignment's quality and the SM's reliability.

4.2. A new IOWA operator for ontology matching

IOWA operator is a dynamic tool where "Ordered" refers to a prerequisite step where scores from various SMs are arranged prior to aggregation. The term 'Weighted Aggregating' implies that the influence of individual input values on the aggregate outcome varies based on assigned weights. The unique aspect of this operator, "Induced", demonstrates the adaptability of these weights as they are dictated by the input values themselves rather than being predetermined. In our methodology, we employ the IOWA operator to provide a dynamic, flexible weighting scheme adept at accommodating the unique reliability of each SM, thereby furnishing an effective approach to assess the quality of selected SMs.

The salient advantage of the OWA operator is its innate capacity to encapsulate a broad spectrum of aggregation behaviours. It surpasses traditional aggregation methods, such as averages, maxima, or minima, by considering not just the magnitude but also the relative importance of each measure. By doing so, it addresses diverse interaction effects among SMs and is able to cope with a variety of situations. The end result is a comprehensive and inclusive aggregation, offering a superior evaluation of SM quality. In particular, given a set of inputs $x_1, x_2, \cdots, x_n$ and a corresponding set of weights $w_1, w_2, \cdots, w_n$, the OWA operator F_{OWA} is defined as follows:

$$F_{OWA}(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} w_i x_i, \tag{4.4}$$

where $x_1, x_2, \cdots, x_n$ is the sorted (in descending order) version of the inputs.

While OWA is already well-suited to the task of aggregating SMs due to its balance between 'and'-like (minimum) and 'or'-like (maximum) aggregation behaviors, it still operates under fixed, predefined weights. This rigidity can limit its sensitivity to the context-specific importance or reliability of the similarity measures, which is often crucial in complex OM tasks. The IOWA operator [34], in contrast, introduces an additional set of induced variables that adaptively determine the ranking of inputs, offering a more flexible aggregation method. In other words, the weights assigned to SMs in IOWA are context-dependent and can vary based on the specific matching task at hand. This capability to adaptively assign weights based on the induced ordering allows IOWA to more accurately reflect the true similarity in situations with significant differences in the importance or reliability of measures. By addressing the fixed weight limitations of OWA, the IOWA operator therefore provides a promising avenue for improving the quality and precision of matching results, accommodating the complex and dynamic nature of the task. Formally, the IOWA operator F_{IOWA} is defined as follows:

$$F_{IOWA}(x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} w_{induced,i} \times x_i, \tag{4.5}$$

where $x_1, x_2, \cdots, x_n$ are the inputs to be aggregated, typically the values or scores obtained from different SMs in the context of OM; $w_{induced,1}, w_{induced,2}, \cdots, w_{induced,n}$ are the weights corresponding to each input, which dictate the importance of each input in the aggregated output, and higher weights signify greater importance; $b_1, b_2, \cdots, b_n$ are the induced variables, which provide additional context that influences the ordering of the inputs and weights; $w_{induced,i}$ and x_i represent the weights and inputs sorted in ascending order and descending order, respectively, based on the ordering induced by $con_1, con_2, \cdots, con_n$.

In this work, the weights in the IOWA operator, determined by the $correctness(\cdot)$ values of SMs, reflect the accuracy of each SM. The correctness of a SM refers to how closely the alignment it proposes corresponds with the correct alignment. Therefore, a higher correctness value means the SM is more reliable or trustworthy, and its outputs should be given more importance in the aggregation process. On the other hand, the induced variables, determined by the $completeness(\cdot)$ values of the SMs, represent the comprehensiveness of each SM. The completeness of a SM indicates the extent to which the SM can identify all valid correspondences between the entities of two ontologies. Therefore, a higher completeness value means the SM is more capable of discovering correspondences, and the outputs of such SMs should be prioritized in the ordering of the aggregation. In essence, using correctness and completeness values to set the weights and induced variables, respectively, ensures that the IOWA operator can aggregate the SMs' outputs in a way that not only gives more importance to more accurate SMs, but also takes into account the comprehensiveness of the SMs. This approach, therefore, leads to a more balanced and effective ontology matching process.

Lastly, let's consider a similarity value vector on an entity pair $x = (0.5, 0.8, 0.7)$ whose elements are respectively determined by three SMs, and its corresponding weights $w = (0.4, 0.3, 0.3)$ whose elements are determined by three SMs' $correctness(\cdot)$ values (Eq 4.2), respectively. We also introduce the induced variables $b = (0.2, 0.1, 0.3)$ whose elements are determined by three SMs' $completeness(\cdot)$ values (Eq 4.1), respectively. The operation of the IOWA operator first involves ordering the induced variables in ascending order: $b_{asc} = (0.1, 0.2, 0.3)$. Based on the ordering of the induced variables, we now rearrange our inputs and weights: $x_{induced} = (0.8, 0.5, 0.7)$, $w_{induced} = (0.3, 0.4, 0.3)$. The IOWA aggregation is calculated as follows: $F_{IOWA}(x) = w_{induced,1} \times x_{induced,1} + w_{induced,2} \times x_{induced,2} + w_{induced,3} \times x_{induced,3} =$

$0.3 \times 0.8 + 0.4 \times 0.5 + 0.3 \times 0.7 = 0.24 + 0.2 + 0.21 = 0.65.$

4.3. Local and global similarity deviation metrics

In our study, we propose two new metrics, local and global similarity deviations, to identify less consistent SMs. Given a SM set, $SM = sm_1, sm_2, \cdots, sm_n$, with corresponding similarity matrices $M^1, M^2, \cdots, M^n$, we first categorize SMs into three groups: syntax-based (SM_{syn}), linguistic-based (SM_{lin}) and structure-based (SM_{str}). Subsequently, we use the Induced Ordered Weighted Aggregating (IOWA) operator to compute four aggregated similarity matrices: M^{all}, M^{syn}, M^{lin}, M^{str}. Each SM, sm_i, has local and global similarity deviations computed as the average absolute difference between its corresponding similarity matrix M^i and its respective local (category-specific) or global aggregated similarity matrices:

$$simDeviation_{local}(sm_i, sm_{categ}) = \frac{\sum |M^i_{j,k} - M^{categ}_{j,k}|}{|O_1| \times |O_2|}, \tag{4.6}$$

$$simDeviation_{global}(sm_i, sm_{all}) = \frac{\sum |M^i_{j,k} - M^{all}_{j,k}|}{|O_1| \times |O_2|}, \tag{4.7}$$

where $|O_1|$ and $|O_2|$ are the number of entities in ontology O_1 and O_2, respectively.

5. Genetic algorithm based similarity measure selector

Inspired by the principles of natural evolution, GA is a search heuristic that is known for its capability to find solutions to complex problems. It operates through mechanisms derived from biological evolution, such as selection, mutation and crossover, to evolve a set of solutions towards an optimum. Applying GA to the task of selecting SMs for OM is particularly advantageous due to several reasons. First, the GA's inherent ability to explore a large search space efficiently makes it suitable for handling the vast combinations of SMs that can be considered for matching entities in ontologies. Second, the evolutionary nature of the GA allows for the iterative refinement of the selected set of SMs, gradually leading to better matching results over successive generations. Third, GAs can effectively deal with the multi-objective nature of the problem, optimizing for different criteria like precision and recall, simultaneously. However, the selection of appropriate SMs for ontology matching is a non-trivial problem, as it requires dealing with a high dimensional space with a large number of potential solutions. When GA gets stuck in the local optima, it is necessary to get user involved to guide the algorithm's search direction.

In this study, we utilize a binary encoding scheme to represent each individual as a SM selection solution within GA. Specifically, the length of a chromosome corresponds to the total number of candidate SMs. Each gene is denoted by a binary value: 1 signifies the selection of the corresponding SM, while 0 indicates non-selection. The pseudo-code of GA-based SM selector is presented in Algorithm 1. It commences by initializing a population with all candidate SMs and setting up an elite individual ind_{elite} based on the evaluation function $Q(\cdot)$ (defined by Eq 4.3). As the iterations progress, a new generation is formed via one-point crossover, bit mutation and roulette wheel selection [35]. Notably, user involvement is incorporated if ind_{elite} remains unchanged for a threshold number of δ generations. This interaction targets the four most inconsistent SMs within ind_{elite}: those with the highest local similarity deviation values across the three SM categories, and the one exhibiting the

maximum global similarity deviation value. The user's decision on the inclusion or exclusion of these SMs informs the update of ind_{elite}, enhancing the alignment's quality and interpretability. Furthermore, an elite preservation strategy is deployed, replacing a randomly chosen individual with ind_{elite} to ensure the survival of high-quality solutions. The process continues until it reaches the maximum generation limit $MaxGen$, after which the final ind_{elite} is returned as the output. This procedure, incorporating elements of evolutionary search and user feedback, effectively navigates the SM selection process, balancing the power of automated reasoning with the insights of human expertise.

Algorithm 1: Genetic Algorithm for Similarity Measure Selection

Input: Population size $size_p$, Crossover rate $rate_c$, Mutation rate $rate_m$, Maximum generation $MaxGen$

Output: Elite individual ind_{elite}

$initialize(Population, size_p)$;
$evaluate(Population)$;
$initialize(ind_{elite})$;
$gen = 0$;
while $gen < MaxGen$ **do**
  $Population' = crossover(rate_c)$;
  $Population' = mutation(rate_m)$;
  $Population = selection(Population')$;
  $update(ind_{elite})$;
  if ind_{elite} *keeps unchanged for* δ *generations* **then**
    for *all the selected similarity measures in* ind_{elite} **do**
      $SM_{local,syn} = \arg\max\{simDeviation^{syn}_{lobal,i}\}$;
      $SM_{local,lin} = \arg\max\{simDeviation^{lin}_{lobal,i}\}$;
      $SM_{local,str} = \arg\max\{simDeviation^{str}_{lobal,i}\}$;
      $SM_{global} = \arg\max\{simDeviation_{global,i}\}$;
      Require User Validate $SM_{local,syn}$, $SM_{local,lin}$,$SM_{local,str}$ and SM_{global};
      $update(ind_{elite})$;
    end
  end
  $saveElite()$;
  $gen = gen + 1$;
end
return ind_{elite};

6. Experimental results

6.1. Experimental configuration

In order to evaluate the efficacy of our proposed MLHGP, we utilize Ontology Alignment Evaluation Initiative (OAEI)'s Benchmark * and three real world sensor ontologies: Semantic Sensor Network

*http://oaei.ontologymatching.org/2016/benchmarks/index.html

(SSN)[†], Commonwealth Scientific and Industrial Research Organisation (CSIRO)[‡] and Marine Metadata Interoperability (MMI)[§]. Our choice to incorporate the SSN, CSIRO and MMI sensor ontologies in our experiment was guided by two key factors: (1) these ontologies encapsulate a considerable amount of overlapping information presented in diverse representations, providing a robust challenge for ontology matching, and (2) the SSN is among the most globally referenced ontologies. Furthermore, it offers an alignment with another high-level ontology—DOLCE ultra lite, which can be used as a gold standard alignment for assessing the quality of a generated alignment, particularly in computing f-measure values. This blend of benchmarking and real-world datasets serves to provide a comprehensive and rigorous evaluation of our proposed methodology. The detail descriptions on the test cases is shown in Table 1.

The parameter setting of IOM-LGSD is empirically setting as follows, which can ensure the highest average f-measure on all test cases:

- The population size: 100;
- The maximum generation: 3000;
- The crossover probability: 0.8;
- The mutation probability: 0.05;
- The generation threshold θ of activating user validation: 200.

Table 1. Description on the test cases.

	Test Case	Description
OAEI's Benchmark	201	Each label or identifier of target ontology is replaced by a random one.
	203	A random, but consistent, typo generator is applied to target ontology's labels and comments.
	204	Different naming conventions (uppercasing, underscore, dash, etc.) are used for target ontology's labels, and the comments have been suppressed.
	205	Target ontology's labels are replaced by synonyms, comments have been suppressed.
	248	Target ontology has scrambled labels + no comments + no hierarchy.
	249	Target ontology has scrambled labels + no comments + no instance.
	250	Target ontology has scrambled labels + no comments + no property.
	251	Target ontology has scrambled labels + no comments + flattened hierarchy.

Continued on next page

[†]https://www.w3.org/2005/Incubator/ssn/ssnx/ssn
[‡]https://www.w3.org/2005/Incubator/ssn/wiki/images/4/42/SensorOntology 20090320.owl.xml
[§]https://mmisw.org/ont/mmi/device

	Test Case	Description
	301	The target ontology is from the real world, which is simpler and closer to the source ontology.
	302	The target ontology is from the real world, which is very similar to the previous one but with different extensions and naming conventions.
Sensor Ontology	SSN	An ontology from W3C for describing sensors, their observations and related processes.
	CSIRO	Australia's national science agency conducting research in a broad range of areas, including sensor technologies in various applications.
	MMI	An initiative promoting data sharing and standardization of metadata in marine science disciplines.

6.2. Results and discussion

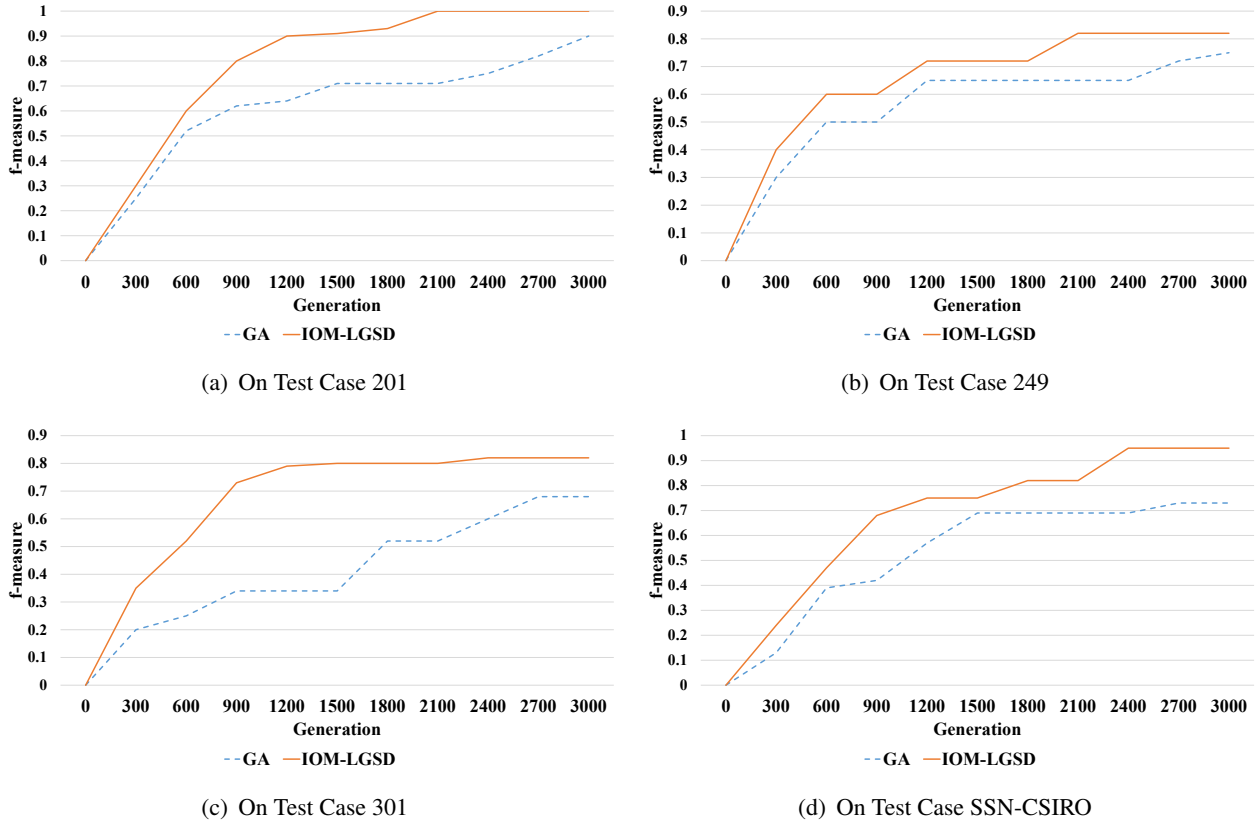
Table 2 presents the f-measure obtained by IOM-LGSD, GA (non-interactive version of IOM-LGSD) and OAEI's participants on all test cases. As shown in the table, IOM-LGSD outperforms the GA and OAEI participants, demonstrating superior effectiveness in all the test scenarios. In particular, for test cases 201–205, IOM-LGSD discovered the optimal alignments, achieving a perfect f-measure score of 1.00. For the more complex test cases 248–301, which pose greater challenges due to sparse entity information in the target ontology, IOM-LGSD continued to excel and demonstrated superior performance compared to the other methods. In the case of matching sensor ontologies, IOM-LGSD's solutions closely approached the optimal matching results, attaining an average f-measure of 0.92, whereas the next best methods, S-Match and CroMatch, achieved f-measure values of 0.82 and 0.80, respectively.

Table 2. Comparison among IOM-LGSD, GA (non-interactive version of IOM-LGSD) and OAEI's participants in terms of f-measure.

	Test case	S-Match [36]	Aroma [37]	RiMOM [10]	ASMOV [38]	CroMatch [39]	GA	IOM-LGSD
OAEI's Benchmark	201	0.86	0.72	0.80	0.80	0.92	0.90	1.00
	203	0.84	0.98	0.94	0.96	1.00	0.99	1.00
	204	0.84	0.83	0.79	0.80	0.86	0.98	1.00
	205	0.85	0.92	0.94	0.95	0.92	0.89	1.00
	248	0.84	0.88	0.81	0.83	0.75	0.85	1.00
	249	0.72	0.77	0.83	0.80	0.72	0.82	0.94
	250	0.84	0.88	0.85	0.87	0.75	0.71	0.92
	251	0.85	0.86	0.80	0.70	0.85	0.80	0.88
	301	0.76	0.72	0.62	0.75	0.62	0.68	0.82
	302	0.75	0.79	0.65	0.71	0.65	0.71	0.88
Sensor Ontology	SSN-MMI	0.80	0.83	0.80	0.82	0.87	0.82	0.90
	SSN-CSIRO	0.80	0.75	0.75	0.72	0.85	0.73	0.95
	MMI-CSIRO	0.85	0.82	0.73	0.73	0.69	0.74	0.92

Figure 1, depicting the comparison between the GA and our proposed IOM-LGSD, elucidates our method's exceptional efficiency. With the number of generations denoted on the horizontal axis and the f-measure on the vertical axis, the figure vividly underscores IOM-LGSD's rapid convergence to

optimal or near-optimal solutions. This behavior significantly outpaces traditional GA across all tested scenarios, underscoring the interactive mechanism's enhanced efficiency within IOM-LGSD.



(a) On Test Case 201

(b) On Test Case 249

(c) On Test Case 301

(d) On Test Case SSN-CSIRO

Figure 1. Comparison between IOM-LGSD and GA (non-interactive version of IOM-LGSD) on Converge Graph.

To conclude, the performance data unequivocally attests to IOM-LGSD's superior efficacy in OM. Its swift convergence, coupled with its consistent outperformance of both GA and the participants of the OAEI, even in complex scenarios with limited entity information, sets a new paradigm in OM methodologies. This compelling evidence illustrates the potential of integrating interactive mechanisms within genetic algorithms, marking a significant stride towards improving both the efficiency and effectiveness of ontology matching.

7. Conclusions and future work

In this work, we explore the challenges of OM in the SW and proposed an innovative approach to enhance the quality of matching results. Ontologies play a crucial role in knowledge representation and sharing in the SW. However, ontology heterogeneity remains a significant issue that hampers the effective integration and interoperability of data. Ontology matching, guided by appropriate SMs, serves as an effective solution to this problem. Despite their effectiveness, the selection and combination of SMs is a complex task, and therefore, requires intelligent mechanisms for optimization. We have presented IOM-LGSD as a novel solution to improve the quality of matching results, which incorporates

a Genetic Algorithm-based SM selector, a Local and Global Similarity Deviation mechanism and an Induced Ordered Weighted Aggregating operator to evaluate the quality of selected SMs. This combination creates an approach that is adaptable, efficient and capable of handling the variety and complexity of ontologies. Through extensive experimentation using the OAEI's benchmark and three real-world sensor ontologies, IOM-LGSD has demonstrated its superior efficiency and effectiveness in ontology matching. It exhibited swift convergence and consistently outperformed the traditional GA and other methods in the field, setting a new benchmark in ontology matching methodologies.

Regarding the IOWA operator, the main constraint could stem from its adaptive nature. While it is designed to adjust to the specific reliabilities of SMs, this might not always be feasible, especially when dealing with complex or rare domains. Furthermore, if the SMs present conflicting results or are skewed, the adaptive nature of the IOWA operator might prove less effective. As for the LGSD, it relies on consistency across SMs to function optimally. In cases where the SMs vary greatly in terms of their results, this could potentially lead to errors or suboptimal results. Moreover, identifying the less consistent SM for user validation might not always lead to improved results, especially if the user lacks comprehensive domain knowledge or expertise. The interaction between the different components of the IOM-LGSD method might also present challenges. While the integration of these components aims to provide a holistic and adaptive approach to ontology matching, it could also lead to potential conflicts or difficulties. For instance, the GA-based SM selector's adaptive determination of user involvement timing might not always align with the LGSD's identification of less consistent SMs. Additionally, user involvement, although beneficial for guiding the selection of SMs and mitigating GA limitations, introduces its own set of challenges. User expertise varies greatly and their decisions can be influenced by subjective bias or lack of sufficient knowledge. This variability can introduce unpredictability into the system, potentially leading to inconsistent performance or unexpected results. In conclusion, while the IOM-LGSD method promises several advantages, its implementation is not without potential issues. Careful tuning, rigorous testing and extensive validation are required to ensure optimal performance across different ontologies and domains.

In the future, we plan to take our validation efforts to a broader spectrum. Despite the positive findings from the OAEI benchmark and real-world sensor ontologies, we assert that extending validation to a more diverse array of sources will fortify the credibility and applicability of our method. We aim to apply our IOM-LGSD mechanism to ontologies across various domains in order to assess its scalability and adaptability and underscore its broad-spectrum relevance. Alongside these efforts, we are committed to enhancing the user-centric aspect of our interactive mechanism. We envision a system that does not just assimilate user feedback but actively tailors its requests for user validation in alignment with the status and results of prior matching tasks. The goal is to deliver a highly personalized ontology matching experience, finely attuned to the particularities of the ontology at hand and the user's distinct knowledge. When it comes to the exploration of additional SMs for OM, we intend to conduct rigorous testing of a wider array of these measures. This will allow us to garner a deeper comprehension of their suitability across diverse contexts and scenarios. Regarding the integration of additional learning algorithms to hone the selection and amalgamation of SMs, we are specifically interested in examining the potential of advanced techniques such as deep learning and reinforcement learning algorithms. Deep learning algorithms have shown effectiveness in managing high-dimensional, large-scale data and could potentially provide nuanced insights into the selection process. On the other hand, reinforcement learning algorithms, with their capability to learn and improve from each decision made, could offer

robust strategies for adaptive SMs amalgamation. By exploring these advancements, we are optimistic about further enhancing the overall efficacy of our ontology matching methodology, particularly in terms of SM selection and combination processes.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62172095), the Natural Science Foundation of Fujian Province (Nos. 2020J01875 and 2022J01644) and Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing (No. CRKL220206), Scientific Research Project of Fujian University of Technology (No. GY-Z23044).

References

1. T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, *Sci. Am.*, **284** (2001), 34–43.
2. N. Guarino, D. Oberle, S. Staab, What is an ontology?, in *Handbook on Ontologies*, Springer, Berlin, 2009. https://doi.org/10.1007/978-3-540-92673-3_0
3. S. Das, P. Hussey, Hl7-fhir-based contsys formal ontology for enabling continuity of care data interoperability, *J. Pers. Med.*, **13** (2023), 1024. https://doi.org/10.3390/jpm13071024
4. A. Sharma, S. Kumar, Machine learning and ontology-based novel semantic document indexing for information retrieval, *Comput. Ind. Eng.*, **176** (2023), 108940. https://doi.org/10.1016/j.cie.2022.108940
5. M. A. Osman, S. A. M. Noah, S. Saad, Ontology-based knowledge management tools for knowledge sharing in organization—A review, *IEEE Access*, **10** (2022), 43267–43283. https://doi.org/10.1109/ACCESS.2022.3163758
6. S. K. Narayanasamy, K. Srinivasan, Y. C. Hu, S. K. Masilamani, K. Y. Huang, A contemporary review on utilizing semantic web technologies in healthcare, virtual communities, and ontology-based information processing systems, *Electronics*, **11** (2022), 453. https://doi.org/10.3390/electronics11030453
7. X. Zhou, Q. Lv, A. Geng, Matching heterogeneous ontologies based on multi-strategy adaptive co-firefly algorithm, *Knowl. Inf. Syst.*, **65** (2023), 2619–2644. https://doi.org/10.1007/s10115-023-01845-2
8. J. Portisch, M. Hladik, H. Paulheim, Background knowledge in ontology matching: A survey, *Semant. Web*, 1–55. https://doi.org/10.3233/SW-223085
9. M. A. Khoudja, M. Fareh, H. Bouarfa, Deep embedding learning with auto-encoder for large-scale ontology matching, *Int. J. Semant. Web Inf. Syst.*, **18** (2022), 1–18.
10. Y. Djenouri, H. Belhadi, K. Akli-Astouati, A. Cano, J. C. W. Lin, An ontology matching approach for semantic modeling: A case study in smart cities, *Comput. Intell.*, **38** (2022), 876–902. https://doi.org/10.1111/coin.12474

11. X. Kou, J. Feng, Y. Wang, W. Cui, A multi-objective particle swarm optimization for matching domain ontologies, *Int. Technol. Lett.*, e405. https://doi.org/10.1002/itl2.405
12. S. Ibrahim, S. Fathalla, J. Lehmann, H. Jabeen, Toward the multilingual semantic web: Multilingual ontology matching and assessment, *IEEE Access*, **11** (2023), 8581–8599. https://doi.org/10.1109/ACCESS.2023.3238871
13. T. Y. Wu, A. Shao, J. S. Pan, Ctoa: Toward a chaotic-based tumbleweed optimization algorithm, *Mathematics*, **11** (2023), 2339. https://doi.org/10.3390/math11102339
14. S. Forrest, Genetic algorithms, *ACM Comput. Surv.*, **28** (1996), 77–80.
15. N. Krishnan, G. Deepak, Easdisco: Toward a novel framework for web service discovery using ontology matching and genetic algorithm, in *Advances in Data Computing, Communication and Security: Proceedings of I3CS2021*, Springer, (2022), 283–291.
16. X. Xue, J. Chen, Matching biomedical ontologies through compact differential evolution algorithm with compact adaption schemes on control parameters, *Neurocomputing*, **458** (2021), 526–534. https://doi.org/10.1016/j.neucom.2020.03.122
17. H. Li, Z. Dragisic, D. Faria, V. Ivanova, E. Jiménez-Ruiz, P. Lambrix, et al., User validation in ontology alignment: functional assessment and impact, *Knowl. Eng. Rev.*, **34** (2019), e15. https://doi.org/10.1017/S0269888919000080
18. T. Y. Wu, H. Li, S. C. Chu, Cppe: An improved phasmatodea population evolution algorithm with chaotic maps, *Mathematics*, **11** (2023), 1977. https://doi.org/10.3390/math11091977
19. A. Dadgar, Y. Baleghi, M. Ezoji, Multi-view data fusion in multi-object tracking with probability density-based ordered weighted aggregation, *Optik*, **262** (2022), 169279. https://doi.org/10.1016/j.ijleo.2022.169279
20. X. Xue, J. Guo, M. Ye, J. Lv, Similarity feature construction for matching ontologies through adaptively aggregating artificial neural networks, *Mathematics*, **11** (2023), 485. https://doi.org/10.3390/math11020485
21. B. Smith, Ontology, in *The Furniture of the World*, Brill, (2012), 47–68. https://doi.org/10.1163/9789401207799_005
22. P. Shvaiko, J. Euzenat, Ontology matching: state of the art and future challenges, *IEEE Trans. Knowl. Data Eng.*, **25** (2011), 158–176. https://doi.org/10.1109/TKDE.2011.253
23. V. M. Tayur, R. Suchithra, Multi-ontology mapping generative adversarial network in internet of things for ontology alignment, *Int. Things*, **20** (2022), 100616. https://doi.org/10.1016/j.iot.2022.100616
24. C. Trojahn, R. Vieira, D. Schmidt, A. Pease, G. Guizzardi, Foundational ontologies meet ontology matching: A survey, *Semant. Web*, **13** (2022), 685–704. https://doi.org/10.3233/SW-210447
25. X. Xue, Q. Huang, Generative adversarial learning for optimizing ontology alignment, *Expert Syst.*, **40** (2023), e12936. https://doi.org/10.1111/exsy.12936
26. J. Fürst, M. Fadel Argerich, B. Cheng, Versamatch: ontology matching with weak supervision, in *49th Conference on Very Large Data Bases (VLDB)*, **16** (2023), 1305–1318.
27. X. Xue, Y. Wang, W. Hao, Using moea/d for optimizing ontology alignments, *Soft Comput.*, **18** (2014), 1589–1601. https://doi.org/10.1007/s00500-013-1165-9

28. J. Berlin, A. Motro, Database schema matching using machine learning with feature selection, in *Advanced Information Systems Engineering*, Springer, (2002), 452–466. https://doi.org/10.1007/3-540-47961-9_32
29. C. C. Kiu, C. S. Lee, Ontodna: Ontology alignment results for oaei 2007, in *Proceedings of the 2nd International Workshop on Ontology Matching (OM-2007) Collocated with the 6th International Semantic Web Conference (ISWC-2007) and the 2nd Asian Semantic Web Conference (ASWC-2007)*, (2007), 304.
30. E. Jiménez-Ruiz, B. C. Grau, Y. Zhou, Logmap 2.0: towards logic-based, scalable and interactive ontology matching, in *Proceedings of the 4th international workshop on semantic web applications and tools for the life sciences*, (2011), 45–46. https://doi.org/10.1145/2166896.2166911
31. U. Thayasivam, P. Doshi, Optima results for OAEI 2011, in *Proc. of 6th OM Workshop*, Citeseer, (2011), 204–211.
32. S. Hertling, Hertuda results for OEAI 2012, *Ontology Matching*, 141.
33. M. C. Silva, D. Faria, C. Pesquita, Extending agreementmakerlight to perform holistic ontology matching, in *The Semantic Web: ESWC 2022 Satellite Events*, Springer, (2022), 31–35. https://doi.org/10.1007/978-3-031-11609-4_6
34. K. Janani, S. Mohanrasu, C. P. Lim, B. Manavalan, R. Rakkiyappan, Ensemble feature selection using Bonferroni, OWA and induced OWA aggregation operators, *Appl. Soft Comput.*, **143** (2023), 110431. https://doi.org/10.1016/j.asoc.2023.110431
35. O. Moroz, V. Stepashko, New two-parametric mutation operator for inductive modelling using combinatorial-genetic algorithm, in *2022 12th International Conference on Advanced Computer Information Technologies*, IEEE, (2022), 76–79. https://doi.org/10.1109/ACIT54803.2022.9913199
36. F. Giunchiglia, A. Autayeu, J. Pane, S-match: An open source framework for matching lightweight ontologies, *Semant. Web*, **3** (2012), 307–317. https://doi.org/10.3233/SW-2011-0036
37. J. R. Gomes, A. L. Gançarski, P. R. Henriques, Omt, a web-based tool for ontology matching, in *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*, 2022. https://doi.org/10.4230/OASIcs.SLATE.2022.8
38. X. Liu, Q. Tong, X. Liu, Z. Qin, Ontology matching: State of the art, future challenges, and thinking based on utilized information, *IEEE Access*, **9** (2021), 91235–91243. https://doi.org/10.1109/ACCESS.2021.3057081
39. N. Mahmoud, H. M. Abdlkader, Enhanced ontology matching for big data integration, *J. Phys.: Conf. Ser.*, **1447** (2020), 012028. https://doi.org/10.1088/1742-6596/1447/1/012028

AIMS Press