Research Article

Xingsi Xue*, Miao Ye, and Qifeng Nian

Matching ontologies with kernel principle component analysis and evolutionary algorithm

https://doi.org/10.1515/dema-2024-0021
received November 14, 2023; accepted April 15, 2024

Abstract: Ontology serves as a structured knowledge representation that models domain-specific concepts, properties, and relationships. Ontology matching (OM) aims to identify similar entities across distinct ontologies, which is essential for enabling communication between them. At the heart of OM lies the similarity feature (SF), which measures the likeness of entities from different perspectives. Due to the intricate nature of entity diversity, no single SF can be universally effective in heterogeneous scenarios, which underscores the urgency to construct an SF with high discriminative power. However, the intricate interactions among SFs make the selection and combination of SFs an open challenge. To address this issue, this work proposes a novel kernel principle component analysis and evolutionary algorithm (EA) to automatically construct SF for OM. First, a two-stage framework is designed to optimize SF selection and combination, ensuring holistic SF construction. Second, a cosine similarity-driven kPCA is presented to capture intricate SF relationships, offering precise SF selection. Finally, to bolster the practical application of EA in the SF combination, a novel evaluation metric is developed to automatically guide the algorithm toward more reliable ontology alignments. In the experiment, our method is compared with the state-of-the-art OM methods in the Benchmark and Conference datasets provided by the ontology alignment evaluation initiative. The experimental results show its effectiveness in producing high-quality ontology alignments across various matching tasks, significantly outperforming the state-of-the-art matching methods.

Keywords: ontology matching, similarity feature construction, kernel principle component analysis, evolutionary algorithm

MSC 2020: 68T20, 90C27, 90C90

1 Introduction

The advancement of the Internet has fostered an era of data explosion, raising the need to structure and make sense of this data deluge. The semantic web (SW) [1] emerges as an extension of the existing web, with the aim of making the data machine readable, thus facilitating data interoperability and knowledge sharing. Ontology is a kernel technique of SW, which is a formal representation of knowledge in a specific domain that includes entities, concepts, and the relationships between them [2]. Currently, ontologies have been widely adopted in a multitude of fields, such as healthcare [3], e-commerce [4], natural language processing [5], and big data

*** Corresponding author: Xingsi Xue,** Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou, Fujian, China, e-mail: jack8375@gmail.com
Miao Ye: Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin University of Electronic Technology, Guilin, Guangxi, China, e-mail: yemiao@guet.edu.cn
Qifeng Nian: School of Big Data and Artificial Intelligence, Fujian Polytechnic Normal University, Fuqing, Fujian, China, e-mail: nianqf@fpnu.edu.cn

analytics [6]. However, the widespread use of ontologies in various domains has led to a critical problem known as ontology heterogeneity [7], which occurs when different ontologies describe similar or identical concepts in various manners. In particular, these differences can stem from inconsistencies in naming conventions, attribute definitions, and even the structuring of relationships, impeding the seamless data integration and knowledge transfer, thereby reducing the efficiency of information systems and semantic services. Ontology matching (OM) [8] is an effective solution to this problem, which aims to identify semantically similar entities across different ontologies, thereby bridging the semantic gaps and ensuring effective communication between heterogeneous ontologies. Despite considerable progress in the development of OM methodologies, the task of generating accurate matching results in heterogeneous scenarios remains a formidable challenge. The core of the difficulty lies in the intricate web of semantic relationships that characterize individual entities within ontologies [9].

A foundation for addressing OM problem is the similarity feature (SF) [10], which serves to quantify the similarity value between two entities. However, the heterogeneity intrinsic to entities across various domains renders the task of identifying a universally effective SF a Sisyphean endeavor. For example, in a medical ontology, the SF might weigh taxonomic similarities more heavily, while in a financial ontology, attribute similarities might take precedence. This disparate weighting of similarity underscores the necessity for the careful construction and nuanced amalgamation of high-level SFs, a process that correlates directly with the precision of OM outcomes [11]. The construction of such high-level SFs involves a meticulously calibrated two-stage process: the first stage involves the selection of SFs that resonate most closely with the specific requirements of the matching task at hand. The second and equally critical stage is the synthesis of these SFs into an optimal ensemble, a structure that can integrate disparate SFs into a harmonious whole. However, it is the interaction between SFs, often a web of complex and subtle interactions, that presents a formidable challenge, one that precludes simplistic or one-dimensional selection strategies [12].

Principal component analysis (PCA) [13] has long been recognized as a powerful technique to reduce the dimensionality of features. However, its classical application assumes a Gaussian distribution of the data, an assumption that often breaks down in OM where the relationships among SFs are inherently complex and nonlinear. This discrepancy calls for a more versatile approach, such as the kernel PCA (kPCA) [14], which transcends the limitations of traditional PCA by using kernel functions. These functions project the data into a higher-dimensional space, allowing the algorithm to unravel the intricate patterns that linear methods overlook. Nevertheless, the effectiveness of kPCA is highly dependent on the choice of the kernel function, as not all kernels are suitable for all data types. Although many kPCA applications have relied on radial basis function kernels [15] or polynomial kernels [16], these may not always be able to faithfully represent the complexity of SF relationships within ontologies. This inadequacy underscores the need for a discerning selection of kernel functions that can truly capture the essence of the underlying relationships among SFs. Given two ontologies under alignment, the pursuit of more sophisticated kernel functions tailored to their heterogeneity characteristics becomes imperative. This pursuit not only promises a significant boost in the performance of kPCA but also opens doors to new insights and methodologies within the field of OM, leading to more accurate, robust, and semantically aware matching methods.

Evolutionary ensemble optimization (EEO) [17] stands out as a transformative approach that fundamentally changes the way we tackle the combination of SFs. Moving away from conventional reliance on random searches or manual adjustments, EEO harnesses the inherent capabilities of evolutionary algorithms (EAs) [18] for systematic exploration and exploitation. This allows for a more intelligent navigation through the vast potential configurations of SF ensembles, promising configurations that are not just feasible but optimally tailored to the task. However, the application of EEO in the real world encounters a significant bottleneck: the dependency on expert-provided reference alignments for the assessment of intermediate alignment quality [19]. This reliance is a luxury that cannot always be afforded outside academic environments, where the accessibility of such expert-verified alignments is a rarity. The paucity of these alignments in practical settings renders traditional EEO methodologies less feasible, thus curtailing their applicability in everyday OM operations. To bridge this gap between theory and practice, it is imperative to make them independent of expert input, which not only broadens its application scope but also provides a robust, self-sufficient tool for OM tasks.

Inspired by the success of PCA and EEO in SF selection and combination for OM, the overall goal of this article is to further explore their capability by developing a novel kPCA-EA method, which seamlessly integrates a kPCA-based SF selector with an EA-based feature combiner to enable automatic SF construction for OM. The main contributions of this work are as follows:

- To enhance the accuracy of OM results, we design a novel framework for automatic SF construction. This framework employs a two-stage sequential process to optimize both the selection of the SF subset and the combination of selected SFs, thereby achieving a more effective and coherent SF construction.
- To adeptly identify the most relevant SF subset for OM, we develop an innovative cosine similarity-based kPCA. This approach is tailored to capture the intricate relationships that are prevalent among various SFs, ensuring a more nuanced and accurate SF selection.
- To strengthen the real-world applicability of the EA-based SF combination, we propose a new evaluation metric to assess the quality of matching results. This metric can evaluate the quality of individuals, guiding the search direction of EA without the intervention of expert.

The proposed kPCA-EA marks a significant advancement in OM by integrating the capabilities of kPCA and EA, which can effectively address the non-linear complexities in SF relationships. As a critical component of kPCA-EA, kPCA elevates the feature selection (FS) process by projecting the original feature space into a higher-dimensional space using a kernel function. This enables kPCA to discern and prioritize the most relevant and distinguishing SFs, capturing complex semantic interrelations in ontological data. Subsequently, the EA undertakes the task of optimally combining these selected SFs into a cohesive ensemble, ensuring an accurate reflection of semantic nuances between ontologies. The interaction between kPCA and EA within kPCA-EA is the cornerstone of this methodology, where kPCA's sophisticated SF selection process is complemented by EA's dynamic and adaptive combination strategy. This synergy not only makes SFs highly relevant to specific OM tasks, but also ensures their optimal integration, significantly enhancing the overall accuracy of ontology alignments.

The structure of the remainder of this article is as follows. Section 2 outlines the definitions relevant to the OM problem and provides an overview of the existing literature in the field. Section 3 introduces our novel two-stage framework for SF construction. Sections 4 and 5 describe the kPCA-based SF selection and the EA-based SF combination approaches for OM, respectively. Section 6 presents detailed experimental results. Finally, Section 7 offers conclusions and sets the stage for future research endeavors.

2 Background

2.1 Ontology and OM

An *ontology* is formally represented as a 3-tuple $O = \langle \mathcal{C}, \mathcal{P}, \mathcal{I} \rangle$ [8], comprising the sets of concepts ($\mathcal{C}$), properties ($\mathcal{P}$), and instances ($\mathcal{I}$), which correspond to objects in the real world. Their union $\mathcal{E} = \mathcal{C} \cup \mathcal{P} \cup \mathcal{I}$ is called *ontology entities*. Due to the absence of uniform design standards, ontologies often suffer from the heterogeneity problem [12], and *OM* addresses this issue by identifying correspondences between semantically similar entities [8].

The task of matching two ontologies can be framed as a binary classification problem that aims to find a classifier $h : \mathcal{E}_1 \times \mathcal{E}_2 \rightarrow \{0, 1\}$. In this scenario, a *correspondence* is specified by a 5-tuple $\langle id, e_1, e_2, r, c \rangle$ [20], where id is the ID of the correspondence, $r \in \{\equiv, \leq, \perp\}$ indicates the relational context between e_1 and e_2, and c signifies the confidence level of the correspondence. The set of such correspondences is called *ontology alignment*. This alignment can be illustrated as a binary matrix: each row and column represents an entity from the respective ontologies, and matrix entries (i, j) denote whether entities i and j are mapped (1) or not (0). Evaluation metrics such as recall, precision, and f-measure [21] require the comparison of this alignment against a *ground truth* to assess its accuracy.

2.2 SF selection and combination for OM

Due to the challenges posed by the high-dimensional search space and complex interactions between SFs in OM, specialized FS methods are essential [12] to improve the quality of matching results. These methods aim to enhance matching accuracy by removing irrelevant or redundant SFs while retaining semantically significant ones, which can be divided into three categories, i.e., embedding-based, filter-based, and wrapper-based FS methods [22]. Todorov [23] developed an embedding-based FS method to filter irrelevant SFs with support vector machines (SVM) [24]. However, the performance of this approach is constrained by SVM limitations in high-dimensional or varied contexts. Belhadi et al. [25] used the wrapper-based FS method to identify the SF subset, which can improve the quality of the chosen SF. They further expanded their work by introducing a pattern mining solution for OM [26], focusing on frequent ontology patterns to refine alignment without the need for exhaustive SF analysis. Later, Yap and Kim [27] used a filter-based FS method to determine the SF subset. They mainly focus on individual SF's confidence, but did not consider scenarios involving complex SF dependencies. In contrast with the other two FS methods, the filter-based approach distinctly stands out due to its intrinsic evaluation mechanism. It evaluates the relevance of features rooted in the data's inherent properties, eliminating the need for external learning algorithms. This inherent efficiency results in faster processing speeds and improved scalability. Given that OM frequently operates in real-time or on-line contexts, the rapidity and adaptability make it particularly apt for such tasks.

Moreover, how to combine the selected SF is critical to the quality of the matching results [8]. In recent years, there has been a growing focus on employing EAs for SF combination in OM [28]. Among these, Genetics for ontology alignments stands out as a genetic algorithm (GA)-based approach focused on optimizing aggregation weights for SFs [29]. Traditional EAs, however, face challenges such as slow convergence and premature optimization. To mitigate these issues, Acampora et al. proposed a hybrid GA that incorporates local search strategies [30]. Although most EA-based approaches concentrate on parameter optimization, including weight aggregation and threshold setting, they often overlook the practical challenges of domain-specific requirements and time constraints. Martinez-Gil and Chaves-González addressed this gap by introducing a symbolic regression-based method using genetic programming (GP) to optimize SF ensemble structures [31], which leverages GP's genotype to explore and select optimal ensemble structures.

Despite significant progress in SF selection and combination for OM, existing solutions still suffer from critical drawbacks that compromise their efficacy and broad utility. First, these methods concentrate mainly on SF selection or SF combination, neglecting the other. This one-sided focus inevitably undermines the overall accuracy of the resulting ontology alignments. Second, current FS approaches do not adequately account for the complex, non-linear interrelations between various SFs in OM scenarios, lowering the quality of the selected feature set. Finally, the existing EA-based SF combination methods lack a comprehensive evaluation mechanism, which hinders their suitability for deployment in real-world applications. To overcome these drawbacks, we introduce a novel kPCA-EA framework that automatically selects and combines SFs to enhance the accuracy of OM results.

3 New framework of two-stage SF construction

Figure 1 illustrates the two-stage SF construction framework, which consists of the SF selection and SF combination stages. In the initial pre-processing, each SF is used to calculate a similarity matrix, whose rows and columns denote the entities from two ontologies and individual elements signify the SF values of two corresponding entities. These matrices are then transformed into similarity vectors by sequentially concatenating rows, which form a vital input for the subsequent stages of SF selection. During the SF selection stage, the similarity vectors are used to construct the kernel matrix, allowing the determination of an optimized SF subset based on its eigenvalues. Subsequently, the SF combination stage employs EA to automatically fine-tune the aggregation weights for the chosen SFs, determining the high-quality alignment.

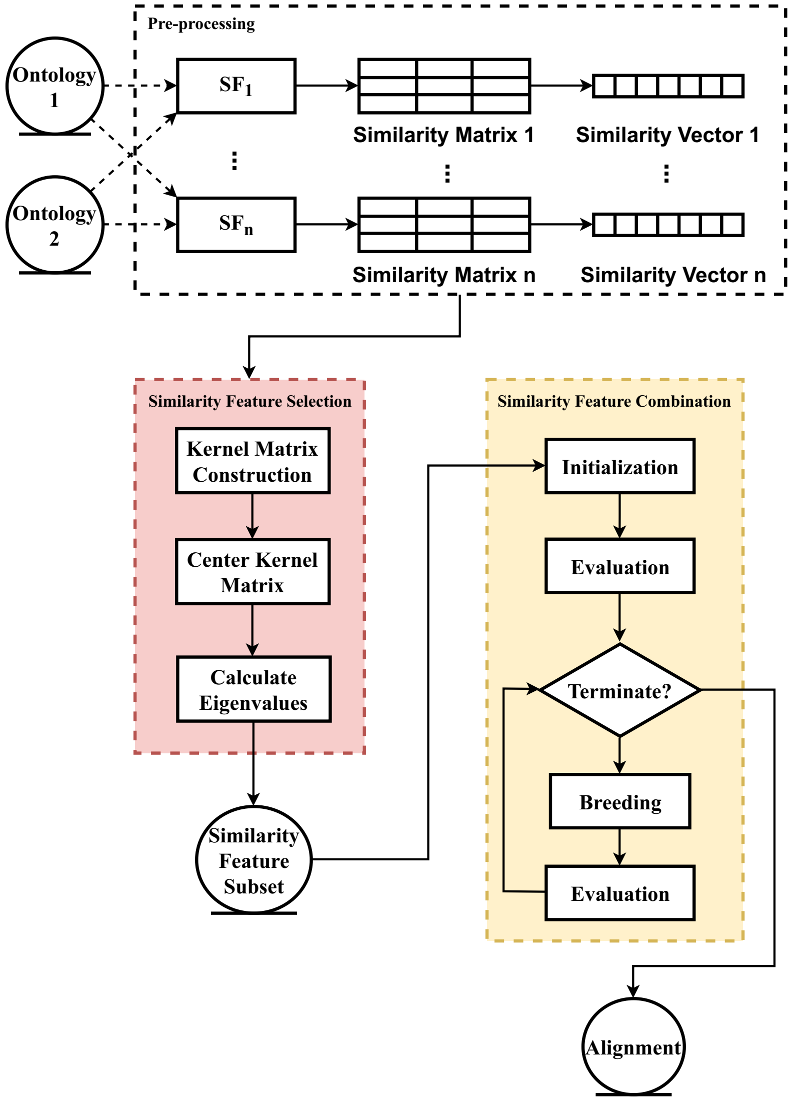

Figure 1: Framework of automatic similarity measure combination.

The proposed two-stage SF construction framework offers a comprehensive and adaptable approach to SF construction for OM. First, it introduces an automatic SF selection process through the construction and centering of a kernel matrix, which eliminates the need for learning process, thereby improving the efficiency of the SF selection. Second, using EA in the SF combination stage, the framework automatically optimizes the aggregation weights for the selected SFs without the requirement for specialized domain expertise, thus contributing to its real-world applicability and scalability.

4 kPCA for SF selection

kPCA has emerged as a leading technique in the refinement of SF selection, which is adept at capturing complex, non-linear relationships among SFs, thereby improving the accuracy of OM outcomes [32]. To determine the relevant SFs, it is important to note that the orientation of their corresponding similarity vectors can be more informative than the magnitudes, where:

- **Orientation**: Refers to the direction in which a vector points in a multi-dimensional space. It captures the relative relationships of its components, irrespective of their absolute values.
- **Magnitude**: Represents the length or size of a vector, corresponding to the overall strength of the similarity.

Although various SFs may yield disparate magnitudes of similarity for identical entity pairs, it is the orientation, i.e., the underlying pattern or trend spanning the entity pairs, that illuminates the true relevance or efficacy of a specific SF. For example, consider two ontologies, O_A and O_B, each comprising three entities: A_1, A_2, A_3 and B_1, B_2, B_3. Let SF$_1$ and SF$_2$ be two SFs used to measure the similarity between these entities. The resulting similarity vectors for SF$_1$ and SF$_2$ are given by: $\mathrm{SF}_1 : \{(A_1 - B_1 : 0.9), (A_2 - B_2 : 0.7), (A_3 - B_3 : 0.8)\}$, and $\mathrm{SF}_2 : \{(A_1 - B_1 : 0.45), (A_2 - B_2 : 0.35), (A_3 - B_3 : 0.4)\}$, respectively. Assuming prior knowledge that A_1 is very similar to B_1, A_2 to B_2 is somewhat similar, and A_3 to B_3 is moderately similar, it becomes evident that despite SF$_2$ producing lower similarity values (lesser magnitude), its orientation aligns more with this knowledge. This underscores that the orientation of SF$_2$ offers a more accurate representation of entity relationships than SF$_1$, emphasizing its potential relevance in OM.

In the realm of OM, selecting the appropriate kernel function for kPCA is pivotal, which should be guided by the domain's unique characteristics and requirements, particularly in discerning relationships between entities. The cosine similarity kernel is chosen for its exceptional capability to prioritize vector orientation over magnitude, aligning well with the demands of SF selection in OM. This kernel emphasizes the directional alignment of vectors rather than their length, crucial in scenarios where entities, despite varying similarity magnitudes, share the consistent relational patterns. By focusing on angular distances, the cosine similarity kernel adeptly captures the relative direction of SF vectors, offering a more accurate reflection of entity relationships. To be specific, given two similarity vectors SV$_1$ and SV$_2$, their cosine similarity $\mathrm{sim}_{\cos}(\cdot)$ is defined as

$$\mathrm{sim}_{\cos}(\mathrm{SV}_1, \mathrm{SV}_2) = \frac{\mathrm{SV}_1 \cdot \mathrm{SV}_2}{||\mathrm{SV}_1|| \times ||\mathrm{SV}_2||}, \tag{1}$$

where $\cdot$ denotes the dot product and || denotes the Euclidean norm of the vector. On this basis, the detailed process of cosine similarity-based kPCA involves several steps:

- **Kernel matrix construction:** Given a set of similarity vectors, construct the kernel matrix K using the cosine similarity between each pair of vectors.
- **Centering the matrix:** The kernel matrix is then centered using the formula:

  $$K_{\text{centered}} = K - 1_N K - K 1_N + 1_N K 1_N, \tag{2}$$

  where 1_N is an $N \times N$ matrix with all entries as $\frac{1}{N}$.
- **Calculating eigenvalues:** Eigenvalues of the centered kernel matrix are calculated to determine the principal components that capture the most variance in the data. These components represent the most relevant SFs.

The kernel matrix captures the nuanced relationships among the SFs, addressing the limitations of linear methods. The centering of the matrix around the origin isolates relevant SF variations, enhancing discrimination. Finally, eigenvalue calculations pinpoint significant SFs, filtering out noise. This integrated approach ensures accurate and robust OM by effectively addressing the intricate relationships between various SFs. Unlike standard linear methods, the kernel matrix in our methodology leverages the cosine similarity to highlight the angular relationships between SFs, providing a more nuanced understanding of their interactions. By centering this kernel matrix around the origin, we effectively isolate and accentuate the variations among SFs that are most relevant for discrimination, ensuring that only the significant variations contribute to the matching process. The subsequent step of calculating eigenvalues from this centered matrix serves to identify and prioritize the SFs that capture the most variance, thereby filtering out less informative or noisy features. This process not only distinguishes our approach from traditional kPCA applications but also optimizes the selection and combination of SFs, ensuring a more accurate and robust framework for OM by directly addressing the intricacies of SF relationships.

Here is an example of cosine similarity-based kPCA. Assume we have three similarity vectors: $SV_1 = [0.1, 0.2, 0.3]$, $SV_2 = [0.4, 0.3, 0.1]$, and $SV_3 = [0.2, 0.1, 0.4]$. First, the kernel matrix K is constructed using cosine similarity between each pair of vectors, resulting in $K = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & 0.7 \\ 0.8 & 0.7 & 1 \end{pmatrix}$. Next, to center this kernel matrix, we use the formula $K_{\text{centered}} = K - 1_N K - K 1_N + 1_N K 1_N$, where 1_N is a 3×3 matrix filled with $\frac{1}{3}$. The centered matrix is obtained as $K_{\text{centered}} = \begin{pmatrix} 0.1 & 0.0 & -0.1 \\ 0.0 & 0.1 & -0.1 \\ -0.1 & -0.1 & 0.2 \end{pmatrix}$. Finally, the eigenvalues of K_{centered} are calculated, yielding $\lambda_1 = 0.35$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.05$. These eigenvalues identify the most pertinent SFs for OM by capturing the maximum variance in the data. Accordingly, we can select the SFs corresponding to the two highest-ranking eigenvalues as our final output.

5 Evolutionary algorithm for SF combination

5.1 SF combination problem

Given an SF set $\mathcal{SF} = \{SF_1, SF_2, \ldots, SF_n\}$, let M_i be the similarity matrix corresponding to SF_i for $i = 1, \ldots, n$, the SF combination problem is to find optimal aggregation weights $\mathcal{W} = \{w_1, w_2, \ldots, w_n\}$ and a threshold $T \in [0, 1]$ that will maximize the alignment quality. The process involves the following steps:

- Aggregate the similarity matrices M_i using the weighted sum strategy to form a new aggregated similarity matrix M_{agg}:

$$M_{\text{agg}} = \sum_{i=1}^{n} w_i \cdot M_i. \tag{3}$$

- Convert M_{agg} into a binary matrix M_{bin} using the threshold T:

$$M_{\text{bin}}(i, j) = \begin{cases} 1, & \text{if } M_{\text{agg}}(i, j) > T, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

The objective is to determine $\mathcal{W}$ and T that maximize the quality of the alignment corresponding to M_{bin}.

5.2 Algorithm overview

EA is a global optimization algorithm renowned for its versatility and robustness, and its inherent capability to explore complex solution spaces makes it particularly suitable for the linear regression problem to combine SFs [28]. In this work, we empirically use the Gray code [33], a well-established binary encoding scheme, to represent both the aggregating weights and a threshold within each individual. Algorithm 1 outlines the pseudo-code of EA for SF combination. Initially, the population $\mathcal{P}$ is initialized and evaluated, and the elite individual $\text{indiv}_{\text{elite}}$ is determined by selecting the individual exhibiting the highest fitness value. In each generation, single-point crossover [34] and bit mutation [35] are used to create a new population $\mathcal{P}''$, and the next generation's population is determined by roulette wheel selection [36]. At the end of each generation, $\text{indiv}_{\text{elite}}$ is updated, which replaces the individual with the lowest fitness values. The algorithm terminates when reaching the maximum generation MaxGen, and outputs $\text{indiv}_{\text{elite}}$ as the final result.

The incorporation of EA in our model uniquely leverages its global optimization prowess, specifically tailored to the nuanced demands of SF combination in OM. Using Gray code for binary encoding, our approach distinctively captures the granularity of aggregating weights and threshold nuances within the evolutionary process. This methodological innovation facilitates a precision-oriented optimization, enabling the EA to adeptly balance and refine the ensemble of SFs. Our model's effectiveness is further amplified through the strategic deployment of evolutionary operations, which can enhance both the diversity and quality of SF combinations. This leads to an optimized aggregation that is not just theoretically robust but practically superior, as evidenced by the enhanced matching accuracy.

Algorithm 1. Evolutionary algorithm

```
Input: Maximum generation MaxGen, Population size $size_p$, Crossover rate $rate_c$, Mutation rate $rate_m$.
Output: Elite individual $indiv_{elite}$.
Initialize the population $\mathcal{P}$;
evaluate($\mathcal{P}$);
Initialize $indiv_{elite}$;
generation gen = 0;
while gen < MaxGen do
    Offspring Population $\mathcal{P}' = crossover(\mathcal{P}, rate_c)$;
    Offspring Population $\mathcal{P}'' = mutation(\mathcal{P}', rate_m)$;
    evaluate($\mathcal{P}''$);
    $\mathcal{P} = selection(\mathcal{P} \cup \mathcal{P}'')$;
    updateElite();
    saveElite();
    gen = gen + 1;
end while
return $indiv_{elite}$;
```

5.3 New fitness function

Conventional evaluation of alignment quality relies on traditional metrics such as recall, precision, and f-measure [37]. However, these metrics require the availability of expert-validated groundtruth-matching results, which are unavailable in practice. To overcome this limitation, we introduce a more flexible evaluation framework grounded in three novel metrics that serve as approximations of these conventional measures. In particular, we assume that all entity pairs have *a priori* probability of constituting a valid mapping, and to formalize this assumption, we introduce two virtual SFs $SF_\top$ and $SF_\bot$. $SF_\top$ unconditionally maps all entity pairs, whereas $SF_\bot$ abstains from mapping any. Given these premises, we define the probability $P(\delta_i)$ of a particular entity mapping δ_i being a true-positive correspondence as follows:

$$P(\delta_i) = \frac{\sum_{k=1,\ldots,\,\mathrm{num}_{\mathrm{SF}},\top,\bot} \mathrm{conf}_k \times \mathrm{SF}_k(\delta_i)}{\mathrm{num}_{\mathrm{SF}} + 2}, \tag{5}$$

where $\mathrm{num}_{\mathrm{SF}}$ represents the total number of SFs used, and $\mathrm{SF}_k(\delta_i)$ equals 1 if the k-th SF designates δ_i as positive based on a threshold determined by the EA, and 0 otherwise. In particular, conf_k is the confidence value of SF_k, which is estimated by analyzing the number of logic conflict mappings within its corresponding alignment. Given an SF SF_k and its corresponding alignment $\mathcal{A}_{\mathrm{SF}}$, an inconsistent subset of alignment $\mathcal{A}_{\mathrm{inconsist}}$ can be determined in the following steps:

- Sort the entity correspondences in $\mathcal{A}_{\mathrm{SF}}$ in descending order of their SF values;
- Add the correspondences with SF values one by one into a temp alignment subset $\mathcal{A}'$, where any correspondences being added do not violate the locality consistency principle [38] with the ones already in $\mathcal{A}'$;
- Add any mappings that conflict with the ones in $\mathcal{A}'$ in the second step to $\mathcal{A}_{\mathrm{inconsist}}$.

On this basis, the confidence value can be defined as the reciprocal of the number of conflicting entity mappings in the corresponding ontology alignment $\mathcal{A}_{\text{inconsist}}$, normalized to the range [0,1]:

$$\text{conf}_k = 1 - \frac{|\mathcal{A}_{\text{inconsist}}|}{|\mathcal{A}|}, \tag{6}$$

where $|\cdot|$ is the number of elements in a set.

On this basis, we define precision as the probability that an entity pair δ_i is true-positive:

$$\text{p}(\delta_i) = \frac{\sum_{i=1,\ldots,|\mathcal{A}|} P(\delta_i)\text{SF}_k(\delta_i)}{\sum_{i=1,\ldots,|\mathcal{A}|} \text{SF}_k(\delta_i)}, \tag{7}$$

where $|\mathcal{A}|$ is the number of entity mappings in the alignment $\mathcal{A}$. Similarly, we define recall as the probability for an entity pair to be judged as true-positive:

$$\text{r}(\delta_i) = \frac{\sum_{i=1,\ldots,|\mathcal{A}|} P(\delta_i)\text{SF}_k(\delta_i)}{\sum_{i=1,\ldots,|\mathcal{A}|} P(\delta_i)}. \tag{8}$$

The f-measure is defined as the harmony mean of pre and rec:

$$f = \frac{2 \times p \times r}{p + r}. \tag{9}$$

Our proposed evaluation metrics provide an effective alternative to traditional metrics such as precision, recall, and f-measure, tailored for EA-based SF combination. At the heart is a consensus model, using an ensemble of SFs to compute $P(\delta_i)$ for a true-positive mapping δ_i. Distinguishing our approach are two “virtual” SFs that act as statistical controls, preventing convergence to misleading values. Furthermore, we incorporate confidence values for each SF, enhancing the accuracy of aggregation by weighting SFs based on their logical consistency. This approach reduces the sensitivity to individual SF performance, resulting in a refined evaluative metric for ontology alignments.

6 Experimental study

6.1 Experiment design

To assess the efficacy of the kPCA-EA approach, we used the Benchmark and Conference datasets provided by the Ontology Alignment Evaluation Initiative (OAEI)[1]. The OAEI Benchmark dataset[2] is meticulously designed for a broad spectrum of heterogeneous OM tasks, offering test cases that vary in complexity, size, and structure:

- Test cases 101–104 encompass nearly identical ontology pairs, posing simpler matching challenges.
- Test cases 201–210 include ontologies with lexical heterogeneities that require advanced matching approaches.
- Test cases 221–247 present ontologies with linguistic variations, adding layers of complexity to matching tasks.
- Test cases 248–262 are the most challenging, consisting of ontologies with pronounced structural heterogeneities.

The OAEI Conference dataset[3] comprises seven ontologies related to conference organization, each with varying class counts and structural intricacies, ekaw (74 classes), sigkdd (49 classes), iasted (140 classes),

[1] https://oaei.ontologymatching.org/2022/conference/index.html
[2] https://oaei.ontologymatching.org/2016/benchmarks/index.html
[3] https://oaei.ontologymatching.org/2022/conference/index.html

conf (62 classes), confOf (38 classes), cmt (36 classes), and edas (104 classes), providing a rigorous testing ground to ascertain the kPCA-EA's adeptness at reconciling real-world ontological variances.

The kPCA-EA configuration for SF selection is as follows:

- Syntax-based SF candidates: Levenshtein distance [39], Jaro distance [40], N-Gram [41], Dice coefficient-based distance [42], and SMOA distance [43];
- Linguistic-based SF candidates: Wu & Palmer distance [44], Path-based distance [45], Leakcock & Chodorow distance [46], and Resnik distance [47];
- Structure-based SF candidates: Similarity flood (SF) distance [48], Neighbor-concept-based distance [49], Instance-based distance [30], and RDF2Vec [50];
- The SF subset siz: 4.

These SF candidates span three major SF categories within the OM field, i.e., syntax-based, linguistic-based, and structure-based [28], capturing similarities between entities from a wide range of perspectives. Furthermore, the EA configuration for the SF combination is empirically optimized to ensure the highest mean quality of matching results across all evaluated cases:

- Population size $size_p$: 60;
- Maximum generation MaxGen: 2000;
- Crossover rate $rate_c$: 0.6;
- Mutation rate $rate_m$: 0.05.

6.2 Sensitivity analysis

The SF subset size is a critical determinant of the efficacy of the kPCA-EA method, significantly affecting both the precision of OM and the computational efficiency. An overly extensive SF subset can encumber the SF combination stage, leading to a prolonged run time that detracts from the method's efficiency. Conversely, a too narrow SF subset may fail to capture essential interactions among features, inadequately addressing the nuanced heterogeneities between ontological entities. Thus, identifying an optimal balance in the size of the SF subset is crucial to maximize the kPCA-EA's capabilities, enabling a synergistic interplay among SFs to effectively navigate and reconcile entity discrepancies.

Figure 2 presents a compelling trade-off between the f-measure and the run time against the size of SF subset. With the size of the subset increasing from 2 to 6, a pronounced increase in run time is observed, indicating that larger subsets exert greater computational demand. In parallel, the f-measure, which serves as a barometer for subset accuracy, ascends with increasing subset size, reaching its zenith at a subset size of 4. Advancing beyond this optimal size, the f-measure stabilizes or experiences a marginal decline, suggesting that an excessively large SF subset may not only fail to elevate performance but could also introduce conflicts among SFs, thus undermining efficacy. Consequently, a subset size of 4 is pinpointed as the ideal configuration, striking a harmonious balance between a commendable f-measure and a manageable computational duration, thereby endorsing both the efficiency and the precision of the OM endeavor.

6.3 Results and discussions

In the experiment, kPCA-EA was compared against OAEI's participants, with results sourced directly from the official OAEI website.[4] The selection of these comparative methods was strategically made to cover a wide range of OM approaches, from basic string-based techniques to sophisticated machine learning algorithms, aiming for a thorough benchmark against leading-edge solutions. These methods were specifically chosen for their proven excellence in navigating the multifaceted challenges of OM as outlined by OAEI, including dealing

[4] http://oaei.ontologymatching.org

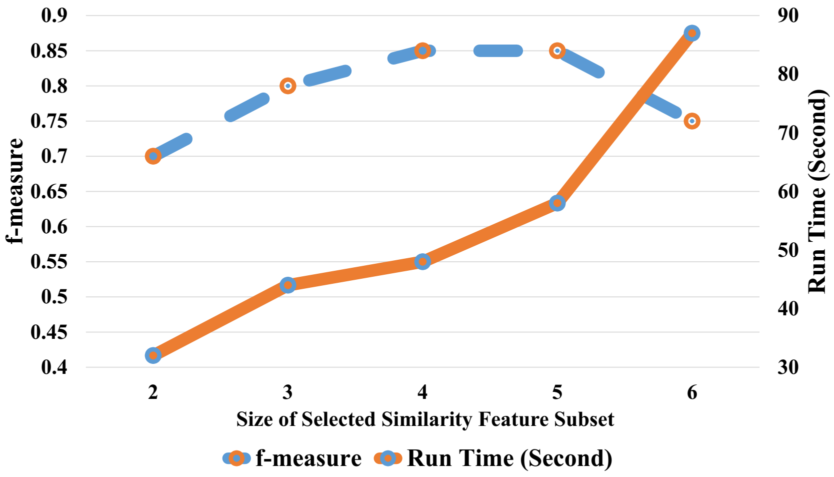

Figure 2: Sensitivity testing on SF subset size.

with structural heterogeneity and lexical variances. This diverse set provides a solid foundation for evaluating the adaptability and efficiency of kPCA-EA. Furthermore, their computational efficiency was a key factor in our selection, establishing a standard against which kPCA-EA's ability to achieve efficient, high-quality ontology alignments could be measured. Our comparative analysis, grounded in the transparent and rigorous evaluation criteria of OAEI, not only demonstrates kPCA-EA's enhanced accuracy in terms of precision, recall, and f-measure but also underscores its suitability for a broad spectrum of real-world OM challenges, thereby underscoring the innovative impact of our approach within the domain.

Table 1 delineates the f-measure scores achieved by kPCA-EA in comparison with its counterparts for all test cases examined. To statistically substantiate the comparative performance, we used the Wilcoxon rank sum test [51] at a significance threshold of 0.05. In the table, the symbols "(+)/(–)/(=)" appended to the outcomes of each competing OM method signify whether kPCA-EA outperformed, underperformed, or achieved similar results to those methods, in a statistically significant manner. In both Benchmark and Conference datasets, kPCA-EA consistently achieves higher precision, f-measure, and recall, highlighting its superior matching accuracy. To further validate the method's advantages, a comparison with non-EAs was conducted. The results, sourced from 30 independent runs, show that kPCA-EA maintains a perfect precision score of 1.00, outperforming competitors in f-measure and recall with minimal standard deviation. These findings indicate kPCA-EA's robustness and reliability in diverse ontological landscapes, making it a versatile tool for OM. The comprehensive performance evaluation also sheds light on the unique benefits of using an EA-based approach in OM. kPCA-EA's superior f-measure and recall scores suggest that the EA component significantly contributes to the method's ability to effectively combine and optimize SFs. In contrast, non-EAs, while effective in certain scenarios, may not provide the same level of adaptability and optimization capabilities as EA-based methods. This comparison not only highlights kPCA-EA's strengths but also its potential limitations, offering insights into specific contexts where traditional or alternative approaches might be more suitable. Overall, this balanced comparison provides a clearer understanding of kPCA-EA's applicability and showcases its potential in a variety of complex real-world OM scenarios.

The experimental results offer compelling evidence supporting the claim that the success of kPCA-EA is largely due to the novel framework we designed for automatic SF construction. The data showcase kPCA-EA's superiority over an array of OAEI competitors, with it consistently achieving higher precision, f-measure, and recall, which are crucial for OM as they collectively measure how accurately and completely the system can identify true matches between entities across ontologies. The framework we introduced plays a critical role in achieving this level of performance due to several factors:

- Two-stage sequential process: Our framework breaks down the SF construction into two strategic stages: selection and combination. By treating these stages sequentially, the framework ensures that each stage is

Table 1: Comparisons among OAEI's participants and kPCA-EA (mean and standard deviation of 30 independent runs)

	Matcher	Precision	f-measure	Recall	Average run time (s)
Benchmark	AML	1.00 (=)	0.57 (+)	0.40 (+)	148 (+)
	CroMatcher	0.95 (+)	0.88 (+)	0.82 (+)	1,100 (+)
	Lily	0.97 (+)	0.90 (+)	0.83 (+)	2,211 (+)
	ATMatcher	0.69 (+)	0.64 (+)	0.51 (+)	127 (+)
	GMap	0.93 (+)	0.68 (+)	0.53 (+)	304 (+)
	YAM++	0.96 (+)	0.89 (+)	0.82 (+)	702 (+)
	RiMOM	0.59 (+)	0.58 (+)	0.58 (+)	105 (+)
	LogMap	1.00 (=)	0.64 (+)	0.47 (+)	194 (+)
	LogMapLt	0.95 (+)	0.66 (+)	0.50 (+)	164 (+)
	Wikimatch	0.97 (+)	0.67 (+)	0.52 (+)	1,845 (+)
	Mamba	0.78 (+)	0.56 (+)	0.44 (+)	3,244 (+)
	XMap	1.00 (=)	0.57 (+)	0.40 (+)	123 (+)
	AOTL	0.89 (+)	0.65 (+)	0.53 (+)	741 (+)
	kPCA-EA	**1.00 (0.01)**	**0.92 (0.01)**	**0.85 (0.01)**	**103 (2)**
Conference	AML	0.82 (+)	0.68 (+)	0.41 (+)	249 (+)
	CroMatcher	0.82 (+)	0.70 (+)	0.44 (+)	732 (+)
	Lily	0.66 (+)	0.44 (+)	0.19 (+)	507 (+)
	ATMatcher	0.69 (+)	0.64 (+)	0.51 (+)	309 (+)
	GMap	0.75 (+)	0.70 (+)	0.55 (+)	672 (+)
	YAM++	0.83 (+)	0.67 (+)	0.38 (+)	480 (+)
	RiMOM	0.83 (+)	0.69 (+)	0.41 (+)	440 (+)
	LogMap	0.76 (+)	0.71 (+)	0.56 (+)	403 (+)
	LogMapLt	0.68 (+)	0.62 (+)	0.47 (+)	320 (+)
	Wikimatch	0.37 (+)	0.20 (+)	0.07 (+)	2,765 (+)
	Mamba	0.79 (+)	0.70 (+)	0.48 (+)	2,174 (+)
	XMap	0.76 (+)	0.65 (+)	0.41 (+)	218 (+)
	AOTL	0.09 (+)	0.11 (+)	0.60 (+)	646 (+)
	kPCA-EA	0.85 (0.01)	0.78 (0.01)	0.71 (0.01)	204 (5)

optimized independently before moving on to the next, allowing for a more refined and targeted approach to SF construction;

- Optimization of SF subset selection: The first stage involves selecting the most pertinent SF subset from a possibly vast feature space. The results demonstrate that kPCA-EA can discern the most informative features, which is evidenced by the high precision scores–signifying that when kPCA-EA predicts a match, it is almost always correct;
- Effective combination of selected SFs: The second stage focuses on the combination of the selected SFs. The superior f-measure and recall scores indicate that kPCA-EA is not only accurate, but also consistently identifies a high proportion of true matches (recall) and balances this with precision (f-measure). This suggests that the SFs are combined in such a way that they complement each other, leading to a more holistic representation of entity similarity;
- Consistency and stability: The minimal standard deviations across multiple runs emphasize the reliability and reproducibility of the framework. This consistency is critical in OM, where varying results can drastically affect the system's trustworthiness;
- Adaptability across ontological domains: The method's robust performance across both Benchmark and Conference datasets indicates its adaptability. This adaptability can be attributed to the framework's ability to handle the unique semantic complexities inherent in different ontologies, optimizing SFs accordingly.

In addition, Table 1 also presents a detailed comparison of the average run time for kPCA-EA and other OAEI participants over 30 independent runs. The experimental results reveal that kPCA-EA not only demonstrates superior performance in terms of precision, f-measure, and recall but also exhibits remarkable efficiency, with significantly lower average run time in both Benchmark and Conference datasets compared to the competing methods. Specifically, in the Benchmark dataset, kPCA-EA's average run time stands at 103 s with a minimal standard deviation of 2 s showcasing its capability to deliver high-quality OM results more rapidly than several other participants. Notably, methods such as Lily, Wikimatch, and Mamba report substantially higher run times, ranging from 2,211 to 3,244 s, which underscores the computational efficiency of kPCA-EA in handling complex OM tasks without compromising the accuracy or the quality of the results. Similarly, in the Conference category, kPCA-EA maintains its computational advantage with an average run time of 204 s and a standard deviation of 5 s, further highlighting its efficiency. Compared to other methods such as Wikimatch and Mamba, which exhibit run times of 2,765 and 2,174 s, respectively, kPCA-EA presents a more time-efficient solution without sacrificing performance metrics.

The experimental results demonstrate the superiority of the kPCA-EA method in the field of OM, showcasing its consistently high precision, f-measure, and recall metrics when compared to a broad spectrum of existing methods. This performance is not limited to specific test cases or competitors but extends across various ontological frameworks, underscoring the method's robustness and adaptability. The inclusion of an EA component significantly enhances kPCA-EA's ability to effectively construct SF, offering a level of adaptability and optimization that traditional and non-EAs struggle to achieve. Moreover, kPCA-EA's computational efficiency further distinguishes it from other methods. It achieves superior matching accuracy and quality at a fraction of the computational cost, making it an efficient solution for complex OM tasks.

6.4 Further analysis

6.4.1 Effectiveness of kPCA-based SF selection

The newly introduced kPCA serves as a pivotal component within the kPCA-EA framework, autonomously ascertaining the optimal SF subset. To validate the effectiveness of this method, kPCA-EA was evaluated alongside seven distinct variations:

- kPCA-EA$_{\text{all}}$: This version of kPCA-EA incorporates all SFs.
- kPCA-EA$_{\text{syn}}$: This variant is tailored to select the SF subset exclusively from syntax-based SFs.
- kPCA-EA$_{\text{lin}}$: In this adaptation, kPCA-EA is inclined toward linguistic-based SFs for selection.
- kPCA-EA$_{\text{str}}$: This kPCA-EA uses kPCA to discern the optimal structure-based SF subset.
- PCA-EA: This variant substitutes kPCA with the classic PCA to determine the SF subset.
- gPCA-EA: In this variant, the kernel function of PCA is replaced by the Gaussian kernel function [52] to determine the SF subset.
- lPCA-EA: This instantiation replaces the cosine similarity kernel function with the Laplacian kernel function [53] to determine the SF subset.

The experimental results encapsulated within Table 2 delineate a compelling narrative of kPCA-EA's excellence over its counterparts in both Benchmark and Conference datasets. In the realm of Benchmark data, kPCA-EA not only attains the highest mean f-measure values but does so with an admirable consistency, as reflected by its notably low standard deviations. This trend of high mean f-measure values continues unabated across the Conference dataset, where kPCA-EA again stands out, underscoring its capability to adeptly navigate the nuanced complexities of varying ontological structures. The statistical symbols accompanying the f-measure values, predominantly marked with "+" next to kPCA-EA's peers, signify the method's statistically significant outperformance. These symbols are not mere annotations but robust affirmations of kPCA-EA's methodological dominance, validated through rigorous statistical analysis. The minimal standard deviations accompanying kPCA-EA's results speak to the method's reliability and the reproducibility of its success, which are pivotal in the field of OM where dependability is paramount.

The experimental results provide robust empirical support for the effectiveness of our innovative cosine similarity-based kPCA within the kPCA-EA framework. The consistent achievement of the highest mean

Table 2: Mean (standard deviation) of f-measure among kPCA-EA$_{all}$, kPCA-EA$_{syn}$, kPCA-EA$_{lin}$, kPCA-EA$_{str}$, PCA-EA, gPCA-EA, lPCA-EA and kPCA-EA over 30 independent runs in all test cases

	Test case	kPCA-EA$_{all}$	kPCA-EA$_{syn}$	kPCA-EA$_{lin}$	kPCA-EA$_{str}$	PCA-EA	gPCA-EA	lPCA-EA	kPCA-EA
Benchmark	101–104	0.92 (0.02) (+)	1.00 (0.02) (=)	1.00 (0.02) (=)	0.93 (0.01) (+)	1.00 (0.01) (=)	1.00 (0.02) (=)	1.00 (0.02) (=)	1.00 (0.01)
	201–210	0.83 (0.01) (+)	0.86 (0.02) (+)	0.79 (0.03) (+)	0.84 (0.01) (+)	0.86 (0.03) (+)	0.84 (0.02) (+)	0.81 (0.01) (+)	0.90 (0.02)
	221–247	0.86 (0.01) (+)	0.82 (0.02) (+)	0.80 (0.03) (+)	0.84 (0.03) (+)	0.87 (0.01) (+)	0.85 (0.02) (+)	0.81 (0.02) (+)	1.00 (0.01)
	248–262	0.63 (0.02) (+)	0.61 (0.01) (+)	0.68 (0.02) (+)	0.65 (0.02) (+)	0.61 (0.03) (+)	0.60 (0.01) (+)	0.63 (0.01) (+)	0.78 (0.01)
Conference	cmt-conf	0.81 (0.01) (+)	0.75 (0.02) (+)	0.71 (0.03) (+)	0.71 (0.01) (+)	0.80 (0.01) (+)	0.79 (0.02) (+)	0.83 (0.01) (+)	0.85 (0.01)
	cmt-confOf	0.80 (0.03) (+)	0.78 (0.03) (+)	0.75 (0.02) (+)	0.74 (0.03) (+)	0.75 (0.02) (+)	0.75 (0.01) (+)	0.82 (0.00) (=)	0.82 (0.01)
	cmt-edas	0.82 (0.02) (=)	0.72 (0.02) (+)	0.71 (0.03) (+)	0.77 (0.02) (+)	0.76 (0.02) (+)	0.71 (0.02) (+)	0.82 (0.01) (=)	0.82 (0.02)
	cmt-ekaw	0.62 (0.01) (+)	0.51 (0.01) (+)	0.53 (0.01) (+)	0.58 (0.01) (+)	0.72 (0.01) (+)	0.52 (0.01) (+)	0.73 (0.01) (=)	0.73 (0.01)
	cmt-iasted	0.82 (0.01) (=)	0.77 (0.02) (+)	0.72 (0.00) (+)	0.77 (0.01) (+)	0.72 (0.01) (+)	0.72 (0.02) (+)	0.78 (0.00) (+)	0.82 (0.01)
	cmt-sigkdd	0.85 (0.02) (=)	0.72 (0.02) (+)	0.74 (0.00) (+)	0.72 (0.02) (+)	0.74 (0.02) (+)	0.74 (0.02) (+)	0.85 (0.00) (=)	0.85 (0.02)
	conf-confOf	0.70 (0.03) (+)	0.67 (0.03) (+)	0.65 (0.01) (+)	0.69 (0.02) (+)	0.67 (0.02) (+)	0.72 (0.01) (=)	0.72 (0.01) (=)	0.72 (0.02)
	conf-edas	0.72 (0.02) (+)	0.62 (0.02) (+)	0.62 (0.02) (+)	0.66 (0.02) (+)	0.68 (0.02) (+)	0.72 (0.02) (+)	0.71 (0.02) (+)	0.75 (0.02)
	conf-ekaw	0.82 (0.01) (+)	0.76 (0.01) (+)	0.72 (0.01) (+)	0.73 (0.01) (+)	0.72 (0.01) (+)	0.72 (0.01) (+)	0.82 (0.02) (+)	0.86 (0.01)
	conf-iasted	0.75 (0.01) (+)	0.69 (0.01) (+)	0.65 (0.02) (+)	0.63 (0.01) (+)	0.61 (0.01) (+)	0.65 (0.01) (+)	0.65 (0.01) (+)	0.73 (0.01)
	conf-sigkdd	0.81 (0.01) (+)	0.76 (0.01) (+)	0.71 (0.03) (+)	0.78 (0.01) (+)	0.79 (0.01) (+)	0.71 (0.01) (+)	0.74 (0.02) (+)	0.84 (0.01)
	confOf-edas	0.67 (0.02) (+)	0.53 (0.02) (+)	0.58 (0.02) (+)	0.52 (0.01) (+)	0.64 (0.02) (+)	0.62 (0.02) (+)	0.69 (0.02) (+)	0.72 (0.01)
	confOf-ekaw	0.74 (0.02) (+)	0.67 (0.03) (+)	0.64 (0.01) (+)	0.68 (0.01) (+)	0.65 (0.01) (+)	0.54 (0.03) (+)	0.65 (0.01) (+)	0.70 (0.01)
	confOf-iasted	0.70 (0.02) (=)	0.61 (0.03) (+)	0.61 (0.03) (+)	0.61 (0.01) (+)	0.61 (0.01) (+)	0.61 (0.01) (+)	0.67 (0.02) (+)	0.70 (0.01)
	confOf-sigkdd	0.81 (0.02) (+)	0.74 (0.02) (+)	0.77 (0.02) (+)	0.71 (0.02) (+)	0.74 (0.02) (+)	0.78 (0.02) (+)	0.78 (0.00) (+)	0.85 (0.02)
	edas-ekaw	0.82 (0.02) (=)	0.75 (0.02) (+)	0.72 (0.00) (+)	0.72 (0.02) (+)	0.76 (0.02) (+)	0.82 (0.02) (=)	0.76 (0.00) (+)	0.82 (0.02)
	edas-iasted	0.84 (0.01) (+)	0.78 (0.01) (+)	0.74 (0.01) (+)	0.74 (0.01) (+)	0.74 (0.01) (+)	0.74 (0.01) (+)	0.78 (0.01) (+)	0.88 (0.01)
	edas-sigkdd	0.67 (0.01) (+)	0.52 (0.01) (+)	0.57 (0.01) (+)	0.55 (0.01) (+)	0.70 (0.01) (+)	0.67 (0.01) (+)	0.70 (0.01) (+)	0.72 (0.01)
	ekaw-iasted	0.61 (0.02) (+)	0.58 (0.02) (+)	0.56 (0.01) (+)	0.56 (0.02) (+)	0.68 (0.02) (+)	0.54 (0.02) (+)	0.68 (0.01) (+)	0.73 (0.02)
	ekaw-sigkdd	0.62 (0.03) (+)	0.57 (0.01) (+)	0.62 (0.02) (+)	0.63 (0.01) (+)	0.52 (0.01) (+)	0.70 (0.01) (=)	0.60 (0.02) (+)	0.70 (0.01)
	iasted-sigkdd	0.68 (0.03) (+)	0.63 (0.01) (+)	0.65 (0.01) (+)	0.60 (0.01) (+)	0.66 (0.01) (+)	0.61 (0.01) (+)	0.66 (0.01) (+)	0.75 (0.01)
	+/−/=	20/0/5	24/0/1	24/0/1	25/0/0	24/0/1	21/0/4	19/0/6	

f-measure values across both Benchmark and Conference datasets is a clear testament to the method's ability to identify the most relevant SF subset for OM, which can be attributed to the following reasons:

- Versatile SF selection: The kPCA-EA, utilizing cosine similarity-based kPCA, adeptly discerns intricate relationships between SFs in various ontological landscapes, including less structured or loosely hierarchical domains. The high f-measure scores in different datasets indicate kPCA-EA's capability to adapt its SF selection to the unique characteristics of each ontology, essential in domains with less defined structures.
- Consistent performance in varied contexts: The low standard deviations in kPCA-EA's f-measure scores demonstrate its consistent and reliable performance, even in complex and non-standardized ontology frameworks. This consistency highlights kPCA-EA's robustness, producing high-quality results across a range of ontological configurations, confirming its effectiveness in various scenarios.
- Nuanced understanding in complex domains: The success of kPCA-EA in various ontological structures suggests its nuanced understanding of SF relationships, crucial for domains lacking clear hierarchies. By focusing on the directionality of SFs through cosine similarity, kPCA-EA ensures the selection of semantically relevant features, leading to more accurate matches in complex and unconventional ontologies.

In conclusion, the experimental results strongly support our claim that kPCA-EA's success stems from its novel framework for automatic SF construction, particularly using cosine similarity-based kPCA. In particular, its versatility in adapting to less structured or hierarchically ambiguous domains showcases the potential for wide-ranging real-world applications, extending its utility beyond traditional OM scenarios.

6.4.2 Effectiveness of approximate evaluation metrics on alignment

The novel approximate metrics for evaluating alignment's quality serve as a pivotal component of kPCA-EA, adeptly guiding the algorithm's search direction autonomously. To demonstrate their efficacy, we compared kPCA-EA against three specialized variants:

- kPCA-EA$_\mathrm{R}$: This embodiment of kPCA-EA supplants the approximate recall with the conventional recall. Drawing a parallel with kPCA-EA$_\mathrm{R}$ allows us to discern the unique advantages offered by the approximate recall.
- kPCA-EA$_\mathrm{P}$: In this variant, the approximate precision integral to kPCA-EA is traded for the established precision. Contrasting this with kPCA-EA$_\mathrm{P}$ accentuates the benefits inherent to our innovative approximate precision.
- kPCA-EA$_\mathrm{F}$: The approximate f-measure in kPCA-EA is substituted with the traditional f-measure. A comparative assessment with kPCA-EA$_\mathrm{F}$ sheds light on the distinct merits of using the approximate f-measure.

The experimental results in Table 3 provide a clear testament to the effectiveness of the approximate metrics employed by kPCA-EA. When evaluated in a comprehensive suite of Benchmark and Conference datasets, kPCA-EA not only matches but also frequently surpasses the performance of its counterparts, kPCA-EA$_\mathrm{R}$, kPCA-EA$_\mathrm{P}$, and kPCA-EA$_\mathrm{F}$, in terms of mean f-measure values. In the Benchmark category, kPCA-EA maintains parity with its counterparts, as indicated by the symbol "=", revealing that the use of approximate metrics does not compromise the quality of the results. In instances where kPCA-EA is marked with a "+", it demonstrates a statistically significant improvement over the traditional metrics, highlighting the benefit of using the approximate approach in those particular scenarios. With respect to the Conference datasets, the strength of kPCA-EA's approximate metrics becomes even more pronounced. Here, the consistent "=" scores across most test cases confirm that the approximate evaluation method holds its ground against the conventional metrics. The occasional "+" outcomes, particularly in comparison with kPCA-EA$_\mathrm{P}$ and kPCA-EA$_\mathrm{R}$, underscore the instances where the approximate metrics not only match but exceed the performance of the traditional metrics, offering a potent alternative for quality alignment evaluation.

The empirical evidence offers a robust validation of the newly proposed evaluation metrics within the kPCA-EA framework. The results across both Benchmark and Conference datasets demonstrate that kPCA-EA achieves equivalent or superior mean f-measure values relative to the variants employing traditional

Table 3: Mean (standard deviation) of f-measure among kPCA-EA$_R$, kPCA-EA$_P$, kPCA-EA$_F$, and kPCA-EA over 30 independent runs in all test cases

	Test case	**kPCA-EA$_R$**	**kPCA-EA$_P$**	**kPCA-EA$_F$**	**kPCA-EA**
Benchmark	101–104	1.00 (0.01) (=)	1.00 (0.02) (=)	1.00 (0.01) (=)	1.00 (0.01)
	201–210	0.86 (0.01) (+)	0.85 (0.02) (+)	0.90 (0.01) (=)	0.90 (0.02)
	221–247	1.00 (0.01) (=)	1.00 (0.01) (=)	1.00 (0.01) (=)	1.00 (0.01)
	248–262	0.78 (0.02) (=)	0.78 (0.01) (=)	0.78 (0.01) (=)	0.78 (0.01)
Conference	cmt-conf	0.85 (0.01) (=)	0.85 (0.02) (=)	0.85 (0.01) (=)	0.85 (0.01)
	cmt-confOf	0.82 (0.03) (=)	0.82 (0.01) (=)	0.82 (0.02) (=)	0.82 (0.01)
	cmt-edas	0.75 (0.02) (+)	0.82 (0.02) (=)	0.82 (0.01) (=)	0.82 (0.02)
	cmt-ekaw	0.73 (0.01) (=)	0.73 (0.01) (=)	0.73 (0.01) (=)	0.73 (0.01)
	cmt-iasted	0.82 (0.01) (=)	0.82 (0.02) (=)	0.82 (0.01) (=)	0.82 (0.01)
	cmt-sigkdd	0.85 (0.02) (=)	0.81 (0.02) (+)	0.85 (0.02) (=)	0.85 (0.02)
	conf-confOf	0.72 (0.03) (=)	0.72 (0.03) (=)	0.72 (0.01) (=)	0.72 (0.02)
	conf-edas	0.75 (0.02) (=)	0.75 (0.02) (=)	0.75 (0.02) (=)	0.75 (0.02)
	conf-ekaw	0.81 (0.01) (+)	0.86 (0.01) (=)	0.86 (0.02) (=)	0.86 (0.01)
	conf-iasted	0.73 (0.01) (=)	0.65 (0.01) (+)	0.73 (0.01) (=)	0.73 (0.01)
	conf-sigkdd	0.84 (0.01) (=)	0.84 (0.01) (=)	0.84 (0.00) (=)	0.84 (0.01)
	confOf-edas	0.72 (0.02) (=)	0.72 (0.02) (=)	0.72 (0.02) (=)	0.72 (0.01)
	confOf-ekaw	0.59 (0.01) (+)	0.70 (0.03) (=)	0.70 (0.01) (=)	0.70 (0.01)
	confOf-iasted	0.70 (0.01) (=)	0.62 (0.01) (+)	0.70 (0.02) (=)	0.70 (0.01)
	confOf-sigkdd	0.85 (0.02) (=)	0.85 (0.02) (=)	0.85 (0.01) (=)	0.85 (0.02)
	edas-ekaw	0.82 (0.02) (=)	0.82 (0.02) (=)	0.82 (0.02) (=)	0.82 (0.02)
	edas-iasted	0.88 (0.01) (=)	0.88 (0.01) (=)	0.88 (0.01) (=)	0.88 (0.01)
	edas-sigkdd	0.72 (0.01) (=)	0.72 (0.01) (=)	0.72 (0.01) (=)	0.72 (0.01)
	ekaw-iasted	0.66 (0.02) (=)	0.66 (0.02) (=)	0.66 (0.01) (=)	0.66 (0.02)
	ekaw-sigkdd	0.55 (0.01) (+)	0.62 (0.01) (=)	0.62 (0.02) (=)	0.62 (0.01)
	iasted-sigkdd	0.75 (0.01) (=)	0.75 (0.01) (=)	0.75 (0.01) (=)	0.75 (0.01)
	+/−/=	5/0/20	4/0/21	0/0/25	

evaluation measures. In the Benchmark set, the new metrics do not detract from the alignment quality, maintaining the integrity of the evaluation while potentially simplifying the computational process. In particular, there are test cases where kPCA-EA delivers statistically significant enhancements over traditional methods, which bolster the claim that our new evaluation metric can, indeed, guide the EA toward more accurate matches. Such improvements are crucial, as they signify the practical viability of the proposed metrics, especially in scenarios where expert knowledge is scarce or difficult to obtain. Within the Conference dataset, the consistent equivalence of kPCA-EA with its traditional metric counterparts across most test cases reinforces the reliability and applicability of the approximate metrics. The occasional superiority is also significant, demonstrating that the approximate metric can not only compete but also excel beyond conventional measures, providing a formidable alternative for evaluating alignment quality.

In summary, the comparative analysis not only corroborates the effectiveness of kPCA-EA's approximate evaluation metrics but also illustrates their potential to autonomously direct the search process toward high-quality ontology alignments. This paves the way for more self-reliant, accurate, and reliable matching systems that can operate effectively even in the absence of expert validation.

7 Conclusion and future work

In this work, a novel kPCA-EA method was developed to automate the SF construction for OM. Different from traditional methodologies, our approach leverages a cosine similarity-based kPCA for SF selection, coupled with an EA incorporating an approximate evaluation metric for optimal SF combination. The performance of the proposed method was evaluated against state-of-the-art OM methods, and our approach showed consistently higher accurate results in various test cases.

In the future, we aim to expand the capabilities of the kPCA-EA framework to address several key areas. First, we plan to enhance the handling of complex correspondences by integrating one-to-many and many-to-many mappings, which will involve developing new algorithms or modifying existing ones within kPCA-EA. This expansion is crucial for addressing more intricate matching scenarios. Second, we intend to integrate natural language processing techniques to incorporate entity descriptions in natural language, providing additional semantic context and improving alignment precision. This will likely involve using advanced NLP methods such as named entity recognition and semantic parsing. Third, we aim to make kPCA-EA more adaptable to ontologies with varying hierarchical complexities, researching methods to assess structural complexity and adjust the process accordingly. Finally, we plan to investigate automated methods for selecting optimal kernel functions for kPCA, potentially using machine learning techniques to learn from past tasks and predict the most effective kernel for new tasks. These enhancements will strengthen kPCA-EA's robustness and versatility, making it a more effective tool for OM across diverse domains.

Acknowledgement: The authors are grateful for the reviewer's valuable comments that improved the manuscript.

Funding information: This work was supported by the National Natural Science Foundation of China (Nos. 62172095 and 62161006), Natural Science Foundation of Fujian Province (No. 2020J01299), Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing (No. CRKL220206), and Scientific Research Project of Fujian University of Technology (No. GY-Z23044).

Author contributions: All authors have accepted responsibility for the entire content of the manuscript and approved its submission.

Conflict of interest: The authors declare no conflict of interest.

Data availability statement: All data generated or analyzed during this study are included in this published article.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, *The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities*, Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web, ACM, New York, United States, 2023, pp. 91–103.

[2] A. Gómez-Pérez and O. Corcho, *Ontology languages for the semantic web*, IEEE Intell. Syst. **17** (2002), no. 1, 54–60.

[3] H. B. Elhadj, F. Sallabi, A. Henaien, L. Chaari, K. Shuaib, and M. Al Thawadi, *Do-care: A dynamic ontology reasoning based healthcare monitoring system*, Future Generation Comput. Syst. **118** (2021), 417–431.

[4] R. V. Karthik and S. Ganapathy, *A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce*, Appl. Soft Comput. **108** (2021), 107396.

[5] M. G. Kersloot, F. J. P. van Putten, A. Abu-Hanna, R. Cornet, and D. L. Arts, *Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies*, J. Biomed. Semantics **11** (2020), 1–21.

[6] A. Castro, V. A. Villagra, P. García, D. Rivera, and D. Toledo, *An ontological-based model to data governance for big data*, IEEE Access **9** (2021), 109943–109959.

[7] M. Mohammed, A. Romli, and R. Mohamed, *Ontology integration by semantic mapping for solving the heterogeneity problem*, International Conference on Information Systems and Intelligent Applications, Springer, Kuala Lumpur, Malaysia, 2022, pp. 93–102.

[8] C. Trojahn, R. Vieira, D. Schmidt, A. Pease, and G. Guizzardi, *Foundational ontologies meet ontology matching: A survey*, Semant. Web **13** (2022), no. 4, 685–704.

[9] I. Osman, S. B. Yahia, and G. Diallo, *Ontology integration: approaches and challenging issues*, Inform. Fusion **71** (2021), 38–63.

[10] X. Xue, J. Guo, M. Ye, and J. Lv, *Similarity feature construction for matching ontologies through adaptively aggregating artificial neural networks*, Mathematics **11** (2023), no. 2, 485.

[11] P. Wang, Y. Hu, S. Bai, and S. Zou, *Matching biomedical ontologies: Construction of matching clues and systematic evaluation of different combinations of matchers*, JMIR Med. Inform. **9** (2021), no. 8, e28212.

[12] P. Shvaiko and J. Euzenat, *Ontology matching: state of the art and future challenges*, IEEE Trans. Knowledge Data Eng. **25** (2011), no. 1, 158–176.

[13] C. Labriiin and F. Urdinez, *Principal component analysis*, R for Political Data Science, Chapman and Hall/CRC, 2020, pp. 375–393.

[14] S. Marukatat, *Tutorial on PCA and approximate PCA and approximate kernel PCA*, Artif. Intell. Rev. **56** (2023), no. 6, 5445–5477.

[15] D. Li, B. Yang, and Y. Zhang, *Dimension-reduction and reconstruction of multi-dimension spatial wind power data based on optimal RBF kernel principal component analysis*, 2020 10th International Conference on Power and Energy Systems (ICPES), IEEE, 2020, pp. 326–332.

[16] H. Sun, G. Lv, J. Mo, X. Lv, G. Du, and Y. Liu, *Application of KPCA combined with SVM in raman spectral discrimination*, Optik **184** (2019), 214–219.

[17] N. O. Nikitin, A. V. Kalyuzhnaya, K. Bochenina, A. A. Kudryashov, A. Uteuov, I. Derevitskii, et al., *Evolutionary ensemble approach for behavioral credit scoring*, Computational Science-ICCS 2018: 18th International Conference, Wuxi, China, June 11-13, 2018 Proceedings, Part III 18, Springer, 2018, pp. 825–831.

[18] N. A. Zolpakar, M. F. Yasak, and S. Pathak, *A review: use of evolutionary algorithm for optimisation of machining parameters*, Int. J. Adv. Manufact. Technol. **115** (2021), 31–47.

[19] Q. Lv, C. Jiang, and H. Li, *Solving ontology meta-matching problem through an evolutionary algorithm with approximate evaluation indicators and adaptive selection pressure*, IEEE Access **9** (2020), 3046–3064.

[20] J. Hao, C. Lei, V. Efthymiou, A. Quamar, F. Özcan, Y. Sun, et al., *Medto: Medical data to ontology matching using hybrid graph neural networks*, Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 2946–2954.

[21] M. Arora, U. Kanjilal, and D. Varshney, *Evaluation of information retrieval: precision and recall*, Int. J. Indian Culture Business Manag. **12** (2016), no. 2, 224–236.

[22] S. P. Maharudra and S. Gandage, *A high-level ensemble feature selection algorithm for mitigating the dimensionality in stress data*, J. Data Acquisit. Process **38** (2023), no. 3, 1064.

[23] K. Todorov, P. Geibel, and K.-U. Kuehnberger, *Extensional ontology matching with variable selection for support vector machines*, 2010 International Conference on Complex, Intelligent and Software Intensive Systems, IEEE, 2010, pp. 962–967.

[24] D.A. Pisner and D. M. Schnyer, *Support vector machine*, Machine Learning, Elsevier, London, United Kingdom, 2020, pp. 101–121.

[25] H. Belhadi, K. Akli-Astouati, Y. Djenouri, J. Chun-Wei Lin, and J. Ming-Tai Wu, *Gfsom: genetic feature selection for ontology matching*, Genetic and Evolutionary Computing: Proceedings of the Twelfth International Conference on Genetic and Evolutionary Computing, December 14–17, Changzhou, Jiangsu, China 12, Springer, 2019, pp. 655–660.

[26] H. Belhadi, K. Akli-Astouati, Y. Djenouri, and J. Chun-Wei Lin, *Exploring pattern mining for solving the ontology matching problem*, New Trends in Databases and Information Systems: ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23, Springer, 2019, pp. 85–93.

[27] C. E. Yap and M. H. Kim, *Instance-based ontology matching with rough set features selection*, 2013 International Conference on IT Convergence and Security (ICITCS), IEEE, 2013, pp. 1–4.

[28] N. Ferranti, S. S. R. Furtado Soares, and J. F. de Souza, *Metaheuristics-based ontology meta-matching approaches*, Expert Syst. Appl. **173** (2021), 114578.

[29] J. Martinez-Gil, J. Montes, E. Alba, and J. F. Aldana-Montes, *Optimizing ontology alignments by using genetic algorithms*, Proceedings of the workshop on nature based reasoning for the semantic web, Springer, Karlsruhe, Germany, 2008, pp. 1–15.

[30] G. Acampora, V. Loia, and A. Vitiello, *Enhancing ontology alignment through a memetic aggregation of similarity measures*, Inform. Sci. **250** (2013), 1–20.

[31] J. Martinez-Gil and J. M. Chaves-González, *A novel method based on symbolic regression for interpretable semantic similarity measurement*, Expert Syst. Appl. **160** (2020), 113663.

[32] S. Bheemireddy, S. S. Durbha, R. L. King, S. K. Amanchi, and N. H. Younan, *An ontology merging tool to facilitate interoperability between coastalsensor networks*, 2009 IEEE International Geoscience and Remote Sensing Symposium, vol. 5, IEEE, 2009, pp. V-367.

[33] K. Zeger and A. Gersho, *Pseudo-gray coding*, IEEE Trans. Commun. **38** (1990), no. 12, 2147–2158.

[34] A. S. Desuky, Y. M. Elbarawy, S. Kausar, A. H. Omar, and S. Hussain, *Single-point crossover and jellyfish optimization for handling imbalanced data classification problem*, IEEE Access **10** (2022), 11730–11749.

[35] A. Rajabi and C. Witt, *Evolutionary algorithms with self-adjusting asymmetric mutation*, International Conference on Parallel Problem Solving from Nature, Springer, Leiden, The Netherlands, 2020, pp. 664–677.

[36] F. Yu, X. Fu, H. Li, and G. Dong, *Improved roulette wheel selection-based genetic algorithm for TSP*, 2016 International Conference on Network and Information Ssystems for Computers (ICNISC), IEEE, 2016, pp. 151–154.

[37] A. Patel and S. Jain, *A partition based framework for large scale ontology matching*, Recent Patents Eng. **14** (2020), no. 3, 488–501.

[38] A. Solimando, E. Jimenez-Ruiz, and G. Guerrini, *Minimizing conservativity violations in ontology alignments: Algorithms and evaluation*, Knowledge Inform. Syst. **51** (2017), no. 3, 775–819.

[39] L. Yujian and L. Bo, *A normalized levenshtein distance metric*, IEEE Trans. Pattern Anal. Machine Intel. **29** (2007), no. 6, 1091–1095.

[40] K. Dreßler and A.-C. Ngonga Ngomo, *On the efficient execution of bounded jaro-winkler distances*, Semantic Web **8** (2017), no. 2, 185–196.

[41] G. Kondrak, *N-gram similarity and distance*, International Symposium on String Processing and Information Retrieval, Springer, 2005, pp. 115–126.

[42] W. Cohen, P. Ravikumar, and S. Fienberg, *A comparison of string metrics for matching names and records*, KDD Workshop on Data Cleaning and Object Consolidation, vol. 3, 2003, pp. 73–78.

[43] G. Stoilos, G. Stamou, and S. Kollias, *A string metric for ontology alignment*, International Semantic Web Conference, Springer, 2005, pp. 624–637.

[44] T. Slimani, B. B. Yaghlane, and K. Mellouli, *A new similarity measure based on edge counting*, Int. J. Comput. Inform. Eng. **2** (2008), no. 11, 3851–3855.

[45] H. Bulskov, R. Knappe, and T. Andreasen, *On measuring similarity for conceptual querying*, Flexible Query Answering Systems: 5th International Conference, FQAS 2002 Copenhagen, Denmark, October 27–29, 2002 Proceedings 5, Springer, 2002, pp. 100–111.

[46] C. Leacock, *Combining local context and wordnet similarity for word sense identification*, WordNet: A Lexical Reference System and its Application (1998), MIT Press, Cambridge, Massachusetts, 265–283.

[47] P. Resnik and D. Yarowsky, *Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation*, Natural Language Eng. **5** (1999), no. 2, 113–133.

[48] K. Ahmed, I. Izadi, T. Chen, D. Joe, and T. Burton, *Similarity analysis of industrial alarm flood data*, IEEE Trans. Automat. Sci. Eng. **10** (2013), no. 2, 452–457.

[49] G. Jeh and J. Widom, *Simrank: a measure of structural-context similarity*, Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 538–543.

[50] J. Portisch, G. Costa, K. Stefani, K. Kreplin, M. Hladik, and H. Paulheim, *Ontology matching through absolute orientation of embedding spaces*, European Semantic Web Conference, Springer, 2022, pp. 153–157.

[51] R. Souto M. de Barros, J. I. González Hidalgo, and D. R. de Lima Cabral, *Wilcoxon rank sum test drift detector*, Neurocomputing **275** (2018), 1954–1963.

[52] K. W. Jørgensen and L. Kai Hansen, *Model selection for gaussian kernel PCA denoising*, IEEE Trans. Neural Networks Learn. Syst. **23** (2011), no. 1, 163–168.

[53] F. Tonin, A. Lambert, P. Patrinos, and J. Suykens, *Extending kernel PCA through dualization: sparsity, robustness and fast algorithms*, International Conference on Machine Learning, PMLR, 2023, pp. 34379–34393.