STM JOURNALS

Journal of
Polymer & Composites

ISSN: 2321-2810 (Online)
ISSN: 2321-8525 (Print)
Volume 14, Special Issue 1, 2026
DOI (Journal): 10.37591/JoPC

https://journals.stmjournals.com/jopc

Research JoPC

Record Linkage in Knowledge Discovery Process Using Angle Based Machine Learning

A. Sairam[1], D. Sasikumar[2], V. Chithambaram[3,*], R. Sendhilkumar[4], A.B Hajira Be[5]

Abstract

Record linkage is a critical data cleansing step in the knowledge discovery process, aimed at identifying and resolving inconsistencies across datasets. This study proposes an enhanced record linkage framework tailored for uncertain and large-scale data using a combination of distance measurement, probabilistic modeling, and semantic reasoning. A novel angle-based distance measurement technique is introduced to optimize matching between candidate records. To further boost match accuracy, a Finite Mixture Model (FMM) is applied as a probabilistic function to assess similarity distributions. An innovative feature of the approach is the integration of ontology-based semantic matching, which uses parent–child (is-a) relationships to validate semantic proximity and eliminate inconsistent record pairs based on agreement scores. Semantic vectors are weighted using the best agreement principle, improving the accuracy of linkage decisions. This hybrid method effectively fuses syntactic, probabilistic, and semantic dimensions, demonstrating superior performance over traditional linkage algorithms. The system is evaluated using complex biomedical datasets, including NCBI GenBank, Gene Ontology, and SwissProt. Experimental results show significant improvements in precision, recall, and F-measure, particularly in noisy or ambiguous data scenarios. The proposed methodology not only addresses key scalability and reliability challenges but also offers a foundation for more accurate data integration in heterogeneous environments.

Keywords: Gene ontology, machine learning, NCBI genbank dataset, record linkage

INTRODUCTION

Mining techniques that process and analyze massive databases have garnered the attention of both the business world and academics. This is primarily due to the massive amounts of data that are being collected by businesses and research projects. One single task that is becoming increasingly important in multiple applications is the matching of records like entities that have been acquired from a wide variety of multiple databases. To enhance the quality of the data, it is necessary to combine and process the information gathered from diverse sources. This enriches the data and facilitates much more detailed analytics. Common examples of record matching include databases of people, such as customers, employees, patients, students, taxpayers, and travelers.

Record linkage improves data integrity and quality by enabling the reuse of conventional data sources. Further, it reduces costs and limits computational efforts during the data acquisition phase [1]. Specifically, in the healthcare sector,

***Author for Correspondence**
V. Chithambaram

[1]Professor, Department of Computer Science Engineering, SIMATS Engineering, Saveetha University, SIMATS, Chennai, Tamil Nadu, India
[2]Professor, Department of Computer Science Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India
[3]Professor, Department of Physics, Rajalakshmi Engineering College (Aotonomous), Thandalam, Chennai, Tamil Nadu, India
[4]Professor, Department of Computer Science Engineering, Thirumalai Engineering College, Kanchipuram, Tamil Nadu, India
[5]Professor, Department of Computer Applications, Karpaga Vinayaga College of Engineering of Technology, Madhuranthagam, Chengalpattu, Tamil Nadu, India

Received Date: June 25, 2025
Accepted Date: August 22, 2025
Published Date: October 16, 2025

Citation: Sairam, D. Sasikumar, V. Chithambaram, R. Sendhilkumar, A.B Hajira Be. Record Linkage in Knowledge Discovery Process Using Angle Based Machine Learning. Journal of Polymer & Composites. 2026; 14(Special Issue 1): S1157–S1170p.

matched records contribute to better policy decisions. Likewise, while traditional data collection is often expensive and time-consuming [2,3], record linkage supports more efficient health surveillance by identifying suspicious patterns, such as disease outbreaks. Statistical agencies also use record linkage to integrate census data for downstream analysis [4]. Additionally, deduplication through record linkage is widely used in business to improve the quality of mailing lists or consolidate data in e-commerce and marketing platforms. Government bodies use it in taxation and social security to detect individuals with multiple registrations. Domains like fraud detection, national security, and crime prevention also show high interest in record linkage [5].

Record linkage techniques have become increasingly relevant in materials informatics. With the rapid growth of materials databases, such as those documenting polymer blends, composite compositions, mechanical properties, and thermal stability, accurately linking and integrating these datasets is crucial. For instance, linking experimental data from polymer tensile tests with simulation data or with chemical structure databases can support better prediction models for new composite materials. This form of intelligent data mining can significantly accelerate the design of next-generation lightweight composites or smart polymer systems.

The main problem associated with record linkage methods is that the presence of missing data leads to misclassification of instances. This challenge is also reflected in materials datasets, where incomplete records, such as missing polymer additive information or fabrication parameters, can make it difficult to establish meaningful relationships. Consequently, finding accurate similarity or distance metrics between linked classes becomes challenging if even one of the linked values is missing. Addressing this issue through robust imputation or probabilistic linkage methods is essential, particularly when optimizing polymer formulations or predicting composite behavior using historical experimental data.

In addition, several issues are found during the selection of record fields, which are discussed below: The main issue is the values, which influences the quality of the generated recorded pairs. Preferably, the fields with errors, missing or variation values is chosen and in field value, the error generates the potential result, which is inserted into an incorrect block. Finally, this leads to missed true record match [6–9]. One such approach to avoid variations and variations is the generation of blocking keys w.r.t the record fields. The records with true matches have a common value which is inserted over the similar blocks.

Another issue during the blocking keys definition is its frequency distribution of the field values, which is used for blocking the keys. This affects generated blocks size and often affects the encodings. The largest blocks, which is generated during the process of indexing results in domination of execution time of the comparison. Since, it contributes a larger number of record pairs. Hence, the use of fields with uniformly distributed values would result in equal block sizes.

The trade-off should be considered while defining the blocking keys. On one hand, large availability of smaller blocks will lead to less generated record pairs. This increases likely missed available true matches. On the other hand, blocking keys with larger blocks generates increased record pairs with more true matches with more candidate record pairs [10-21].

The Record Linkage techniques used for indexing are used to remove the duplicate datasets splits the database records into overlapping blocks. These matches are then inserted into the same set of overlapping block and it gets non-matched with dissimilar blocks.

The main objective of the proposed study is shown below:

- The main objective of the study is to improve the record linkage in uncertain datasets using distance measurement and probabilistic function.
- To estimate the distance between the linked datasets using angle similarity measurement, where the datasets are clustered into groups using certain angle.

- To improve the record linkage strategy to link the effective linked records using Finite Mixture process, which improves the accuracy at the time of clustering process. To link well the record pairs in cluster, the proposed study aims at using ontology concept with parent–child (is-a) relationship, which helps to estimate the agreement and non-agreement pairs.

The proposed study improves the record linkage method through the concept of distance measurement and probabilistic function. An angle-based measurement is used to find the matching between the linked records. The Finite Mixture Model is used as a probabilistic function to improve the accuracy of the proposed method in providing proper matches.

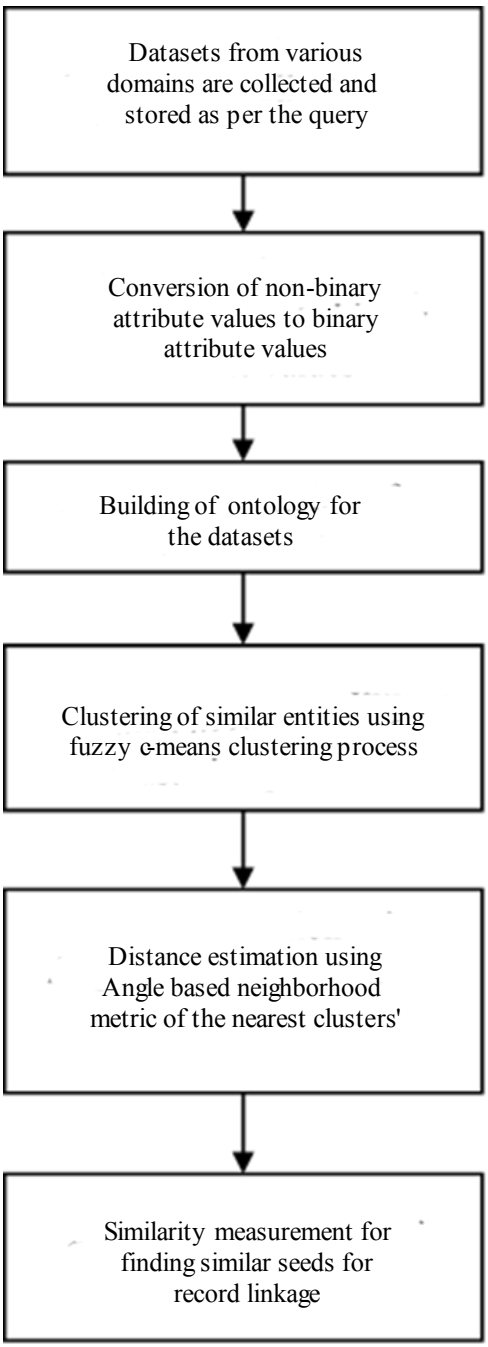

Figure 1. Proposed record linkage process using angle-based similarity and ontology-based semantic matching [1].

PROPOSED ONTOLOGY-BASED RECORD LINKAGE

It is possible to think of record links as a function that takes in sets of identified records and returns sets of predicted matches (where two records are predicted to refer to each other) and sets of predicted no matches (where two records are predicted to refer to different people) as an output. Record links can be thought of as a function that takes in sets of identified records. To put it another way, record links are a function that accepts sets of identified records as input and returns sets of predicted records as output respectively. The concept of record linkage can be dissected into its component parts. Figure 1 provides a concise overview of the record link process, and the subsequent paragraphs contain additional information that can be found in the following:

Treatment of Missing Data

The absence of specific data occurs whenever the value of the field that belongs to the data holder is not recorded. This lack of recording results in the absence of the data. This could occur as a result of the absence of specific characteristics or the lack of sufficient information. Both of these possibilities are possible. The problem of missing data, which has an impact on a variety of different aspects of the record linkage system, is discussed in this section.

Missing Data and Field Weighting

For this investigation, field weighting is determined by an algorithm that takes into consideration the dependent probability of field agreement in accordance with the status that corresponds to each domain. This algorithm is used to determine field weighting. To be more specific, for each field to carry out its functions, it is assigned two values. The probability of the value of the ith field is considered to be the true matches among record pairs; and the probability of the value of the ith field is considered to be the true non-matches among record pairs. This algorithm however does not indicate the factor in the algorithm of missing field values. The study considers two options based on disagreement and missing elements, which are discussed below:

Treat Missing as Disagreement

The algorithm for field weighting includes a set of vectors, representing the record pair similarities. First, when comparing fields for vectors, the missing value of the field is treated like a disagreement. Basically a field value receives a null similarity value, if it is compared with a record value where the field value of it is missing. This method implicitly discounts the discriminatory powers of fields, because they are not seen as being acceptable.

Exclude Missing and Discount

The record inputs algorithm contains all field values but lead to an overestimation of the power in the field. If a field's weighting method uses only samples that show field values, but the field value mostly lacks, then the discriminatory power estimates are not accurate. The field value is estimated at the same time. Therefore, the proposed system aims at the reduction of the weight by each field missing values based on completed records.

Steps Involved in Record Linkage

It is possible to think of record links as a function that takes in sets of identified records and returns sets of predicted matches (where two records are predicted to refer to each other) and sets of predicted no matches (where two records are predicted to refer to different people) as an output. Record links can be thought of as a function that takes in sets of identified records. To put it another way, record links are a function that accepts sets of identified records as input and returns sets of predicted records as output respectively. The concept of record linkage can be dissected into its component parts to facilitate better comprehension. Therefore, record linkage is to accurately categorize record pairs as either Predicted Matches or Non-Matches. This is the objective of record linkage. Because of this, the predicted matches are referred to as true matches, and the non-matches are referred to as true non-matches. the classification of the two different types of record pairs is based on this classification.

Proposed Blocking Model

The major contribution of the proposed study consists of distance and probabilistic functions are introduced to improve the record-linkage. In this case, the distance function is angle-based measurement of distance similarity and the probabilistic function is the finite mixture model (FMM). This gives more precision than the normal clustering process. The ontology concept is further developed to connect the best pair within a cluster accurately. for this process to be successful, it is necessary for a parent and child to have a good relationship with one another. The distance that exists between each pair that is contained within the semantic vector, which also includes the parent concept, is what determines the weight of the vector. In conjunction with ontology, the utilization of the distance function allows for the determination of the routes that are the most optimal. To be more specific, in this particular instance, non-agreement pairs are removed from the database to facilitate efficient recording connection recording.

The key issue is the hybrid form that, as opposed to traditional works, combines distances and probabilistic records. The procedure to improve the record link for large databases also tries to be standardized. The current work focuses on keeping the exact similarity of the datasets with binary relations. The following steps are taken to achieve the proposed model more feasible.

We begin the process by constructing the ontology framework that will allow us to access the databases. This is the first step in the process. In this scenario, the reference datasets are utilized to identify datasets from a variety of domains. This is done to facilitate accurate data collection.

Similarly, the non-binary characteristics of the dataset are transformed into binary values during the second step of the process. By utilizing the numbers 1 and 0, these binary values provide information regarding the availability of the given objects.

Fuzzy C-means clustering is the third method that we use to group together the components of the group that are comparable to one another. This method is used to group together the components of the group.

It is necessary to cluster datasets that are situated in close proximity to the cluster that contains the data to accomplish the fourth step, which is to reduce the squared distance. to determine the location of the cluster center, a neighborhood distance metric that is based on angles is utilized.

Following the formation of clusters that are devoid of datasets that have been incorrectly classified, the fifth step is to construct the ontology and make use of similarity measurement to locate seeds that are comparable to one another.

Ontology Matching Problem

In the last decade, ontology matching became a vibrant field of research. Existing methods discuss various techniques for ontological matching, improvements or complete matching. These system in particular illustrate a paradigm common to all systems that probably use a name-based alignment method. This paradigm is a sequential process, starting with the analysis of various types of evidence, in most cases with the emphasis on the labels involved, and generating a series of weighted matching hypotheses as an intermediary result. A subset of hypotheses generated are chosen as final output from the intermediate result. In the first stage, calculation, aggregation, spreading, and any other method to refinish similarity values are dominated. In the second phase, the techniques used range from thresholds to the choice of consistent subsets that can be optimal for a single objective function. The intermediate result is mostly modeled on several correspondences with confidence scores. These trust ratings are aggregated values derived from the analysis of the tokens on the ontology labels. Using various examples, we argue that the problem of extraction should be different to make tokens and logical entities. If the acceptance or rejection of the correspondence is not possible, it cannot be exploited if two tokens have (or do not have) the same meaning. Any reasonable mining should, however, conform to the underlying assumptions. This can be ensured only if the assumptions can be expressly shaped.

Proposed Ontology Model

The root and seed nodes are determined by means of tables. It is similar in finding the root for every concept of ontology that could have more than one route for the root node. Here, the distance between the root and the seed for each path or route is unique. This is a total of the available seed borders used to allocate the concept weight along the path. If the concept of ontology has a varying design, the final weight shall be highly assigned and, if the data or document has been formed, the semantic vector representation shall be computed and the similarity distance shall be calculated. For many paths or borders, maximum weights for each concept are assigned.

If there is only one-on-one alignment, i.e., matching in the source ontology with the one in destination and vice versa is made through single entity comparison. The ontology correspondence shown in the present study is as follows:

$$\begin{cases} \max & MF(x) \\ s.t. & X = \left(x_1, x_2, ..., x_{|O_1|},\right)^T \\ & x_i \in \left\{1, 2, ..., |O_1|\right\},\ i = 1, 2, ..., |O_1| \end{cases}$$

where |O1| is considered as the ontology entity sets cardinalities and |O2| is considered as the ontology entity sets cardinalities, xi is considered as the ith correspondence pair. The study aims to increase the membership function, which is required to estimate the value of X. An interactive activation network node in the context of ontology mapping presupposes that in ontology O1 a concept C1i can be mapped into the concept C2j in ontology O2. The initial node activation can be considered as the similarity of two concepts between (C1i, C2j). The node activation can be updated by the simple rule, where ai indicates the activation function of node i, which is represented as ni, and the node's net input is neti.

$$a_i(t+1) = \begin{cases} a_i(t) + net_i\left(1 - a_i(t)\right) & net_i > 0 \\ a_i(t) + net_i\left(a_i(t)\right) & net_i < 0 \end{cases}$$

The neti is obtained from the three different sources, i.e., it is obtained from neighbor, bias, and external inputs.

$$net_i = i_{str}\left(\sum_j w_{ij} a_j + bias_i\right) + e_{str}\left(input_i\right)$$

where wij is the weights of connection exist between the documents ni and nj, aj is defined as the activation function of the node nj, biasi is defined as the bias of a node ni, istr is defined as the constants that opts for contribution from the internal source inputs that has to be manipulated readily and estr is defined as the constants that opts for contribution from the external source inputs that has to be manipulated readily.

Fuzzy C-means Clustering

The algorithm able to calculate the squared error, which allows us to determine the weighted total of the distances that exist between the centers of the cluster and the elements that are relevant to the fuzzy cluster system. Because of this, we will be able to find the most effective solution possible. A connection exists between the m-number and the membership rates of the performance index. This connection is statistically significant. With the increasing m the partition tends to become fuzzier and the fuzzy c means clustering algorithm for any m number is given as (1, ∞), where the algorithm tends to converge. The conditions required to reach the lower value of Equation 4 is given below:

$$U_{ik} = \left(\sum_{j=1}^{c}\left(\frac{\left\|x_k - v_i\right\|}{\left\|x_k - v_j\right\|}\right)^{\frac{2}{m-1}}\right)^{-1} \quad \forall i, k$$

Record Pair Comparison

At the stage of record pair comparison, the fields in a record pair that are comparable to one another are combined into a single measurement of the degree to which the record pair is comparable. This is done to maximize the accuracy of the measurement. By using the weights that are associated with agreement and disagreement for each field, it is also possible to determine a similarity score for the record pair. This score can be determined by using the weights. In this article, we will discuss an alternative approach to comparing record pairs. When the fields of two records are in agreement with one another, the similarity weights of those records are added to their total score. When the contract weights of the two records are consistent with one another, this is the result. In situations where the agreement weight of one field is in conflict with the agreement weight of another field, the dissimilarity weights of record pairs that correspond to each other are added to the final score. This ensures that the final score is always accurate.

Asymmetric Angle Based Similarity Measurement (ASM)

When attempting to locate the cluster nodes that are situated in the closest proximity to the cluster heads, it is beneficial to make use of a neighborhood distance measure that is based on angles. It is necessary to make use of an angle-based measurement of the neighborhood distance to ascertain the distance that is the shortest overall. The formula (s + 1)-1 can be utilized to compute the weighted matrix, which is also referred to as the shortest distance. The reason for this is that s is the equivalent of the shortest distance shown in Figure 2. To further elaborate, when taking into account N training samples, the following is the measurement of the neighborhood distance based on the angle, utilizing the circumference angle value (□): (a)

Similarly, the nodes that are located within this high-energy region will be considered cluster centers rather than other nodes, and the center of the line that connects both circles will be considered the center of the circle. The red line nodes that are still present in the neighborhood are the ones that represent the distance nodes throughout the neighborhood. to ascertain the distance between the seed and the cluster center, angle-based metrics are utilized for determining the distance between the two. We make it a point to keep track of the seeds and to ascertain the distance that is the shortest starting point from the geographic center of the cluster. A value that is comparable can be discovered. It is common practice to use the same

Weight matrix calculation and choose the weighted maximum seed whenever multiple queries are generated. It is accepted that this behavior is something that people do. When the weight matrix is being computed, the semantic vector is assigned to the seed that was chosen because it possesses the highest weight throughout the process. This seed was chosen because it was previously selected. Consequently, the overall distance that the vectors are separated by is computed as a result of this.

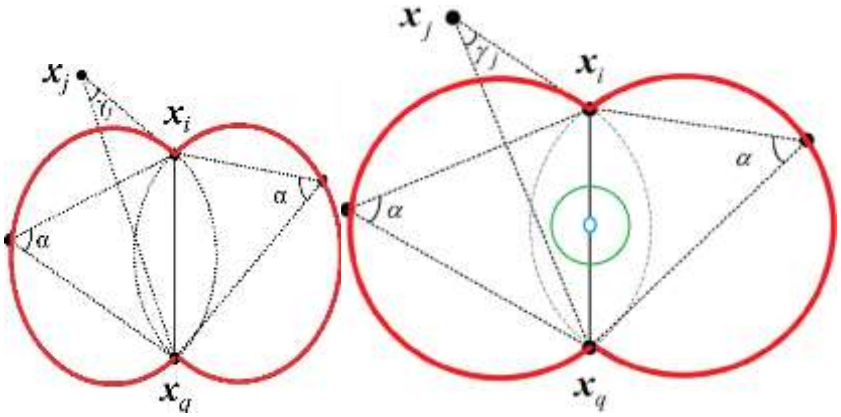

Figure 2. Distance estimation using angle-based neighborhood metric for similarity computation [4].

For determining the distance between the neighborhoods, the clustering process takes into consideration the distance between each neighborhood. The utilization of neighborhood distance estimates that are obtained based on angles makes it possible to discover a semantic similarity metric among the nearest neighbors. This is made possible by the utilization of neighborhood distance estimates. The angle measurement of the neighborhood distance is the one that is specifically chosen when conducting an analysis of similarities in a setting that contains one or more classes. This is because the angle measurement is the one that is most accurate.

Ontology Semantic Measurement Using Clustering Model – Distance Phase

The primary goal is to eliminate constraints through the utilization of direct concept-to-concept matching that is established on the basis of semantic similarity. For carrying out the process, two separate datasets are utilized, and the similarities and differences between concepts that are identical are taken into consideration. When it comes to ontology, the process of connecting the ideas (references) is referred to as the ontology match. The links that are available are taken into consideration when determining the degree of similarity between the two data sets. The distance between concepts that are close together and the degree to which they are semantically similar is inversely proportional to the degree to which they are located in close proximity to one another. Find out who your closest neighbors are by calculating the distance between them after taking into account the angle at which they are located in the neighborhood. In a space model, the datasets that are contained within it are represented by space vectors that have N dimensions, where N is the total number of concepts that are available to the user. The concept that is accessible is represented by the number 1, while the concept that is not accessible is represented by the number 2. As a result of the fact that the vector only contains information that is represented by binary values, there is no discernible change in the vector. It is possible to determine the degree of similarity between two datasets by computing the cosine difference between the vectors that are associated with each of the datasets. During the process of comparing two datasets, this is done. to get around this limitation, a standard vector space model makes use of an N-dimensional vector that represents the dataset, in addition to a unity representation for the idea or seed that is extracted from the corresponding dataset. This allows the model to circumvent the limitation. to improve the vector description, it is possible to inhibit the vectors that have weights that are not zero in this attribution. This helps to ensure that the vector description is improved. The proportions of the progenitors on the plant are reflected in the dimensions of the seed, which reflect those proportions. The weight (w) of the ideas is one of the relationships that are involved in the relationship that exists between the seed and its ancestor. In this instance, the weight is directly proportional to the non-orthogonal component of the least common predecessor. to ascertain the distance that exists between them, we make use of a minimum distance vector. Using semantic data, the cosine pair relationship that exists between two sets of data (a and b) can be utilized for defining the similitude measurement. This can be accomplished by using the relationship between the two sets of data.

$$Similairity(a,b) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} (a_i)^2} \sqrt{\sum_{i=1}^{N} (b_i)^2}}$$

where, n is regarded as the union of entire concepts occurring in both a and b datasets.

Record Pair Classification

It is necessary to draw a classification limit between record pairs to differentiate between predicted matches and non-matches when performing the record pair classification step. This is done to differentiate between the two types of records. It is possible that the conventional approach to record linking, which makes use of clerical evaluation to determine the actual match status of record pairs, could incorporate a third intermediate class. This is a possibility. to classify registry pairs into predicted and predicted non-match sets, four different methods are utilized. These methods are reciprocity, Force A, Force B, and the most similar. Reciprocity is the most similar method. According to the similarities

that exist between the records, these methods can be established. One of the potential drawbacks of the similar classification method is that it may result in the consolidation of multiple records from dataset B into a single record, or vice versa. This could come about because of the approach. The assumption that the A and B datasets are de-duplicated, on the other hand, suggests that this should not take place.

RESULTS AND DISCUSSION

The experiments that are described in this study made use of not just one but several computers throughout the course of the experimentation sessions. An in-depth explanation of these digital assets is provided in the following paragraphs. It is necessary to have several machines to carry out the experiments, which are not only extensive but also complicated. Similarly, for each group of experiments to be considered comparable, they are all carried out on the same machine if they are to be considered comparable. Even though not all of the time that is reported in this thesis is comparable, this is the case. After being described in greater detail in the subsequent sections of the dissertation, the snapshots that were used for the experiments are documented after they have been described.

Datasets Description

The Car Evaluation Database is a straightforward hierarchical decision-making model, which serves as the basis for the database's implementation. There is a preference for using lowercase letters when printing input attributes. There are four different ideas that are incorporated into the model. These are CAR, PRICE, TECH, and COMFORT. The concept that constitutes the target is achieved through the utilization of these concepts as intermediaries. There is a connection between each concept and its lower-level offspring in the model that was initially developed. As a result of the fact that the fundamental concept structure is already known, it is possible that techniques for structure discovery and constructive induction will find this database to be particularly useful.

Experimental Evaluation

Out of the candidate record pairs that are generated, there are four metrics that are utilized to determine the quality and completeness of the candidate record pairs. There are a total of nM record pairs that are associated with each other, while there are nN record pairs that are not associated with each other. Using is how the linking of two databases is accomplished.

$$nM + nM = nA \times nB$$

Utilizing is one method that can be utilized to accomplish the deduplication of databases.

$$nM + nN = nA(nA-1)/2$$

Quality Measures

For evaluating the quality of record links, the accuracy measurement that is typically utilized is insufficient. The reason for this is that matches, and non-matches are typically not balanced within the weight vector set W. This is the reason why this exists. Given the high number of non-matches, an overly optimistic evaluation of the accuracy and performance outcomes would be overshadowed by the magnitude of the problem. This happens because the problem is so widespread. To accomplish the evaluation of the quality of the classifier, the F-measure, which is a harmonic mean of recall and precision, is instead utilized to accomplish the following:

$$\text{F-measure} = 2PR/(P + R), \text{ with}$$

$$\text{Precision} = TP/(TP + FP) \text{ and}$$

$$\text{Recall} = TP/(TP + FN).$$

where,

TP is defined as the total number of true positives (true matched record pairs, which is classified as matches),

TN is defined as the total number of true negatives (true non-matched record pairs, which is classified as non-matches),

FN is defined as the total number of false negatives (true matched record pairs, which is classified as nonmatches), and

FP is defined as the total number of false positives (true non-matched record pairs, which is classified as matches).

With the assistance of an evaluation of similarity between two sets of data—one that is similar and one that is different—the system that has been proposed is able to determine whether the two sets of data are comparable to one another. There are several datasets that are presented in Table 1 [5], including the weight matrix and the shortest distance that was measured using the method. These datasets are displayed in the table. When compared to the distance between other databases, the shortest distance between D3 databases is substantially shorter than the distance between other databases. Therefore, when the weight matrices of the other databases are compared to the weight matrices of the D3 database, the weight matrix of the D3 database presents itself as particularly noteworthy.

The weights of two distinct datasets (ai and bi) that are contained within a single database can be utilized to determine whether the datasets are in agreement with one another. This can be accomplished by using the weights of the datasets simultaneously shown in Table 2. Comparing the degree of agreement or disagreement across all datasets is analogous to the process of evaluating the degree of similarity between database variables to determine the degree of similarity between the variables. The results of a comparison between the dataset pertaining to artificial intelligence and the bi dataset that is included in the D1 dataset are presented in Table 3. This table contains the results of the comparison. The similarity measurement of datasets in which there is no agreement significantly decreases. This is the case because there is no agreement. Similarly, when you compare datasets that are comparable, you will notice that the seeds are quite comparable to one another. This is something that you will notice. We have samples that provide evidence of agreement at 1:9, as well as samples that provide evidence of disagreement. Both types of samples are presented here. Datasets that use agreement and non-according strategies are comparable to one another, as demonstrated in Figure 3, which also displays the accuracy calculated between the various modes of the proposed method. Figure 3 also displays the accuracy calculated between the various modes.

Table 1. Angle based ontology measurement [5]

	Shortest distance	**Weight matrix**
D_1 with 576 instances	4	$(4+1)^{-1} = 0.2$
D_2 with 576 instances	3	$(3+1)^{-1} = 0.25$
D_3 with 576 instances	2	$(2+1)^{-1} = 0.33$

Table 2. Weights of FMM for agreement and disagreement [14]

Database	a_i	b_i	**Agreement**	**Disagreement**
D_1 with 576 instances	0.9987	0.0004	8.9260	-5.1974
D_2 with 576 instances	0.9821	0.0010	8.6227	-5.1783
D_3 with 576 instances	0.9728	0.0020	8.2324	-4.7285

Table 3. Evaluation of similarity between the collected databases [6,7].

Dataset combinations	**Similarity between the variables**
$D_1(a_i)$, $D_2(b_i)$	0.1034
$D_1(a_i)$, $D_1(b_i)$	0.9342
$D_3(a_i)$, $D_2(b_i)$	0.0923
$D_2(a_i)$, $D_4(b_i)$	0.1223
$D_2(a_i)$, $D_2(b_i)$	0.9423
$D_3(a_i)$, $D_3(b_i)$	0.8923
$D_4(a_i)$, $D_4(b_i)$	0.9143

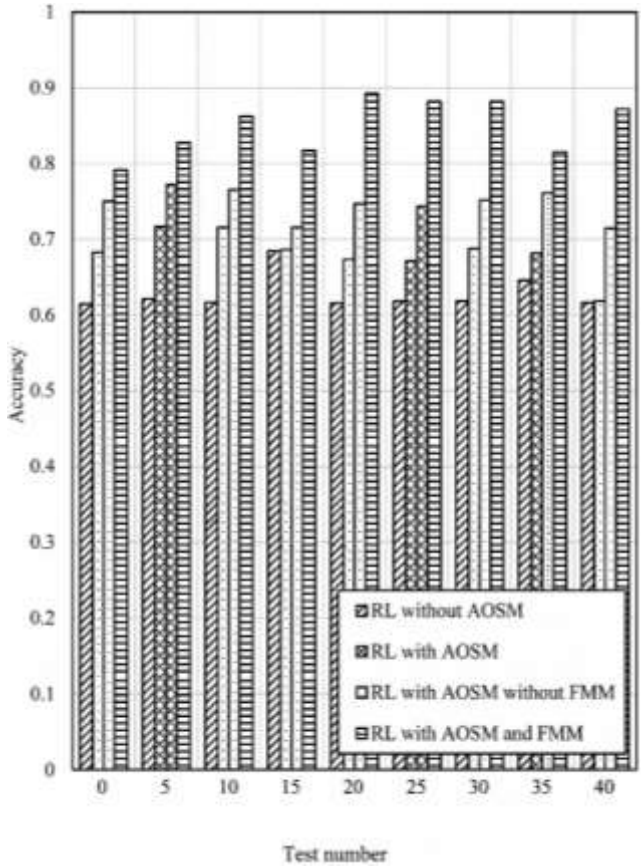

Figure 3. Accuracy for dataset 1 [7].

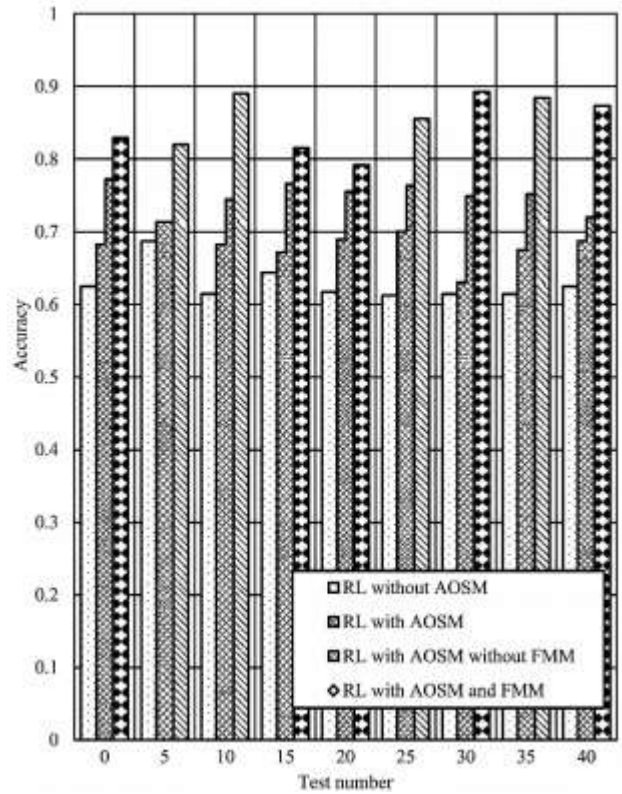

Figure 4. Accuracy for dataset 2 [2,3,14].

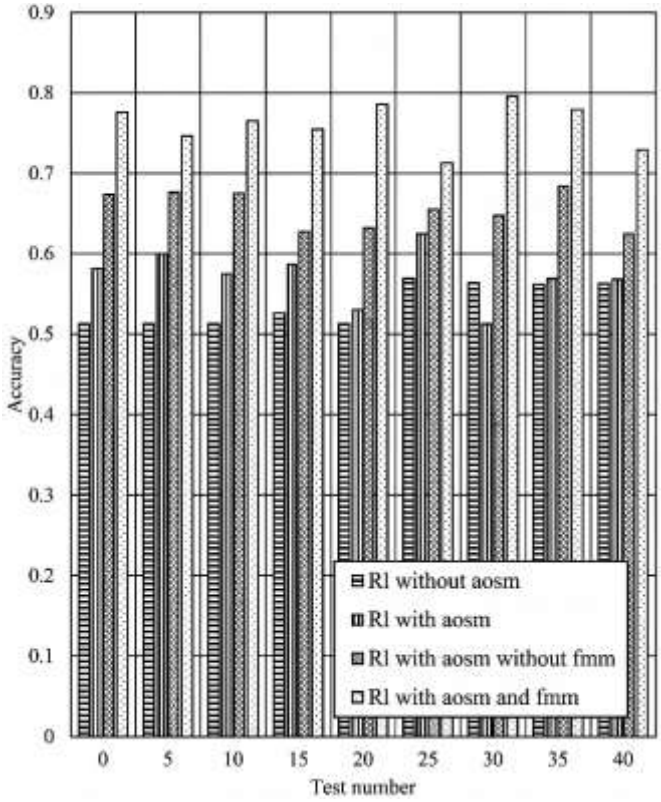

Figure 5. Accuracy for dataset 3 [2,14,20].

In Figure 3, we see an illustration of the implementation of the accuracy for dataset D1 using a variety of different modes of the method that was proposed. When training sets do not accurately represent the output that is desired, the accuracy of non-semantic record links decreases. This is because the output that is desired is not accurately represented. The output that is desired is a description of the datasets that are more comparable to one another than the others. Similarly, when training samples that are incorrect are utilized, the accuracy of the other modes that do not make use of AOSM and FMM is significantly diminished. This is the case. The accuracy of the method that was proposed is maintained at a level that is between 0.80% and 0.91% across all of the datasets, which means that this is completely respected. A comparison is made between the approach that has been proposed and other methods that are considered to be more conventional. These methods include supervised record linkage [2], probabilistic record linkage [3], and graph-based record linkage [7].

When it comes to dataset D1, the accuracy is applied in several different applications of the method that has been proposed shown in Figure 4. When training sets do not accurately represent the output that is desired, the accuracy of non-semantic record links decreases. This is because the output that is desired is not accurately represented. The output that is desired is a description of the datasets that are more comparable to one another than the others.

Similarly, when training samples that are incorrect are utilized, the accuracy of the other modes that do not make use of AOSM and FMM is significantly diminished provided detail in table 4. This is the case. The accuracy of the method that was proposed is maintained at a level that is between 0.80% and 0.91% across all of the datasets, which means that this is completely respected. A comparison is made between the approach that has been proposed and other methods that are more conventional. These methods include supervised record linkage [3], probabilistic record linkage [2], and graph-based record linkage [7]. An exhaustive investigation into the accuracy, recall, and F-measure of the method that was proposed is carried out. When compared to the method that was proposed, the performance of these alternative methods, which include supervised record linkage, probabilistic record linkage, and graph-based record linkage, is extremely poor.

Table 4. Comparison over various database [2,4,14,20]

RL Methods	D1 with 576 instances		
	Precision	***Recall***	***F-measure***
RL + AOSM + FMM	0.96	0.97	0.96
RL with AOSM	0.93	0.97	0.95
Graph based record linkage	0.92	0.77	0.84
Probabilistic record linkage	0.98	0.75	0.93
Supervised record linkage	0.97	0.9	0.93

RL Methods	D2 with 576 instances		
	Precision	***Recall***	***F-measure***
Proposed RL + AOSM + FMM	0.97	0.94	0.95
RL with AOSM	0.98	0.92	0.95
Graph based RL	0.75	0.98	0.85
Probabilistic RL	0.9	0.6	0.72
Supervised RL	0.98	0.88	0.93

RL Methods	D3 with 576 instances		
	Precision	***Recall***	***F-measure***
Proposed RL + AOSM + FMM	0.69	0.6	0.64
RL with AOSM	0.58	0.74	0.65
Graph based RL	0.45	0.74	0.57
Probabilistic RL	0.36	0.86	0.5
Supervised RL	0.43	0.75	0.55

The proposed system shows the snapshots of the simulation carried out in Java platform. The simulation is carried out using cars datasets with 1728 instances combined.

CONCLUSION

In this study, the Record Linkage is considered as an important tool for data mining over large data. It is seen that Record Linkage offers multiple advantages for matching and identifying relevant records and eliminates irrelevant records over larger datasets. It is seen from the study that the linkage information is available in more than a single location. The study eliminates such record linkage errors, incomplete datasets and unagreeable records and this increases the scalability of the system. The present study uses an improved record linkage technique using the concept of ontology-based record mapping and angle-based similarity measurement. The proposed method improves the reliability of finding similar record links between two different distant datasets. For instance, linking data between polymer blend ratios and final mechanical output, or mapping nano-reinforcement properties in composites to thermal degradation behavior, can be more accurately achieved using the proposed record matching model.

The study is found to have a robust design that effectively maps and classifies the related and non-related records. The study is said to have a comprehensive design that utilizes optimal linkage strategy to improve the quality of obtained records. The study considers the quality of databases, size of each file, uniqueness of identifiers and penalties liked with FP and FN links. It also considers computational time and linkage software programs To simulation.

Therefore, by applying this technique to composite and polymer datasets, researchers can accelerate material discovery, improve predictive modeling, and facilitate reuse of historical data for material innovation and performance optimization.

REFERENCES

1. Bellomarini L, Fayzrakhmanov RR, Gottlob G, Kravchenko A, Laurenza E, Nenov Y, et al. Data science with Vadalog: Knowledge Graphs with machine learning and reasoning in practice. Future Gener Comput Syst. 2022;129:407–22.
2. Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng. 2021;35(1):614–33.
3. Yesodha KRK, Jagadeesan A, Gowrishankar V, Jiwani N, Yuvaraj N. The smart optimization model for Supply Chain Management using IoT Applications. In: 2024 15th Int Conf on Computing Communication and Networking Technologies (ICCCNT). IEEE; 2024. p. 1–6.
4. Jagadeesan A, Gowrishankar V, Yesodha KRK, Yuvaraj N. Enhanced Supply Chain Management using IoT based Predictive Analysis. In: 2024 15th Int Conf on Computing Communication and Networking Technologies (ICCCNT). IEEE; 2024. p. 1–6.
5. Mohammed AS, Mallikarjunaradhya V, Sreeramulu MD, Boddapati N, Jiwani N, Natarajan Y. Optimizing Real-time Task Scheduling in Cloud-based AI Systems using Genetic Algorithms. In: 2024 7th Int Conf on Contemporary Computing and Informatics (IC3I). IEEE; 2024. p. 1649–53.
6. Pont S, Bond DM, Shand AW, Khan I, Zoega H, Nassar N. Risk factors and recurrence of hyperemesis gravidarum: A population-based record linkage cohort study. Acta Obstet Gynecol Scand. 2024;103(12):2392–400.
7. Hassani H, Entezarian MR, Zaeimzadeh S, Marvian L, Komendantova N. An oversampling-undersampling strategy for large-scale data linkage. Front Big Data. 2024;8:1542483.
8. Cybulski L, Chilman N, Jewell A, Dewey M, Hildersley R, Morgan C, et al. Improving our understanding of the social determinants of mental health: a data linkage study of mental health records and the 2011 UK census. BMJ Open. 2024;14(1):e073582.
9. Boyd A, Evans K, Turner E, Flaig R, Oakley J, Campbell K, et al. UK Longitudinal Linkage Collaboration (UK LLC): The National Trusted Research Environment for Longitudinal Research. Int J Popul Data Sci. 2025;10(1).
10. Bailey G, Lee A, Ahmed S, Scanlon I, Cowley L, Stuart A, et al. Improving opportunities for data linkage within Children Looked After administrative records in Wales. Int J Popul Data Sci. 2025;10(1).
11. Nielsen TC, Nassar N, Boulton KA, Guastella AJ, Lain SJ. Estimating the prevalence of autism spectrum disorder in New South Wales, Australia: A data linkage study of three routinely collected datasets. J Autism Dev Disord. 2024;54(4):1558–66.
12. Cash RE, Crowe RP, Swanton M, Boggs KM, Goldberg SA, Sullivan AF, et al. Creation of a novel national dataset through linkage of EMS transport destination and verified ED capability. Prehosp Emerg Care. 2025;just-accepted:1–8.
13. Tan YCRS, Jin A, Seow LHA. Association between body mass index, diabetes and breast cancer incidence: a population-based cohort study. Lancet Reg Health West Pac. 2025;55.
14. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform. 2015;56:80–6.
15. Li T, Zhang L, Lu W, Hou H, Liu X, Pedrycz W, Zhong C. Interval kernel fuzzy C-means clustering of incomplete data. Neurocomputing. 2017.
16. Mandacaru PMP, Andrade AL, Rocha MS, Aguiar FP, Nogueira MSM, Girodo AM, et al. Qualifying information on deaths and serious injuries caused by road traffic in five Brazilian capitals using record linkage. Accid Anal Prev. 2017;106:392–8.
17. Chi Y, Hong J, Jurek A, Liu W, O'Reilly D. Privacy preserving record linkage in the presence of missing values. Inf Syst. 2017;71:199–210.
18. Jurek A, Hong J, Chi Y, Liu W. A novel ensemble learning approach to unsupervised record linkage. Inf Syst. 2017;71:40–54.
19. Smith D. Secure pseudonymisation for privacy-preserving probabilistic record linkage. J Inf Secur Appl. 2017.
20. Abril D, Torra V, Navarro-Arribas G. Supervised learning using a symmetric bilinear form for record linkage. Inf Fusion. 2015;26:144–53.
21. Lu Y, Sinnott RO. Semantic privacy-preserving framework for electronic health record linkage.

