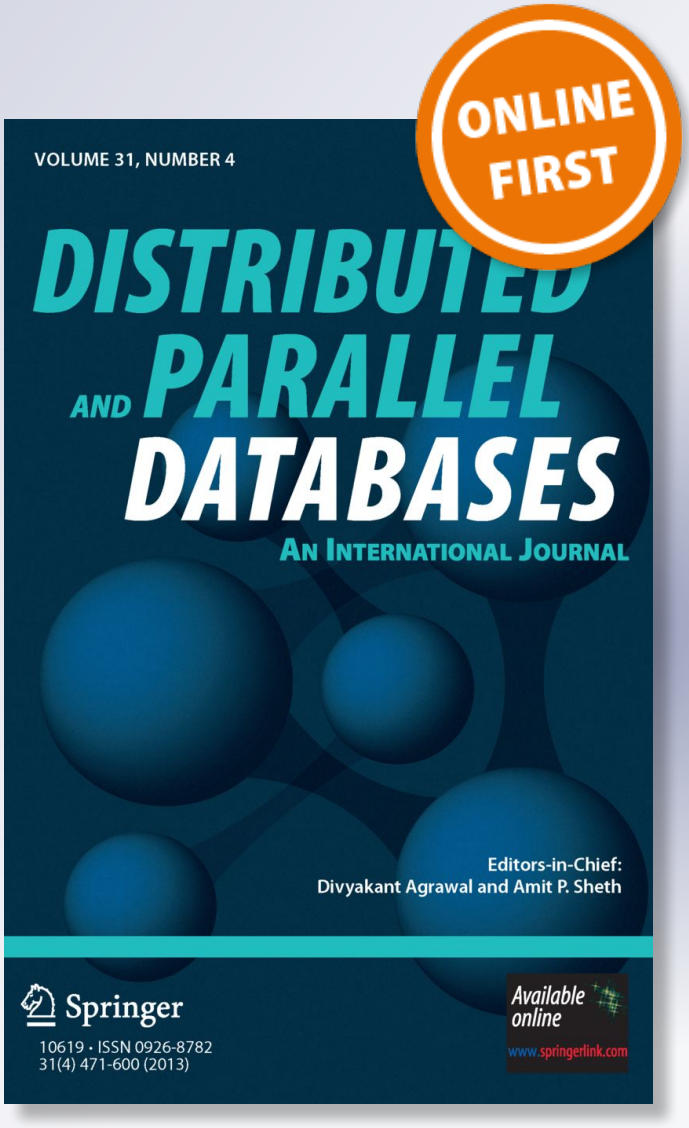
Resolving Entity on A Large scale: DEtermining Linked Entities and Grouping similar Attributes represented in assorted TErminologies

K. A. Vidhya & T. V. Geetha

Distributed and Parallel Databases
An International Journal

ISSN 0926-8782

Distrib Parallel Databases
DOI 10.1007/s10619-017-7205-1



Springer

Distrib Parallel Databases
DOI 10.1007/s10619-017-7205-1

Resolving Entity on A Large scale: DEtermining Linked Entities and Grouping similar Attributes represented in assorted TErminologies

K. A. Vidhya[1] · **T. V. Geetha**[1]

Abstract The tremendous growth of the World Wide Web (WWW) accumulates and exposes an abundance of unresolved real-world entities that are exposed to public Web databases. Entity resolution (ER) is the vital prerequisite for leveraging and resolving Web entities that describe the same real-world objects. Data blocking is a popular method for addressing Web entities and grouping similar entity profiles without duplication. The existing ER techniques apply hierarchical blocking to ease dimensionality reduction. Canopy clustering is a pre-clustering method for increasing processing speed. However, it performs a pairwise comparison of the entities, which results in a computationally intensive process. Moreover, conventional data-blocking techniques have limited control over both the block size and overlapping blocks, despite the significance of blocking quality in many potential applications. This paper proposes a Real-Delegate (Resolving Entity on A Large scale: DEtermining Linked Entities and Grouping similar Attributes represented in assorted TErminologies) that exploits attribute-based unsupervised hierarchical blocking as well as meta-blocking without relying on pre-clustering. The proposed approach significantly improves the efficiency of the blocking function in three phases. In the initial phase, the Real-Delegate approach links the multiple sets of equivalent entity descriptions using Linked Open Data (LOD) to integrate multiple Web sources. The next phase employs attribute-based unsupervised hierarchical blocking with rough set theory (RST), which considerably reduces superfluous comparisons. Finally, the Real-Delegate approach eliminates a redundant entity by employing a graph-based meta-blocking model that represents a

✉ K. A. Vidhya
avidhya06@gmail.com

T. V. Geetha
tv_g@hotmail.com

[1] Department of Computer Science Engineering, Anna University, Chennai, India

Published online: 09 September 2017

redundancy-positive block and removes overlapping profiles effectively. The experimental results demonstrate that the proposed approach significantly improves the effectiveness of entity resolution compared with the token blocking method in a large-scale Web dataset.

Keywords Big data · Web data · Entity resolution · Linked Open Data · Hierarchical blocking · Rough set theory · Meta-blocking · Query processing

1 Introduction

Over the past decade, the popularity of Web data publishing has generated an enormous volume of heterogeneous data. This massive generation of Web data on an unprecedented scale potentially creates Big data. To realize the prospects of Big data, integration of the massive volume of heterogeneous data is essential [1]. Due to the loose schemata of Web data, the unification of similar entity profiles is a prerequisite for many potential applications. Web content providers desire to employ their own schema and entity identifiers rather than accessing and re-using preexisting content using Linked data principles. This content leads to overlapping and heterogeneous data on the Web that adversely degrades the quality of Web entities. Thus, entity resolution (ER) is paramount for identifying the data objects that represent the same entities [2]. Matching the similarity of entities in a large dataset is a computationally challenging task. To cope with the large volume of data that scales quadratically, data-blocking is the prominent method to enhance the quality of query processing by restricting the number of pairwise comparisons. This process is executed by segregating the overlapping entities and performing entity comparisons exclusively within a block.

The majority of the existing blocking techniques assume that the schema and its features are well known in advance. The schema and qualitative features are the essential requirements for selecting the most distinctive attributes that facilitate the key value-based entity assignment to the blocks. However, traditional blocking methods are incompatible with heterogeneous Web data. To overcome this incompatibility, some blocking methods concentrate on controlling the block size of the ER. Hence, conventional data-blocking techniques only partially ameliorate the ER from the perspective of mitigating the overlapping blocks [3,4]. Hierarchical clustering is the ideal way to limit both the block size and generation of overlapping blocks and to ease the dimensionality reduction in Big data. Regardless of the control of the trade-off between block quality and block size in the clustering process [5,6], the existing hierarchical blocking methods need to apply pre-clustering, such as canopy, for further dimensionality reduction in de-duplication. Iteration of the overall co-occurring entities in the associated blocks makes the ER computation process an arduous task. Another problem that requires a focus on dimensionality reduction is pairwise comparisons that induce tolerable redundant and superfluous comparisons.

The existing blocking techniques employ a pre-clustering method that consumes additional time and cost and likely increases the complexity of the entire process. Moreover, the meta-blocking techniques restructure the blocks using token blocking and generate many blocks that lead to overlapping across multiple contexts. Hence, optimal reconciliation of large data by resolving the entities and eliminating super-

fluous and redundant comparisons is a crucial process. To address pre-clustering and token blocking issues, the Real-Delegate approach presents a novel blocking method that considers the attribute-based pairwise comparisons in unsupervised hierarchical blocking. Linked Open Data (LOD) is the most popular method to expose and interlink similar entities. The Real-Delegate introduces the optimal attribute selection using the Rough Set Theory (RST) with the blocking concept. Instead of entire attributes, the optimality in attribute blocking using RST reduces superfluous comparisons. The integration of generic meta-blocking in the ER system is inevitable, regarding the recall, but not adequate for precision. The novelty of Real-Delegate in meta-blocking is to avert the redundant comparisons of the domain values for unique profiles, instead of replicating all the domain values and to maintain the recall without degrading precision.

The main contributions of the Real-Delegate are as follows:

- The proposed rough set attribute-based unsupervised hierarchical blocking exploits an optimum set of attributes avoiding the pre-clustering algorithm there by improving the process overhead and thereby ensuring the maximum effectiveness while resolving the schema-agnostic Web entities of the entity resolution framework.
- The optimal attributes selection proposed by the Real-Delegate approach represents the context of large-scale Web entities with the reduced set of attributes, which improves the efficiency of resolving the heterogeneous web entities.
- The proposed map-reduce based intelligent meta-blocking scheme further reduce the search space while block processing in the Web data there improving the time complexity of the ER system.
- By providing the optimized non-linear query comparison structure, the Real-Delegate maintains the trade-off between the effectiveness and efficiency of the ER system, thereby achieves highly scalable ER system.

2 Related works

The ER is a traditional problem in the computer science field, and numerous ER methods have been proposed to tackle the shortcomings in the large-scale Web environment [7]. Variants of ER techniques are predominant and include string similarity, similarity of transformations, and entity relationship. Among the conventional ER technologies, blocking is one of the essential techniques that scales ER to a large dataset and reduces the search space.

2.1 Entity matching techniques

Entity matching is also known as record linkage, duplicate identification, reference reconciliation, and entity resolution and plays a crucial role in data cleaning and integration. Several existing studies [8,9] independently examine the blocking methods of the entity matching techniques by analyzing the entity profiles that are either identical or not.

The supervised and unsupervised learning approaches are frequently applied in the ER process. Most of the works employ learning-based ER methods that require

entity pairs as the training data to determine the class of the match. The supervised learning-based ER approaches [10] fail to deliver effective results for large-scale ER problems due to the lack of training data for the evolving topics. To mitigate the burden of entity pair labeling, the entity-matching approaches focus on unsupervised learning and employ explicit relations to inherently identify the relationships. The work in [6] employs the Blocking Key Value (BKV) and sorted neighborhood method, respectively, to generate blocks of similar entities within a specific size and dynamically update the BKV for real-time ER. The BKV-based ER technique groups the key value-based entities in each group in which the values depend on the selected attributes. Further, the sorted neighborhood method arranges the input records in an order based on its BKV along with the fixed window size. Even though this framework produces the blocks with the satisfaction of size constraints, the variation in the attribute values affects the quality of the sorting key-based matching. Due to the heterogeneity of Web sources, manipulation of inconsistent and incomplete data is necessary during ER. A statistical approximation method in [11] explores the different types of entity descriptions in Uniform Resource Identifier (URI) that include infixes, attributes, and predicates to accomplish the effective ER of large-scale Web data. This method is based on the relational machine learning model of an RDF by exploiting the contextual information and measuring the structural similarity of entities. Even though this part of the ER process improves the performance, the sequential steps of coupled matrix and tensor factorization extend the processing time.

2.2 Blocking techniques

The StringMap method [12] is one of the major mapping-based blocking techniques. It maps the key of each record with the multidimensional Euclidean space and then possibly groups the similar objects into the same cluster. Canopy clustering [13] is one of the unsupervised pre-clustering algorithms, which employs a random tuple selection method. It creates a canopy and generates blocks using a distance metric, but it leads to the generation of overlapping blocks. Later, the character-based q-grams method has been utilized to precisely map identical entities by applying the similarity between the two strings [14]. Then, the researchers focus on the token-based n-grams method [15] that measures the similarity between each token of two strings to match the entities. For instance, consider that ‘John Smith’ is a string in which ‘John’ and ‘Smith’ are the tokens. To increase the accuracy of blocking, the existing studies focus on the suffix array blocking method [16]. The suffix array blocking method matches similar objects based on the suffix of each string in which the blocking method needs to assign the suffix length parameter for entity matching process. As a result, it facilitates the identification of dissimilarities between the two strings such as ‘John Smith’ and ‘John Snith’. The iterative blocking in [17] reflects the results of the ER to other blocks, which substantially diminishes the processing time of other blocks in generating repeated record matches. However, the aforementioned mapping-based blocking methods are not suitable for handling heterogeneous Web entities due to the assumption of data that pertain to the same schema.

Though the approach in [5] examines the clustering complexity regarding the size constraint in the small dataset, it fails to provide the benefits of size-based clustering. In

addition, since this approach performs the entire pairwise comparison, it becomes an arduous task while blocking in large-scale datasets. The spectral neighborhood (SPAN) approach in [18] employs unsupervised learning of hierarchical clustering to mitigate the number of comparisons in each block that is based on the size constraint and determines similar entities that are based on the neighborhood relation. However, the performance degrades with the increasing size of the dataset, especially when increasing the number of duplicates. The CBLOCK approach in [19] automatically blocks large-scale datasets and considerably reduces the complexity of the de-duplication process by exploiting hash function and applying the splitting and merging methods. However, the use of a labeled training data is impractical for all cases, especially the Web data. To have an effective and automatic key selection, the work in [20] employs an unsupervised learning method. However, this technique is not efficient enough to handle the heterogeneous Web entity collections with the small block size. In Big data framework, dimensionality reduction that includes selection and extraction of features is crucial to diminishing the search space to identify the optimal features. The rough set theory-based feature selection approach [21] selects an important feature known as 'reduct' from the Wisconsin breast cancer dataset. Using such reducts, further processes are carried out along with the support vector machine (SVM), which averts the redundant features in the results and, consequently, improves diagnostic accuracy. The hybrid decision-support system [22] exploits the rough set theory and extreme learning machine to select the reducts in the hepatitis disease dataset. However, such learning-based selection of reducts is not suitable for large-scale Web data because it is likely to mislead the decision support system.

The homogeneity and a-priory schema-based blocking techniques are inappropriate for heterogeneous and schema-agnostic Web data. The schema-agnostic blocking is independent of the input schema, which relies on the attribute values to handle the heterogeneous information spaces. For instance, the Resource Description Framework (RDF) data-based semantic indexing [23] reduces the false hit rate and significantly improves the real hit rate. The token blocking-based efficient entity resolution (TBEER) approach [24] uses an intelligent attribute-agnostic blocking mechanism to present a scalable solution. However, this method creates blocks with a high level of redundancy because it considers a single token and disregards the associated attribute. In addition, it leads to performance issues because it involves additional comparisons in the overlapping blocks. Later, TYPiMatch [25] learns keys and key values to determine the ideal representative attributes. It improves the effectiveness as well as efficiency of the ER by learning both the subtype and subtype-specific key values in unsupervised blocking. Even though these methods reduce the number of comparisons, this reduction is necessary for achieving the reduction of oversized blocks in a reasonable time. To do so, the URI semantics blocking method in [26] significantly diminishes the number of comparisons by exploiting entity identifiers and relationships between entities while building the blocks. This Prefix-Infix-Suffix (PIS)-based blocking method also maintains the effectiveness of blocking in the Web data. The HAshed RecoRd LinkAge (HARRA) approach [27] is the Locality-Sensitive-Hashing (LSH)-based record linkage method that improves speed of large data collection. Even though the LSH method provides high scalability in the homogeneous information spaces, it has to deal with the heterogeneous web entities.

2.3 Block processing techniques

To further enhance the generated blocks and discard the additional comparisons across the blocks, the existing methods introduce different block processing techniques. The unsupervised block processing techniques focus on the optimal block sizes; the techniques include meta-blocking [28], comparison propagation [29], and clustering method [6]. Meta-blocking transforms a set of blocks into a new set of blocks with reasonable minimum comparisons. The comparison propagation method discards all redundant comparisons that overlap the entities in a single block rather than in multiple blocks and, thus, reduces the search space. The unsupervised graph-based meta-blocking method [30] partially overcomes the constraint in balancing precision and recall through the identification of the likelihood of profiles based on the edge weight of the node in which each node refers the entity profile. Several ER approaches have employed an iterative method [17] and a supervised method [31] to optimally reduce the repeated comparisons. The supervised method-based block processing attempts to accurately identify the non-matching edges by exploiting the training data as the composite pruning models. Although the current ER research considerably resolves Web entities, accomplishing an effective and efficient ER is still a major constraint in the Web data.

Entity resolution of the highly heterogeneous information spaces (ER-HHIS) approach in [32] targets to reduce the unnecessary pairwise comparisons by focusing on the effectiveness and efficiency layers. Utilizing block pruning and purging methods, it considerably reduces the number of pairwise comparisons. However, the performance of this approach is not as efficient when compared to the Real-Delegate approach. By exploiting the hierarchical clustering and meta-blocking techniques, the Real-Delegate approach effectively utilizes the optimal attributes alone and, thus, significantly reduces the number of pairwise comparisons. The benchmarking work [33] provides the MapReduce solutions for the token blocking method to reduce the processing complexity. Even though it reduces the processing time during clustering, it leads to the imprecise results since it fails to create a separate block for two different contexts when there is a token that matches two contexts of a lexical pattern. However, the Real-Delegate approach diminishes the search space by applying the MapReduce function on the generated blocks. Its hierarchical clustering method facilitates the block processing in the descending order of significant attributes, which effectively produces the results in a reasonable time when searching the information in the Web database. The paper [34] applies the MapReduce for parallelized meta-blocking to reduce the superfluous and redundant comparison. Though this approach considerably reduces the overhead and MapReduce jobs, it lacks efficient cleaning of overlapping profiles. Hence, the Real-Delegate approach is being employed to exploit the MapReduce technique to clean the overlapping profiles, and by utilizing the metadata of each block hierarchically, effective query processing can be accomplished. For instance, if metadata of the first block do not match the query, the Real-Delegate system ignores the sequential sub-blocks, where the first block is the most significant attribute-based block. The work [35], which is an extended work of [34], differs in the implementation aspects and marginally improves the time efficiency. The paper [36] merely considers the prefix, infix, and suffix in its original lexical structure. The

different resources comprise a property table in different alignment such as 'Birthdate' and 'date of birth'. The Real-Delegate approach employs the URI in the decomposed forms of prefix, infix, and suffix to identify the relevant URI links that appearing in the variety of Web resources. Hence, it utilizes Linked Open Data (LOD) to determine the similar URI links by considering the names of the attributes.

3 Problem formulation and overview of Real-Delegate approach

The core of entity resolution is to describe real-world objects. A collection of web URI (E) includes entity profiles. Each profile $p<A, V>$ is uniquely divided into a set of attribute-value pairs using LOD, and those sets of pairs describe a real-world object. The Real-Delegate includes blocking and meta-blocking with comparison reduction.

Definition Given E, the aim of Real-Delegate is to group all profiles $\in$ E into the block, and these profiles describe similar real-world objects so that the comparisons of co-occurring profiles are limited. However, each profile 'p' consists of a vast number of attributes (M), which creates extensive computational complexity. The novelty of the Real-delegate system is that it can extract the optimal attributes ($N \in M$) that have a maximum dependency with a real-world object. Compared with the token blocking, the Real-Delegate reduces the computational complexity from $O(M \times p) - O(\{M - N\} \times p)$. Since, the traditional token blocking groups all similar attribute values into blocks, regardless of the associated name of the attribute, it produces overlapping blocks with additional comparisons.

3.1 Blocking

To illustrate the functionality, consider the web entity profiles ($p = 4$) in Fig. 1, where A1 to A4 represent the attributes and V1 to V4, the values of web entity profiles. To select an 'N' number of optimal attributes, the Real-Delegate measures the inter-dependency score between attributes, which is shown in Fig. 1(b). In essence, the Real-Delegate removes the uncertain attribute values of every pair of attributes based on an equivalent relationship. In Fig. 1(a), the attribute A1 has the same value from V2 to V4; the attribute A2 has the same value from V2 to V4. The equivalence relationship of profiles {P2, P3, P4} according to attribute A1 and A2 is called a specific attribute pair. The Real-delegate discards the attributes that have a less-weighted average inter-dependency degree among the attribute pairs. Fig. 1(c) eliminates the attribute 1 for comparison reduction. Despite the uncertain attribute pair reduction using the inter-dependency measure, the ER system is not adequate without extending the comparison reduction in a block to its concurrent attributes. Ordering the attributes with hierarchical blocking is of paramount importance for extending the ER.

Most existing efforts on ER systems only focus on comparison reduction in a block. These schemes have conducted the comparisons of O(Blocks), regardless of attributes. Nonetheless, there is a possibility to considerably reduce the unnecessary comparisons in the concurrent attributes using the hierarchical blocking schematic representation. Figure 1(d) illustrates the attribute linkage regarding different profiles. The attribute

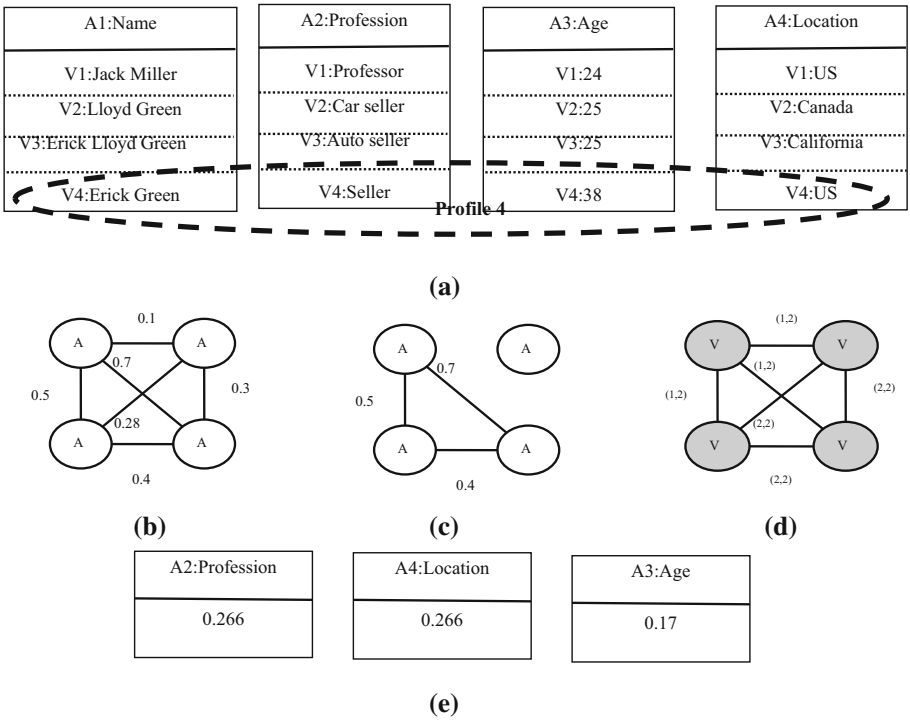


Fig. 1 Blocking concept in Real-Delegate approach

linkage is measured in terms of intra-dependency measure for attribute A2, where (V1, V2) are the non-matching values represented as the (1,2) domain values. Moreover, the matching values of V2, V3, and V4 are covered under domain value 2. The total domain value of A2 is 2, and the indiscernibility score is 0.266 [AVG(0.5,0.7,0.4)/2)]. According to this score, the attributes A2, A4, and A3 are arranged in order. The values of entity profiles ({V1},{V2, V3, V4}) have two domains according to A2. For a domain value of {V2,V3,V4}, A3 has ({V2,V3},{V4}) two domains. This schematic extension up to the Nth attribute builds hierarchical blocking and reduces the comparisons from O(Blocks) to $\{O(Blocks_A) - O(Blocks_i\}$, where 'I' represents the attribute of a non-matching profile value. The Real-Delegate does not apply the comparison for blocks in subsequent other attributes.

3.2 Meta-blocking

The following metrics are the important performance measures.

(i) *Precision* It considers the true positives of matching comparisons and false positives of superfluous and redundant comparisons. Precision = $|P(B)|/||B||$, where $|P(B)|$ and $|B|$ represent the number of similar entity profiles and blocks, respectively.

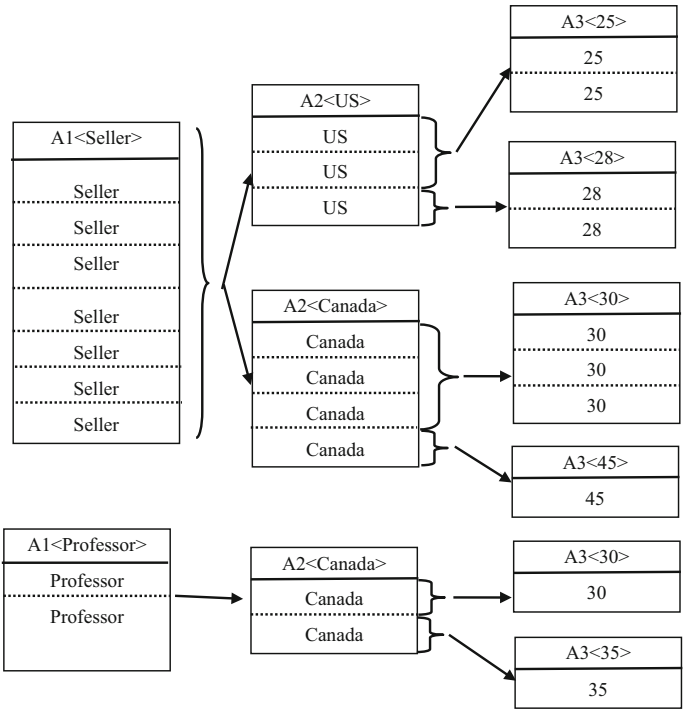


Fig. 2 Meta-blocking in the Real-Delegate approach

(ii) *Recall* Number of existing duplicates in B. Formally, recall = |P(B)|/|P(E)|, where P(E) represents the set of all similar entity profiles.

Most meta-blocking studies place every entity profile into multiple blocks and make a new set of blocks for emphasizing recall. However, this reduces precision due to the placement of every pair of duplicate profiles in a separate block. This factor collapses the blocking schema and escalates the unnecessary comparisons while retaining the better recall.

The Real-Delegate reduces the unnecessary comparisons without degrading the precision and recall by restructuring a block collection, which is the redundancy-positive blocks. Meta-blocking splits the domain values of an attribute with respect to the preceding attribute and compares the values of the blocking graph in its subsequent attributes, only when the preceding attribute leaves the matching profiles in its block. This algorithm discards the comparison of succeeding attribute values that correspond to a non-matching preceding attribute of the same profiles. Figure 2 illustrates the meta-blocking of Real-delegate approach. According to the blocking concept, nine entity profiles with three attributes are arranged hierarchically. The attribute A1 has two blocks with the domain values of ‘seller’ and ‘professor’. If a query matches the value ‘seller’, it encourages to compare the values of A2 ‘US’ and ‘Canada’. Otherwise, the A2 and A3 values for the corresponding profiles remain uninvolved in the comparisons.

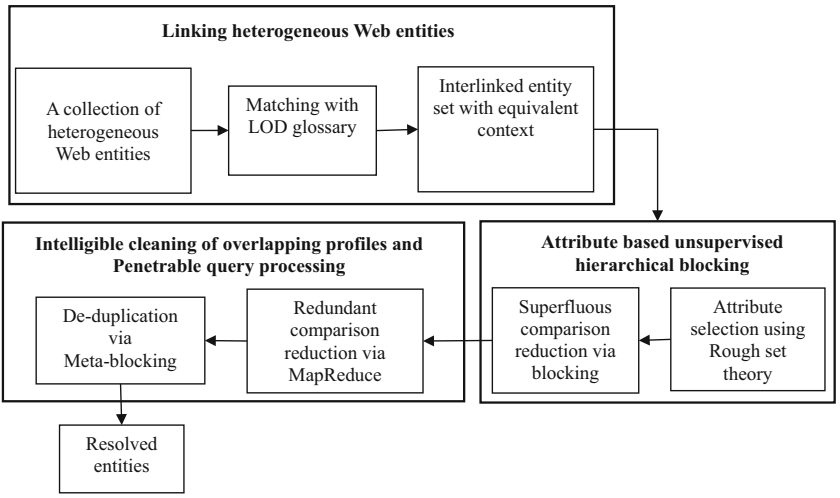


Fig. 3 The Real-Delegate methodology

Instead of replicating the profiles, the proposed meta-blocking technique allows the duplication of domain values, only when the unique preceding attribute values are present. For example, the domain value of 'age 30' is replicated in both 'seller' and 'professor' in the Real-Delegate. Thus, the Real-delegate effectively manages the Web entities and ensures search space reduction without degrading the precision and recall.

4 Real-delegate methodology

Effective Big data management is necessary to reconcile identical heterogeneous entities and discard the duplicates. In large Web data, reducing the query delay is one of the critical processes because there are occurrences of duplicate files and superfluous and redundant comparisons. The conventional ER methods de-duplicate the entities by averting the redundant pairwise comparisons. However, it fails in eliminating the superfluous comparisons. To remedy such a constraint, this work suggests a novel attribute-based blocking technique together with meta-blocking that clusters similar entities to facilitate effective query processing in Web data. Figure 3 shows the overall process of the Real-Delegate methodology.

Linking heterogeneous sources: The Real-Delegate approach determines a set of homogeneous entities or an equivalent entity in the Web data using the LOD source. LOD is one of the knowledge bases that comprises a huge number of knowledge sources such as GEOname, DBpedia, and DrugBank. LOD exploits the URI of each entity description and links the input entities with desired knowledge source to generate a set of relevant entity descriptions that carry similar context or attributes. The mapping phase of LOD produces unique entities, where each linked entity set

has a unique set of attributes. Section 4.1 discusses the detailed description of this phase.

Attribute-based unsupervised hierarchical blocking: The Real-Delegate approach judiciously selects the optimal blocking key that is based on the rough set theory and various criteria of the key. According to the optimal attribute selection, the Real-Delegate measures the indiscernibility score for each attribute to hierarchically cluster similar entities into blocks and, thus, efficiently reduces the superfluous comparisons while querying the Web pages. Section 4.2 presents the detailed procedure for generating the hierarchical blocks.

Intelligible cleaning of overlapping profiles and penetrable query processing: The Real-Delegate approach mitigates both the redundant comparisons and duplicate entities using a graph based meta-blocking. The graph based meta-blocking utilizes the domain value of an attribute to tag the meta information of the block, which enables effective mapping of those entities that appear exclusively in the corresponding blocks, and its sub-blocks, which averts repeated mapping of the redundant entities. This reduces the search space and facilitates penetrable query processing. The meta-blocking of the Real-Delegate approach is described in Sect. 4.3.

4.1 Linking heterogeneous Web entities

The Real-Delegate approach employs LOD to identify and link the related entities among the voluminous Web data sources. LOD is the largest and freely available knowledge base that contains multiple knowledge sources [37]. It contains data, which are updated by a variety of sources, and includes different domain information such as books, films, companies, online communities, scientific publications, clinical trial, statistical, scientific data, and music. To ease the LOD matching process, the Real-Delegate system utilizes the URI of the entity collections in the Web data instead of using the raw data of entity descriptions. The URI incorporates three components such as the Prefix, Infix, and Suffix parts. The Prefix part refers the information about a specific domain; the Infix part indicates the local identifier; and the Suffix part represents either the format details, such as .rdf, or the named anchor.

The Real-Delegate approach builds the representation model for unified heterogeneity of Web data by employing three steps, as shown in Fig. 4. Initially, it matches the URI link of each entity description with the linked data source and generates the relevant properties along with the mapping value. The Real-Delegate approach extracts the matching evidence from entity identifiers using the LOD source. LOD matches the input URI link or keyword with its sources and provides a set of relevant URI links under a specific domain. The Web entities, which are collected from heterogeneous Web pages, include information, news, advocacy, business, and entertainment. The entity link mapping enables the creation of a list of relevant links while integrating the multiple Web sources. Second, LOD conflates the related links based on the mapping value and merges the equivalent entity descriptions of URIs from different sources. It unifies the equivalent entity descriptions from a set of relevant entity descriptions based on both the explicit and implicit equivalence relationship of mapping value. The

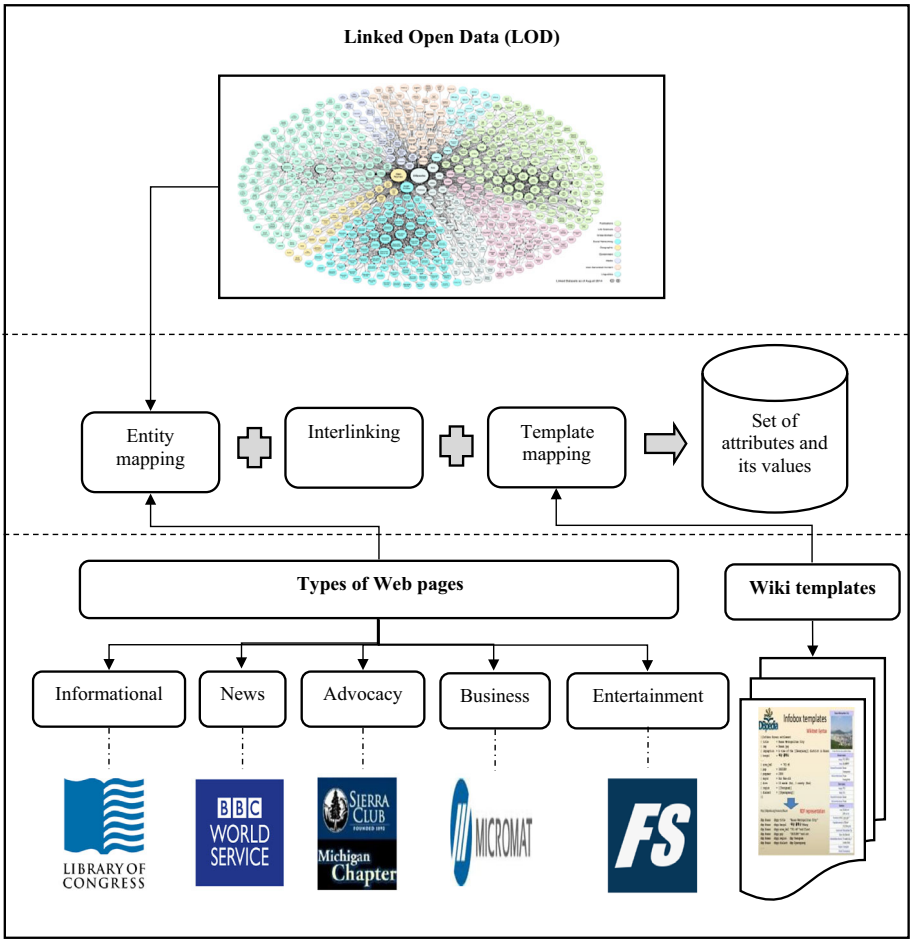


Fig. 4 Linking heterogeneous sources

mapping value provides a specific domain information for each URI to the linked data source. For instance, the mapped value refers the WordNet type retrieved from the LOD mapping phase. Thus, it creates different groups in which the features of each group can be varied with one another so that each group contains the most relevant and equivalent entity descriptions of the Web data.

Even though the linked source provides a set of relevant entity descriptions and URI information along with its corresponding main class, the received information about the URI alone is not adequate to resolve the entities in the Web data. Third, the Real-Delegate approach exploits the additional information of each entity description to differentiate the identical entity profiles instead of URI, wherein additional information represents the retrieved properties of each URI. To obtain the context-based raw value of the retrieved properties, the Real-Delegate employs the Wikipedia template information of each created domain group. This template-matching process facili-

tates the attribute-based blocking while mapping each entity with the corresponding attributes. It interlinks the properties of each entity description with the fields in the corresponding domain template. For instance, the entity descriptions in the 'music' domain group are linked to the attributes in the 'music' template.

4.2 Attribute-based unsupervised hierarchical blocking

Conventional ER techniques employ canopy as a pre-clustering step for unsupervised hierarchical blocking to increase the processing speed of large-scale real-world datasets. However, this is likely to provide compute-intensive results because it captures merely pairwise comparisons and needs an additional clustering method to ensure precise block generation. The measurement of inter and intra-dependency of attributes leverages the hierarchical blocking with substantial processing speed. The Real-Delegate approach exploits an attribute-based unsupervised hierarchical blocking to attain an effective dimensionality reduction. It targets the most reliable and distinguishable attribute selection to precisely conflate similar entities into blocks. The dependency relationship between the attributes paves the way to select an optimal attribute as block key that has high dependency value across entire attributes. Thus, it groups the large-scale data set into blocks. However, single level blocking is hard to scale for the Web data. Therefore, there is a need for extending the clustering level to maintain better trade-off between ER quality and block size. Single attribute-based blocking provides only a minimum number of blocks and partially resolved entities, whereas the entire attribute-based blocking generates the highly resolved entities and blocks. Thus, the Real-Delegate approach decides an optimal set of attributes and applies the hierarchical blocking, which is based on the dependency degree of selected attributes, and conflates the relevant entity descriptions or entities. This process considerably reduces the dimensionality in the clustered Web data.

4.2.1 Rough set theory-based attribute selection

The selection of an attribute using RST [21,22,39,40] and clustering of the possibly overlapped entities into a single block considerably alleviates the comparison of large-scale Web data. The RST measures the indiscernibility relation between the attributes and determines the optimal set of attributes for hierarchical blocking. The indiscernibility relation reveals the equivalence relation of entities according to each attribute. The RST considers the domain value of each attribute as the blocking key criteria. For instance, the attribute 'job' has several domain values such as seller, dealer, designer, and developer. These key criteria are used to determine the indiscernibility relation between the pair of attributes to select the optimal blocking key. RST measures the indiscernibility relation, which involves the inter and intra-dependency degree, and identifies the equivalent class of entities to entire that attributes use the attribute key criteria, as shown in Eq. (1). $V(Ai_1)$ and $V(Ai_2)$ implies the value of entity collections of records '1' and '2' under attribute 'A_i'.

$$\text{Inter}\,(A_i) \;=\; \{(Ai_1, Ai_2) \in R \times R \,|\forall A_i, V\,(Ai_1) \;=\; V\,(Ai_2)\} \tag{1}$$

The Real-Delegate RST-based optimum set of key selection involves three processing steps such as defining key criteria in each attribute, determination of the inter-dependency degree of each attribute, and selection of the optimal set of highly dependent attributes. First, the Real-Delegate approach considers the domain values to define the key criteria. Second, it determines the inter-dependency degree of each attribute using the key criteria.

$$\begin{aligned} &\text{Let}, A_i \in A; X_i \subseteq R_i; \left(R_i \in A_i \& R_j \in A_j\right) = k_{ij} \in K; \\ &\qquad [x]_{k_{ij}} \subseteq V\left(R_i\right), \text{where 'ij'} \text{ varies from 1 to } i * j \\ &U\_V\left([x]_{k_{ij}}\right) = \left\{ x \in R \,\middle|\, [x]_{k_{ij}} \subset \left(X_{Ri} \cup X_{Rj}\right) \neq \{\} \right\} \qquad (2) \\ &L\_V\left([x]_{k_{ij}}\right) = \left\{ U\_V\left([x]_{k_{ij}}\right) \right\} - \left\{ R\left([x]_K - [x]_{k_{ij}}\right) \right\} \qquad (3) \\ &\text{Pos\_V}\left([x]_{k_{ij}}\right) = \bigcup_{k_{ij} \in X_i} L\_V\left([x]_{k_{ij}}\right) \qquad (4) \end{aligned}$$

where $[x]_{kij}$ represents a set of entity descriptions, in which kij denotes the pair of domain value that belong to A_i and A_j. 'K' indicates the domain values of all feasible pairs. The upper approximation consists of entity descriptions in a specific class 'V' of two consecutive attributes ($[x]_{kij}$), which is denoted as $U\_V([x]_{kij})$. The lower approximation ($L\_V([x]_{kij})$) incorporates a set of entity descriptions of the information table, excluding the uncertainty of entity descriptions belonging to value 'V' of a specific domain ($[x]_{kij}$). The positive region ($Pos\_(V([x])$) refers a set of entity descriptions that appear in all classes of the lower approximation set in [x]. Eqs. (2–4) reveals the underlying requirement of the inter-dependency degree measurement to select the optimal set of attributes using the RST. Equation (5) yields the inter-dependency degree between the attributes that are based on the uncertainty value of equivalent classes, which are obtained using Eq. (3). $\theta(X_i)$ denotes the value of X_i, which is the total number of domain values in attribute 'X_i'. 'n' and 'b' represent the number of domain values in X_i and X_j, respectively. In Eq. (6), $|\theta(X_{in})|$ refers to the number of records or entities in the corresponding n^{th} domain value of the X_i attribute.

$$\begin{aligned} &\text{Dep}_{\text{inter}}\left(A_i, A_j\right) = -\sum_{n=1}^{|\theta(X_i)|} P_n * \sum_{b=1}^{|\theta(X_j)|} Q_b \qquad (5) \\ &P_n = \left(\frac{|\theta\left(X_{in}\right)|}{X_i} \right) \qquad (6) \\ &Q_b = \left(\frac{\left|L\_V\left([x]_{k(nb)}\right)\right|}{|X_{in}|} \right) \log \left(\frac{\left|L\_V\left([x]_{k(nb)}\right)\right|}{|X_{in}|} \right) \qquad (7) \end{aligned}$$

Third, the Real-Delegate approach analyzes the attributes based on the degree of dependency values. It arranges the attributes of the weighted average dependency degree in descending order and, using Eq. (8), determines a blocking attribute that has a high dependency degree. In Eq. (8), $|\{\text{Dep}_{\text{Inter}}(A_i)_H\}|$ and $|\{\text{Dep}_{\text{Inter}}(A_i)_L\}|$

Table 1 Sample of the information system

	A1	A2	A3
R1	$A1_1$	$A2_1$	$A3_1$
R2	$A1_1$	$A2_1$	$A3_1$
R3	$A1_1$	$A2_2$	$A3_2$
R4	$A1_1$	$A2_2$	$A3_2$
R5	$A1_2$	$A2_3$	$A3_2$
R6	$A1_2$	$A2_3$	$A3_2$
R7	$A1_2$	$A2_3$	$A3_2$
R8	$A1_2$	$A2_3$	$A3_1$
R9	$A1_2$	$A2_1$	$A3_1$

represent the number of high and low inter-dependent attributes of A_i; 'α' and 'β' denotes the weighted parameters, i.e., $(\alpha + \beta = 1)$ and $\alpha > \beta$; $|\{Z\}|$ represents the average number of entities excluding uncertainty (σ) across the pair of attributes. The threshold value divides the high and low inter-dependent values of each attribute. This inter-dependency degree measurement helps find an optimum set of attributes.

$$\mathrm{Dep}_{\mathrm{inter}}(A_i) = \frac{\alpha * \left|\left\{\mathrm{Dep}_{\mathrm{inter}}(A_i)_H\right\}\right| + \beta * \left|\left\{\mathrm{Dep}_{\mathrm{inter}}(A_i)_L\right\}\right|}{|\{Z\}|}$$
$$\text{where } Z_{Ai} = \sum_{j=1}^{n} (R - \sigma)_{A_{ij}} \tag{8}$$

Most conventional attribute selection methods contemplate only the inter-dependency value of attributes in the dependency degree measurement. However, the system tends to generate many blocks and extracts inaccurate inherent information from the blocks. This process leads to redundant and superfluous comparisons while searching the results for the query. Hence, the Real-Delegate approach incorporates the Indiscernibility relation between attributes in the optimal set of key selection, instead of considering the inter-dependency value of attributes alone and, thus, it improves the efficiency of blocking in the Web data. Indiscernibility relation implies that the inter-dependency and intra-dependency value of the attributes contemplated together.

For instance, Table 1 shows the definite number of entity descriptions in rows and attributes in the column including corresponding cell values, as a two-dimensional representation. Consider, the 'Music artist' domain information is as shown in Table 1, in which A1 is age, A2 is occupation, and A3 is instrument. $A1_1$= less than 30 and $A1_2$= greater than 30. $A2_1$=Singer, $A2_2$=composer, and $A2_3$=Song writer. $A3_1$=Violin, and $A3_2$=Piano.

From the abovementioned steps, Eqs. (2) to (4) indicate the parameters that are required for selecting the optimal set of attributes, which are based on the RST. The common entities that belong to each domain value are retrieved for each pair of attributes. Eqs. (5–7) determine the inter-dependency degree between the attributes that are based on the domain values in each attribute. 'ω_A' represents the inter-dependency

score of each attribute, which is measured using Eq. (8). Finally, the Real-Delegate approach arranges the attributes that are based on the threshold value of '$\gamma(\omega_A)$', creates the hierarchical blocking structure. Algorithm.1 illustrates the steps for an optimal set of attribute selection using RST.

Step 1: Upper_$(V([x]_{k1})_{A1=A11}$ = {R1, R2, R3, R4} ; Upper_$(V([x]_{k1})_{A1=A12}$= {R5, R6, R7, R8, R9} Upper_$(V([x]_{k1})_{A2=A21}$ = {R1, R2, R9} ; Upper_$(V([x]_{k1})_{A2=A22}$ = {R3, R4}; Upper_$(V([x]_{k1})_{A2=A23}$ = {R5, R6, R7, R8}
Step 2: Lower_$(V([x]_{k1})_{A2=A11}$ = {R1, R2, R3, R4} ; Lower_$(V([x]_{k1})_{A2=A12}$ = {R5, R6, R7, R8}
Step 3: Pos_$(V([x]_{k1})_{A1}$={R1, R2, R3, R4, R5, R6, R7, R8}
Step 4: Using Equation(5),
$Dep_{Inter}(A_1, A_2)$={(R1, R2), (R3, R4), (R5, R6, R7, R8)}
$Dep_{Inter}(A_1, A_3)$={(R1, R2), (R3, R4, R5, R6, R7), (R8)}
$Dep_{Inter}(A_1) = \omega_{A1}$ using equation (8)
Step 5: Repeat Step 4 for 'A_2' and 'A_3
$Dep_{Inter}(A_2) = \omega_{A2}$; $Dep_{Inter}(A_3) = \omega_{A3}$
Step 6: Sorting list = OA ⊆ A => (OA > γ (ω))

Algorithm 1: Steps for selecting the optimal set of attributes using RST

4.2.2 Indiscernibility relation ensures unsupervised blocking

The optimal set of key-based blocking facilitates de-duplication and accomplishes the Web data normalization through the reduction of pairwise comparisons. The Real-Delegate blocking method merely focuses on the large-scale data clustering without exploiting the either completely or partially trained information. Hence, this blocking is termed as the unsupervised blocking method. The Real-Delegate approach concentrates on the inter-dependency as well as intra-dependency value to restrict a large number of feasible pair-wise comparisons. The selected optimum sets of attributes are further sorted via intra-dependency, which represents the number of domain values in each attribute. The sorted attributes with respect to the dependency relationship in clustering facilitate the partitioning of the massive Web data collection while ensuring the dimensionality reduction. Equation (9) rearranges a set of attributes by jointly considering the inter-dependency and intra-dependency.

$$\text{IND}(A_i) = \frac{\text{Avg.Dep}_{\text{Inter}}(\text{OA})}{\text{Dep}_{\text{Intra}}(A_i)} \quad \text{where OA} = \{OA_1, OA_2, \ldots, OA_N\} \tag{9}$$

where OA represents a set of optimum attributes, and 'N' refers to the total number of optimum attributes for blocking. The optimal blocking key or IND relation of the attribute is the ratio between the average inter-dependency (Dep_{Inter}) of an optimal set of attributes and intra-dependency (Dep_{Intra}) of the corresponding attribute. The Real-Delegate approach clusters the Web entity collections that are based on the high

indiscernibility relation of the optimal key attribute, which is referred to as the first level attribute-based blocking. A further level of clustering is based on the arrangement of the optimal set of attributes from higher to lower indiscernibility relation. The indiscernibility relation value-based hierarchical clustering inherently organizes the attributes from a lower to higher intra-dependency value to avert the overlapping entities in hierarchical blocking. This arrangement indicates that the domain value count of the succeeding attribute is greater than or equal to the domain value count of its preceding attribute, which also shows that multi-level blocking is in a divisive hierarchical blocking manner. The Real-Delegate approach recursively applies the attribute clustering technique to obtain the further level of clusters until reaching the outermost attribute in the optimal set, excluding several attributes that have indefinite domain values in the arrangement of attributes. However, the generated block substantially overlaps with each other and leads the system to numerous redundant comparisons. Hence, the Real-Delegate approach suggests meta-blocking to clean the overlapping profiles and consistently reduce unnecessary comparisons.

4.3 Meta-blocking: per-block intelligible cleaning of overlapping profiles and penetrable query processing

ER is a relatively expensive and tedious process in high-dimensional Web data. The hierarchical attribute blocking method results in the redundancy-positive blocks, which indicates that there is a common sharing of blocks between two identical entity profiles. Each block leads three different types of entity comparisons, such as duplicate, redundant, and superfluous, when searching the information with coarse granularity in the Web data. Duplicate entities and redundant comparisons refer to the repeated comparisons of complete and partially matching profiles or entity descriptions, respectively. In contrast, the superfluous comparison implies the unnecessary comparison between the non-matching profiles. The Real-Delegate approach reduces only the superfluous comparisons using attribute-based multi-level blocking. To considerably reduce the dimensionality of the entity descriptions, it focuses on discarding both the redundant comparisons and duplicate entities via graph-based meta-blocking. If each block contains matching entities in a sequential order, the query processing needs to compare the matching entities linearly in the same dimension, which results in redundant comparisons. When employing meta-blocking for dimension reduction, the significant information, which is available in the generated redundancy-negative blocks, has to be preserved. The redundancy-negative block refers to the most identical entity profiles that share only one block. The graph model-based Meta-blocking maps and reduces the linear comparisons of matching entities in each attribute via the conversion of sequential entity descriptions into different ranges that facilitate restricted non-linear comparisons.

4.3.1 Entity de-duplication and redundant comparison reduction via meta-blocking

In the presence of Big data, the traditional token-based meta-blocking is an arduous task because it necessitates the creation of an exclusive block for every distinct token

in the Web data. To accomplish such constraint, the Real-Delegate approach employs attribute-based unsupervised blocking instead of traditional token blocking to generate a minimum number of blocks. Even though such blocks restrict superfluous comparisons, the redundant comparisons and duplicate entities have to be effectively discarded through the tagged metadata of each block to ensure considerable reduction. Metadata facilitates the optimized query processing through tagging of relevant criteria in each block according to the sequential entity IDs, which are represented in the bounded interval. The bounded interval represents the absolute difference between the IDs of the initial and final entity descriptions of each block. When a query matches the description that is found first in the block, the query pointer enters the next block through the bounded interval, and thus it initiates non-linear and non-redundant comparisons. The Real-Delegate effectively combines the domain value of the attribute and the list of entity collections in a block to generate a block ID for each block. In each block, the attributes are represented as 'A_n', and they vary from '1' to many till an optimal number of attributes is obtained. The domain value is denoted as D_A, and the list of entity IDs for D_A is $\{D_A(R_1), D_A(R_2)...., D_A(R_m)\}$, and the entities in each D_A are denoted as R_m. The proposed approach divides the entire Web data into blocks according to the D_A of an optimal set of attributes and eliminates the superfluous or non-matching comparisons. The Real-Delegate approach maintains a metadata for blocks of the optimal set of attributes based on the Bounded Interval (BI) of these attributes. For instance, the bounded interval of the defined set D_A is $\{R_1 - R_m\}$. According to the bounded interval of an optimal set of attributes, the Real-Delegate creates the non-linear query comparison model using the comparison pruning method. Consider three columns (d) with 'm' entities, and these columns contain 1, 2, and 4 blocks, respectively. A set of BI forms a group of the columns $G = \{(R_1 - R_{m/1})^1$, $((R_1 - R_{m/2})^2$, $(R_{m/2} - R_m)^2)$, and $((R_1 - R_{m/4})^3$, $(R_{m/4} - R_{m/2})^3$, $(R_{m/2} - R_{3m/4})^3$, $(R_{3m/4} - R_m)^3)\}$. By applying pruning to every 'G', the redundant comparisons are eliminated. For instance, the comparison pruning on 'G' results in $\{(R_1^1)$, $(R_1^2, R_{m/2+1}^2)$, $(R_1^3, R_{m/4+1}^3, R_{m/2+1}^3, R_{3m/4+1}^3)\}$ in terms of entities. Thus, the non-linear query comparison model of the Real-Delegate reduces the 't' number of comparisons using the evaluation technique, as shown in Eq. (10). Where y_x represents the total number of blocks in the column 'x'.

$$t = (m \times d) - \sum_{x=1}^{d} y_x \tag{10}$$

Moreover, the Real-Delegate approach precisely identifies the duplicate entities and removes them using metadata. It performs metadata matching of the entities within a specified number of attributes and facilitates the de-duplication. If a pair or set of entities in each metadata uniformly matches the other metadata of the entire attributes, the Real-Delegate approach declares those entities as duplicate entities and removes them from the Web entities.

For instance, consider a Web database comprising a set of blocks with the meta information such as 'singer', 'composer', and 'song writer' sub-blocks under the attribute of 'occupation' in the 'music' domain. The bounded interval of each sub-

block is 'singer(R1-R5)', 'composer(R6-R10)', and 'song writer(R11-R15)'. Then, the next hierarchy level, 'composer(R6-R10)' has the sub-blocks of 'violin(R6-R7)' and 'piano(R8-R10)', and 'singer(R1-R5)' has the sub-blocks of 'classical(R1-R3)', and 'rock(R4-R5)'. Similarly, other blocks are also in the form of hierarchical clustering. If a query contains a 'composer' as a keyword, the Real-Delegate approach explores the generated blocks with the intention of reducing the redundant and superfluous comparisons. When observing the mapping of a query keyword with a first entity in the 'singer(R1-R5)' block, the Real-Delegate approach mismatches the corresponding data in the first instance and attempts to match the query with a first entity in the 'composer(R6-R10)' block. It reduces the superfluous comparisons in searching R2, R3, R4, and R5 in the 'singer' block and its sub-blocks. In the case of redundant comparison reduction, after matching the query keyword with R6, the Real-Delegate maps its sub-blocks of 'violin(R6-R7)' and 'piano(R8-R10)'. If a similar query keyword 'composer' arrives in the next query, the Real-Delegate employs the mapped entities retained from the previous mapping process, which is found in the 'composer' block. Thus, the Real-Delegate approach considerably reduces redundant comparisons and enables effective hierarchical penetration to its subsequent attributes.

4.4 Real-delegate blocking and meta-blocking in the MapReduce framework

The Real-Delegate approach executes blocking and meta-blocking in the MapReduce framework [30]. Attribute-based unsupervised hierarchical blocking divides a set of attributes into multiple subsets and allocates each subset to the mapper. After executing the RST-based indiscernibility measurement in each mapper, the reducer provides a set of optimal attributes in each domain. Likewise, the Real-Delegate executes meta-blocking in the MapReduce framework and ensures redundancy comparison-negative querying. For example, Fig. 3 reveals the pairwise comparison reduction using MapReduce in which the attribute-based generated blocks are given as the input of the MapReduce method. In Fig. 5, A_1 has one domain value that contains three entity descriptions, namely, R_1, R_2, and R_3; A_2 has two domain values in which entities R_1 and R_2 are grouped under one domain value, and R_3 is grouped under another domain value; A_3 separately has three entities of R_1, R_2, and R_3 under three domain values.

When a query request arrives, the Real-Delegate approach generates an accurate result via the process of compression of pairwise comparisons. The Map phase considers the representation of the viable pairwise comparison structure of each block and transforms it into the resembling structure. The initial value of the entity in each block is nearly equivalent to the value of the other entity descriptions in a block in which the value represents similar features or attributes. The reduced phase generates the optimized or non-linear query comparison structure using the resemblance through the process of pruning the redundant comparisons. Exploiting of resemblance feature in the Real-Delegate approach facilitates the distinguishing of the redundant and non-redundant entity comparisons that are based on the two matching states such as '0' and '1'. Each value of the optimized query comparison structure depends on the current entity of a block as well as on its preceding entities within a block. The value of the

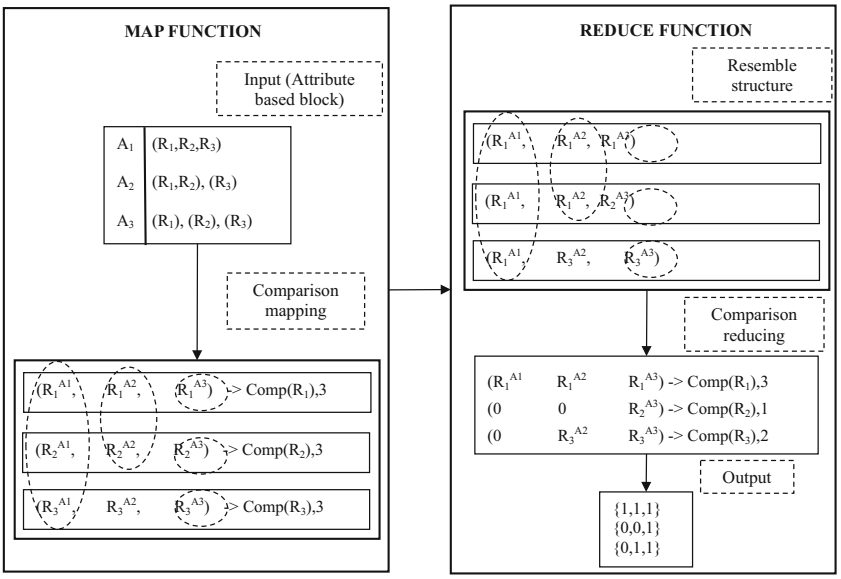


Fig. 5 Pairwise comparison reduction via the MapReduce method

query optimization structure is crucial for identifying whether the entity comparison is redundant or non-redundant. If the current entity is similar to any one of the preceding entities, the Real-Delegate approach assigns the value '0'; otherwise, it assigns the value '1'. When the succeeding entities become similar, the Real-Delegate approach does not change the state '0' until reaching the dissimilar entity description. It retains a new set of entity collections that reveal the entities performing comparison according to the query. After performing the comparisons in one block, it terminates the entity comparisons from the current block to its succeeding block of an attribute and continues the entity comparison from the initial entity of its succeeding block within an attribute. Thus, the Real-Delegate approach effectively utilizes the MapReduce function in Hadoop and improves the query processing qualitatively.

5 Experimental evaluation

The experimental evaluation compares the proposed Real-Delegate approach with the conventional TB-EER method [24] and analyzes the effectiveness and efficiency of the proposed approach. Moreover, the Real-Delegate approach is compared to the existing ER methods with various blocking techniques such as Locality-Sensitive-Hashing (LSH) [27], and Prefix-Infix-Suffix (PIS) [26].

5.1 Experimental setup

The experimental framework implements both the Real-Delegate and existing TB-EER approach on a Hadoop platform to assess the performance. The framework employs

Apache Hadoop 1.2.1, Hbase 0.94.16, and Java 1.8.0 from OpenJDK. The evaluation process runs on a Linux Ubuntu 12.04 LTS 64-bit machine, 2.9 GHz Intel CPU and 32 GB memory.

The Real-Delegate runs the proposed algorithm using the MapReduce and stores the restructured meta-blocks by availing Hbase. The prototype of the Real-Delegate approach is as follows. Initially, it employs the LOD source to find the equivalent relations of entity descriptions from a variety of sources. The LOD source requires the Resource Description Framework (RDF) type of the URI link for each entity description to effectively interlink the Web entity descriptions that have the owl:sameAs links of URI. For instance, 'DBpedia' and 'Geonames' use two different URIs of http://dbpedia.org/resource/Berlin and http://sws.geonames.org/2950159/, respectively, to identify a similar real-world entity of 'Berlin'. LOD links this type of URI links that facilitate the Real-Delegate to resolve the Web entities accurately. The Real-Delegate approach utilizes the information from the corresponding Wikipedia infobox templates of entity description to perform the attribute-based hierarchical clustering algorithm. The infoboxes contain the field information for each category, which facilitates the categorization of a set of URIs into a specific set of URIs with equivalent attributes under the main class extracted from the LOD. The Real-Delegate approach measures indiscernibility score and selects optimal attributes to generate blocks hierarchically. Then, it applies the meta-blocking technique to reduce the number of comparisons within a block on Hbase. The Real-Delegate meta-blocking method provides the input and output data of the MapReduce jobs for the Hbase in Hadoop platform. Thus, the MapReduce concept-based pairwise comparison reduction efficiently restricts the search space in the large-scale Web data.

5.1.1 Dataset

The performance of the Real-Delegate approach is evaluated using the DBpedia and BBC datasets. The Real-Delegate approach exploits the raw infoboxes from DBpedia version 3.5 to implement the attribute clustering algorithm. The Real-Delegate experimental framework combines the DBpedia and John Peel Sessions at BBC [38]. It links the artists in the BBC John Peel Sessions dataset to the corresponding information in DBpedia. Table 2 illustrates the information about both datasets. The DBpedia version 3.5 infoboxes contain 27,011,880 RDF triples, 1,638,149 entity descriptions, 31,857 attributes, and 16.49 average attribute-value pairs per description. John Peel Sessions contain the data related to the live musical performances conducted by the John Peel on the BBC Radio One. The John Peel Session dataset consists of 2,087 owl:sameAs links to DBpedia, but the type of entities restricts peel Session in terms of artists and songs. The evaluation framework employs the Gold Standard to assess the performance of the proposed algorithm. To precisely measure the accuracy, the proposed evaluation framework constructs the Gold Standard by exploiting the average value while frequently running the proposed algorithm on the dataset.

The Real-Delegate system employs the column-oriented database of HBase, and thus, HBase Query is more appropriate for the query-processing phase rather than using other structured query languages. HBase comprises a set of tables in which each table contains multiple column families, and each column family includes a set of rows

Table 2 Statistics of the Datasets

Datasets	Statistics	
DBpedia Infoboxes	RDF triples	27,011,880
	Entity descriptions	1,638,149
	Attributes	31,857
	Average attribute-value pairs per description	16.49
John Peel Sessions at BBC (DBTune)	RDF triples	277000
	owl:sameAs links to DBpedia	2,087
	Distinct DBpedia resources	1143

and columns. The Real-Delegate system sends the query in the HBase query format to optimally retrieve the appropriate information from the HBase storage system.

5.1.2 Evaluation metrics

The proposed evaluation framework employs performance metrics, such as Precision, Recall, F-measure, reduction ratio (RR) and Response time, to assess the performance of the Real-Delegate approach. Scalability is the significant factor that has a profound influence on accuracy. Accordingly, this work evaluates the precision and recall of the number of Web entity collections to demonstrate the accuracy of the system without compromising the accuracy with scalability. Regardless of the scalability, minimizing the processing time via optimal selection of attributes is necessary even when the number of Web entity collections is nominal. Thus, this work evaluates the RR and response time by varying the number of attributes.

- **Precision** Precision is also referred to as the measurement of pair quality.

$$\text{Precision} = \frac{\text{Number of mapping comparisons between the Gold Standard and the Real-Delegate system}}{\text{Number of total comparisons in the Real-Delegate system}}$$

- **Recall** It measures the pair completeness value.

$$\text{Recall} = \frac{\text{Number of mapping comparisons between the Gold Standard and the Real-Delegate system}}{\text{Number of total comparisons in the Gold Standard}}$$

- **F-measure** It is the weighted harmonic mean of precision and recall.

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Reduction ratio (RR)** It measures the reduction in the search space in terms of pairwise comparison reduction.

$$\text{RR} = 1 - \frac{\text{Generated pairwise comparisons}}{\text{Initial pairwise comparisons}}$$

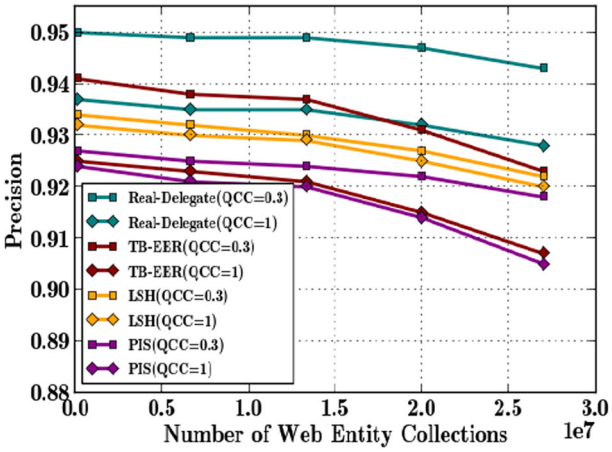


Fig. 6 Number of web entity collections versus precision

- **Response time** It is the total time taken by the user query on the Web data to deliver the desired information to the corresponding query.

5.2 Experimental results

The experimental results validate the effectiveness of the proposed approach by applying real-time scenarios with a different number of entity collections and attributes.

5.2.1 Number of web entity collections versus precision

Figure 6 illustrates the precision of both the Real-Delegate and several existing blocking methods while testing on the John Peel sessions dataset with the DBpedia. In this scenario, the number of Web entity collections is varied from 119010 to 27011880, and the Query Cross-Correlation (QCC) is in the range of 0.3 and 1. QCC refers the correlation value of the Web query across several domains. Overall, both the precision of the Real-Delegate and TB-EER consistently declines when the number of Web entity collections increases. When increasing the number of Web entity collections from 119010 to 27011880 and QCC=1, the precision value of the TB-EER approach rapidly falls by 1.94% after reaching a specified number of entities, whereas only a marginal decrease of 0.96% precision is observed in the Real-Delegate approach while applying the attribute-based hierarchical blocking method. This is because, the attribute-agnostic blocking method of TB-EER does not considerably affect the performance when QCC has the minimum value, but it creates greater uncertainty when a single token correlates with the multiple domains. The precision values of the LSH and PIS blocking methods are nearly 1.43% higher than those of the token blocking method when the QCC of TB-EER is 1 because the LSH and PIS methods employ the contextual similarity methods to attain the dimensionality reduction in terms of reducing the redundant and unnecessary pairs.

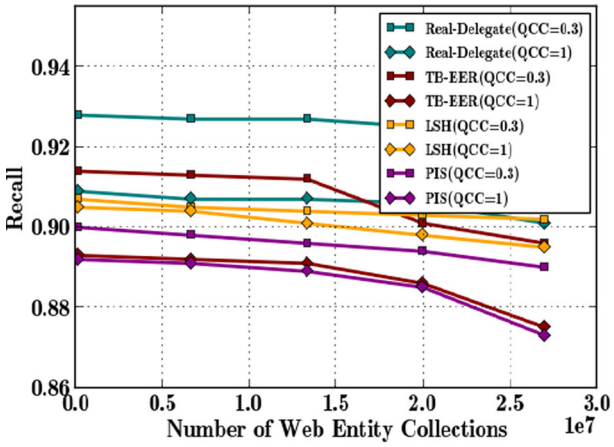


Fig. 7 Number of web entity collections versus recall

5.2.2 Number of web entity collections versus recall

Figure 7 depicts the recall of the Real-Delegate and baseline approaches while varying the number of Web entity collections and the QCC. To examine the effect of the proposed algorithm, the experimental evaluation varies the values of QCC. The recall value descends when the number of Web entity collections increases from 119010 to 27011880. The recall of Real-Delegate approach has a negligible impact even when the QCC value is 0.3 and is varied. It maintains the recall in the range of 92.1–92.8% for all sample sizes of entity collections when QCC is 0.3. This recall reflects the effectiveness and scalability of the proposed approach under different fluctuations. The number of entity collections linearly increases the total number of pairwise comparisons while querying the results. As a result, the Real-Delegate approach achieves a 2.6% higher recall value than the TB-EER approach when the number of entity collection is 27011880 and QCC is 1. The performance of the TB-EER approach becomes similar to the PIS method when the QCC=1 while increasing the number of web entity collections. The reason is that the PIS method matches the prefix and infix of the link, which leads to a match with similar links and, consequently, establishes the entity linkage effectively even when there is a variation in the prefix and suffix information.

5.2.3 Number of attributes versus F-measure

(i) F-measure with the impact of QCC

Figure 8 illustrates the F-measure of the Real-Delegate approach while comparing the TB-EER, LSH, and PIS approaches with the impact of QCC. While increasing the number of attributes, the Real-Delegate approach exploits an optimal set of attributes and effectively maintains the performance. When the QCC value is 0.3 and the number of attributes varies from 5 to 25, the Real-Delegate approach manages the F-measure value to maintain the variation within 0.53%, whereas the TB-EER manages with the

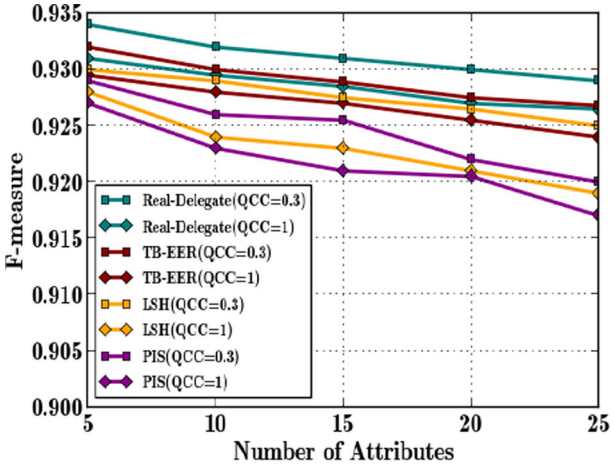


Fig. 8 Number of attributes versus the F-measure with the impact of QCC

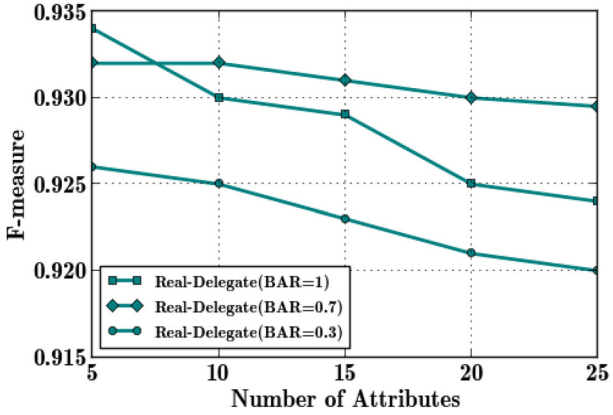


Fig. 9 Number of attributes versus F-measure with the impact of BAR

variation of 0.64%. The selection of the optimal set of attributes in the Real-Delegate approach facilitates search space reduction and improves the contextual search when there is an increasing number of attributes as well as QCC. In the existing blocking techniques, the LSH method outperforms the PIS method when increasing the number of attributes since LSH has not been compromised with the scalability while ensuring accuracy, whereas in other methods, such as token blocking and the PIS blocking methods, the coverage of the Web data is limited. The F-measure value of LSH is nearly equal to the Real-Delegate method when dealing with various QCC values. The Real-Delegate approach marginally decreases its performance only by 0.48% even when the QCC is 1, but the LSH degrades it to 0.53% even when QCC has a minimum value.

(ii) F-measure with the impact of the BAR

Figure 9 indicates the F-measure of the Real-Delegate approach while varying the blocking attribute ratio (BAR) from 0.3 to 1, and the number of attributes varies

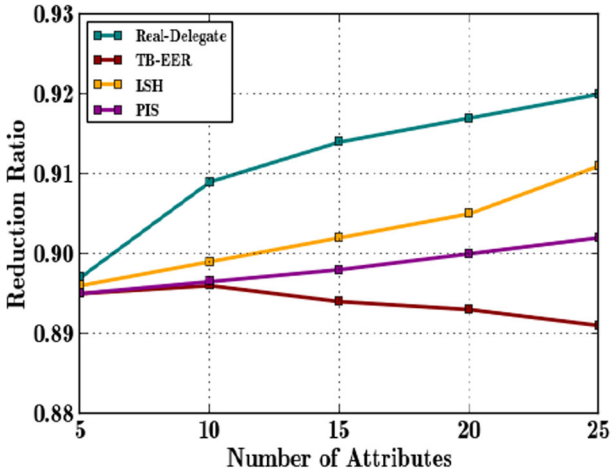


Fig. 10 Number of attributes versus the reduction ratio

from 5 to 25. The BAR is defined as the ratio between the number of attributes for blocking and the total number of attributes per entity in the Web entity collections. Increasing the number of attributes degrades the F-measure value since abundant attributes escalate the complexity of query processing in the Web data. The F-measure of the Real-Delegate approach to be maintained with the minimum variation for the BAR is 0.7, and it shows drastic changes when the BAR is at 0.3 and 1. The proposed approach maintains an optimum set of attribute-based blocking to provide both efficient and effective query processing. When the BAR is minimum, the system is unable to offer accurate results across the number of attributes.

5.2.4 Number of attributes versus the reduction ratio

Figure 10 illustrates the reduction ratio (RR) of the Real-Delegate and several baseline ER approaches when the number of attributes varies from 5 to 25 in the Web entity collections. The experimental results suggest that when the number of attributes is high, RR is also high. The outcome of RR implies that the higher RR value provides better performance in which the RR value of the Real-Delegate approach increases when the number of attributes is added. Both the attribute based hierarchical blocking and meta-blocking methods significantly reduce the superfluous and redundant comparisons among the Web entity collections. Hence, the Real-Delegate enables the proposed system to achieve a higher RR value. When the number of attributes is 25, the Real-Delegate approach achieves 3.25% of RR than the TB-EER approach. The TB-EER approach requires additional comparison due to the consideration of token blocking method, and this may lead to mismatches. Both the existing LSH and PIS methods escalate the reduction ratio compared with the baseline TB-EER method. The probabilistic nature of the LSH method effectively creates the block with the most similar entities, which averts unnecessary comparisons in each block.

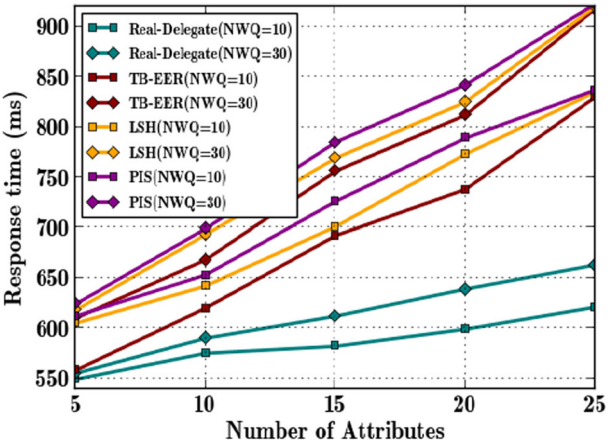


Fig. 11 Number of attributes versus the response time

5.2.5 Number of attributes versus the response time

Figure 11 demonstrates the response time of the Real-Delegate and several existing ER approaches while increasing the number of attributes in the Web entity collections and the Number of Web Queries (NWQ) submitted by the users. The experimental results of Figure 11 show the variation of response time for NWQ is in the range of 10 to 30. The response time extends while increasing the number of attributes and NWQ. In the Real-Delegate approach, the response time gradually increases while the number of attributes vary from 5 to 25, and NWQ is set to 10. The attribute-based hierarchical blocking and meta-blocking has enabled the system to compare within the relevant comparisons rather than searching the entire generated blocks. When varying the number of attributes from 5 to 25 and when NWQ is 30, the Real-Delegate approach prolongs the response time from 555 to 663 ms only. However, in the TB-EER approach, the time is severely extended to 918 ms from 610 ms. When the number of attributes is 25, the Real-Delegate approach provides the result within 621 ms, but the PIS method consumes 832 ms. However, the PIS method provides very similar results as found in the token blocking method at the point of NWQ = 10, and the number of attributes is 25. This is because the token blocking method probably considers the needless tokens. The consideration of subject URIs in the PIS method assists in improving the response time performance.

5.2.6 Number of web entity collections versus the response time

As shown in Figure 12, the response time increases while adding the number of Web entity collections that represent the scalability of the Real-Delegate with the existing blocking approaches. The Real-Delegate approach manages to retain the response time at a reasonable value even when the scalability of the system is high due to the reduction of unnecessary and redundant pairwise comparisons. However, the TB-EER approach needs to compare the number of pairwise comparisons in each block,

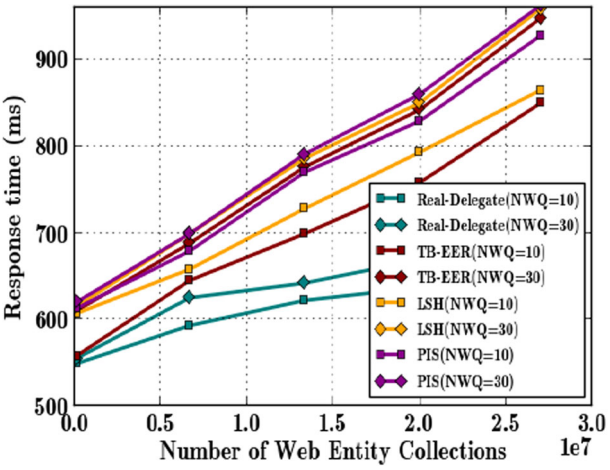


Fig. 12 Number of web entity collections versus the response time

which makes the system to extend the response time while increasing the number of Web entity collections. The Real-Delegate approach reduces the resulting time query to 257 ms while comparing the TB-EER approach when the number of Web entity collection is 27011880, and 30 NWQ are present. The LSH method precisely estimates the coherence among the real-world entities, which is very similar to the meta-blocking of the Real-Delegate approach. Hence, it can manage the related entities effectively and, thus, leverages the system to provide an immediate response. However, while comparing the Real-Delegate and TB-EER approaches, the LSH method lacks the support of the heterogeneous information spaces, which deteriorates the performance.

6 Conclusion

This study proposes an effective ER methodology in the context of loosely structured, highly heterogeneous, and voluminous Web data. The proposed Real-Delegate approach employs the LOD, attribute-based unsupervised hierarchical blocking, and meta-blocking to resolve the constraints of the ER process of the Web data. In contrast to the existing blocking methods, the Real-Delegate approach employs URI information of the entity descriptions in the loose schema to effectively link the homogeneous Web entities. Attribute-based unsupervised hierarchical blocking reduces the search space and eliminates the process of superfluous comparisons while querying the results. Moreover, the meta-blocking method efficiently restricts the redundant pairwise comparisons using the metadata of each block. Eventually, the Real-Delegate approach resolves Web entities while ensuring higher effectiveness and efficiency. The experimental evaluation reveals that the proposed Real-Delegate approach outperforms the existing token blocking-based ER approach in terms of Recall and Response time. The experimental results of the real-world datasets illustrate that the Real-Delegate system yields a pairwise comparison reduction to the extent of 3.2%

more than that of the TB-EER method while ensuring a 93% recall. The future work of the Real-Delegate approach is as follows:

The future work must contemplate extending the Real-Delegate approach with the support of recent terms update to address the emerging web entity trends and applications. It must focus on linking similar web entities from all web information sources with context-dependent decision-making rather than considering Wikipedia alone. The future work must plan for including temporal information-based indexing to reduce the search space more effectively.

References

1. Dong, X.L., Srivastava, D.: Big data integration. IEEE 29th International Conference on Data Engineering (ICDE), pp. 1245–1248 (2013)
2. Stefanidis, K., Efthymiou, V., Herschel, M., Christophides, V.: Entity resolution in the Web of data, ACM Proceedings on WWW, pp. 203–204 (2014)
3. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. Knowl. Data Eng. **24**(9), 1537–1555 (2012)
4. Steorts, R.C., Ventura, S.L., Sadinle, M., Fienberg, S.E.: A comparison of blocking methods for record linkage. In: Domingo-Ferrer, J., et al. (eds.) Privacy in Statistical Databases, pp. 253–268. Springer, Berlin (2014)
5. Zhu, S., Wang, D., Li, T.: Data clustering with size constraints. Knowl. Based Syst. **23**(8), 883–889 (2010)
6. Fisher, J., Christen, P., Wang, Q., Rahm, E.: A clustering-based framework to control block sizes for entity resolution, 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 279–288 (2015)
7. Christophides, Vassilis, Efthymiou, Vasilis, Stefanidis, Kostas: Entity resolution in the web of data. Synth. Lect. Semant. Web **5**(3), 1–122 (2015)
8. Kopcke, H., Rahm, E.: Frameworks for entity matching: a comparison. Data Knowl. Eng. **69**(2), 197–210 (2010)
9. Papadakis, G., Ioannou, E., Niederée, C., Palpanas, T., Nejdl, W.: To compare or not to compare: making entity resolution more efficient, ACM Proceedings of the International Workshop on Semantic Web Information Management, p. 3 (2011)
10. Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures, ACM Proceedings of the ninth SIGKDD international conference on Knowledge discovery and data mining, pp. 39–48 (2003)
11. De Assis Costa, G., de Oliveira, J.M.P.: A relational learning approach for collective entity resolution in the web of data, ACM Proceedings of the 5th International Conference on Consuming Linked Data, vol. 1264, pp. 13–24 (2014)
12. Li, C., Jin, L., Mehrotra, S.: Supporting efficient record linkage for large data sets using mapping techniques. World Wide Web **9**(4), 557–584 (2006). Springer
13. McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of highdimensional data sets with application to reference matching, ACM Proceedings of the Sixth SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 169–178 (2000)
14. Ukkonen, E.: Approximate String Matching with q-grams and Maximal Matches, Theoretical Computer Science, vol. 92, pp. 191–1211. Elsevier Science Publishers Ltd., Essex (1992)
15. Gravano, L., Ipeirotis, P.G., Koudas, N., Srivastava, D.: Text joins in an RDBMS for web data integration, ACM Proceedings of the 12th International Conference on WWW, pp. 90–101 (2003)
16. Vries, T., Ke, H., Chawla, S., Christen, P.: Robust record linkage blocking using suffix arrays and Bloom filters, ACM Transactions on Knowledge Discovery from Data, vol. 5, No. 2 (2011)
17. Whang, S.E., Menestrina, D., Koutrika, G., Theobald, M., Garcia-Molina, H.: Entity resolution with iterative blocking, ACM Proceedings of the SIGMOD International Conference on Management of data, pp. 219–232 (2009)
18. Shu, L., Chen, A., Xiong, M., Meng, W.: Efficient spectral neighborhood blocking for entity resolution, IEEE 27th International Conference on Data Engineering, pp. 1067–1078 (2011)

19. Sarma, Das A., Jain, A., Machanavajjhala, A., Bohannon, P.: An automatic blocking mechanism for large-scale de-duplication tasks, 21st ACM International Conference on Information and Knowledge Management, pp. 1055–1064 (2012)
20. Ramadan, B., Christen, P.: Unsupervised blocking key selection for real-time entity resolution, Springer International Publishing on Pacific-Asia Conference on Knowledge, pp. 574–585 (2015)
21. Chen, H.-L., Yang, B., Liu, J., Liu, D.-Y.: A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Syst. Appl. **38**(7), 9014–9022 (2011)
22. Kaya, Y., Uyar, M.: A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Appl. Soft Comput. **13**(8), 3429–3438 (2013)
23. Nin, J., Muntes-Mulero, V., Mart ınez-Bazan, N., Larriba-Pey, J.-L.: On the use of semantic blocking techniques for data cleansing and integration, IEEE 11th International Symposium on Database Engineering and Applications, pp. 190–198 (2007)
24. Papadakis, G., Ioannou, E., Niederee, C., Fankhauser, P.: Efficient entity resolution for large heterogeneous information spaces, ACM Proceedings of the Fourth International Conference on Web Search and Data Mining, pp. 535–544 (2011)
25. Ma, Y., Tran, T.: Typimatch: type-specific unsupervised learning of keys and key values for heterogeneous Web data integration, Sixth ACM International Conference on Web Search and Data Mining, pp. 325–334 (2013)
26. Papadakis, G., Ioannou, E., Niederee, C., Palpanas, T., Nejdl, W.: Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data, ACM Proceedings of the Fifth International Conference on Web Search and Web Data Mining, pp. 53–62 (2012)
27. Kim, H.S., Lee, D.: HARRA: fast iterative hashed record linkage for large-scale data collections, ACM Proceedings of the 13th International Conference on Extending Database Technology, pp. 525–536 (2010)
28. Papadakis, G., Koutrika, G., Palpanas, T., Nejdl, W.: Meta-blocking: taking entity resolution to the next level. IEEE Trans. Knowl. Data Eng. **26**(8), 1946–1960 (2014)
29. Papadakis, G., Ioannou, E., Niederee, C., Palpanas, T., Nejdl, W.: Eliminating the redundancy in blocking-based entity resolution methods, ACM Proceedings of the 2011 Joint International Conference on Digital Libraries, pp. 85–94 (2011)
30. Papadakis, G., Papastefanatos, G., Palpanas, T., Koubarakis, M.: Scaling entity resolution to large, heterogeneous data with enhanced metablocking, In EDBT, pp. 221–232 (2016)
31. Papadakis, G., Papastefanatos, G., Koutrika, G.: Supervised metablocking. ACM proceedings of the VLDB **7**(14), 1929–1940 (2014)
32. Papadakis, G., Ioannou, E., Palpanas, T., Niederee, C., Nejdl, W.: A blocking framework for entity resolution in highly heterogeneous information spaces. IEEE Trans. Knowl. Data Eng. **25**(12), 2665–2682 (2013)
33. Efthymiou, V., Stefanidis, K., Christophides, V.: Benchmarking blocking algorithms for Web entities. IEEE Trans. Big Data (2016). doi:10.1109/TBDATA.2016.2576463
34. Efthymiou, V., Papadakis, G., Papastefanatos, G., Stefanidis, K., Palpanas, T.: Parallel meta-blocking: realizing scalable entity resolution over large, heterogeneous data, IEEE International Conference on Big data (Big data), pp. 411–420 (2015)
35. Efthymiou, V., Papadakis, G., Papastefanatos, G., Stefanidis, K., Palpanas, T.: Parallel meta-blocking for scaling entity resolution over big heterogeneous data. Information Systems **65**, 137–157 (2017)
36. Papadakis, G., Demartini, G., Fankhauser, P., Kärger, P.: The missing links: discovering hidden same-as links among a billion of triples, ACM Proceedings of the 12th International Conference on Information Integration and Web-based Applications and Services, pp. 453–460 (2010)
37. Bizer, C., Heath, T., Berners-Lee, T.: Linked data—the story so far. In: Semantic Services, Interoperability and Web Applications: Emerging Concepts, pp. 205–227 (2009)
38. http://dbtune.org/bbc/peel/
39. Vidhya, K.A., Geetha, T.V.: Rough set theory for document clustering: a review. J. Intell. Fuzzy Syst. **32**(3), 2165–2185 (2017)
40. Vidhya, K.A., Geetha, T.V., Aghila, G.: Text document classification using Rough Set theory and Multi-level Naïve Bayes. Int. J. Appl. Eng. Res. **10**(75), 331–336 (2015). (IJAER)