Indian Journal of Science and Technology, *Vol 8(34), DOI: 10.17485/ijst/2015/v8i34/75103, December 2015*

*ISSN (Print) : 0974-6846*
*ISSN (Online) : 0974-5645*

# A Semantic Deduplication of Temporal Dynamic Records from Multiple Web Databases

**R. Parimala Devi[1]* and V. Thigarasu[2]**

[1]Department of Computer Science, Karpagam University, Coimbatore - 641021, Tamil Nadu, India;
parimaladeviphd@gmail.com
[2]Department of Computer Science, Gobi Arts and Science College, Gobichettipalayam - 638453,
Tamil Nadu, India; vthigarasu@gmail.com

## Abstract

**Objective:** The main objective of this paper is to improve the true positive level of record deduplication using Ontology based MHMM-Fuzzy clustering approach. **Methods/Statistical Analysis:** Most of the record deduplication system in literature used genetic programming based record deduplication which combined different pieces of evidence extracted from the data content. However, the accuracy of the system is low. To overcome this problem, a Multiple Hidden Markov Model (MHMM) is proposed and it is used to increase the accuracy and also to identify joint duplicate records. In this model, if the database has multiple columns, it performs the deduplication for the all columns which can degrade the performance of the system. To solve this problem, MHMM-Fuzzy Clustering based record deduplication is introduced. In this system Fuzzy clustering is performed through multiple observations from the Hidden Markov Model. Then the duplicate data are grouped into one cluster according to their fuzzy logic and it can be eliminated easily. However ,the true positive level of the system is low. To improve the true positive level, Fuzzy Ontology based semantic similarity is incorporated in MHMM-Fuzzy Clustering approach. This implies the improvement of the true positive level of the model. Thus, it increases the efficiency of deduplication function that identifies the records of replica and duplications. **Findings:** Multiple Hidden Markov Model (MHMM) based record deduplication, MHMM-Fuzzy clustering based record deduplication and Ontology based MHMM-Fuzzy clustering approach are applied on Cora Bibliographic dataset and Restaurants dataset. The performance measures are evaluated in terms of precision, recall, f-measure, execution time and accuracy results. **Applications/Improvements:** Thus the current research achieves improved result on record deduplication is better than previous works in terms of precision, recall, f-measure, execution time and accuracy results.

**Keywords:** Hidden State Sequence, Membership Function, Observation Sequence, States, Semantic Deduplication

## 1. Introduction

### 1.1 Record Deduplication

Record deduplication[1] is a dedicated data compression approach which is used for removing duplicates from various sources[2]. Duplicate data holding such mistakes as spelling, erroneous data linked with a field, unfinished or out-dated data.

### 1.2 Record Deduplication in Data Mining

Data mining is the technology which extracts the useful information needed by the organization for taking a well again assessment. The enormous development in the data base size is affected by trouble of dirty data. Due to these unclean data in the database causes variety of problems such as quality loss, increased cost and performance ruin[2].

The above mentioned problems are avoided by discarding "dirty data" from the data source. The dirty data is the data with replicas, with no uniform representation, etc. It requires technical efforts to manage them. By avoiding them, the overall speed and system performance will be increased.

** Author for correspondence*

The problem of discovering and removing dirty or duplicate records from a data base is called as record deduplication. It is also known as data cleaning and record matching.

The duplicate record is categorized into three types[4]. There are

- Fully Duplicated Records.
- Erroneous Duplicated Records.
- Partially Duplicated Records.

In fully duplicated records, the two rows indicate the same real world entity. In Erroneous duplicated records due to the mistake of data entry operator's, they appeared as dissimilar. The partially duplicated records are partially duplicates but there are dissimilar from the original records.

The most important challenge in this task is designing a function that can resolve when a pair of records refers to the same entity in spite of various data discrepancy[5].

The record deduplication[6] is the method of recognizing same individual across various data sources or warehouse. There is a variety of schemes to record deduplication. They are:

- Adhoc or domain knowledge schemes. It is based on area knowledge and utilizes declarative languages.
- Training based schemes. It is based on supervised or semi-supervised learning.

## 2. Data Deduplication Advantages

- Reduced storage capacity is necessary for a certain amount of data[7].
- Ability to store considerably more data on the given amount of disk.
- Restore from disk rather than tape may develop ability to meet Resurgence Time Objective (RTO).
- Network bandwidth savings (some implementations).
- Lower storage-management and energy costs resulting from reduced storage requirements.

**Gayathri et al.** presented a firefly algorithm which is used to record deduplication. This Meta heuristic algorithm is motivated by a flashing behaviour of fireflies. Each and every firefly is attracted by other fireflies which one has high brightness. The brightness can be decreased according to distance increases. Here, the objective function (f(x)) depends on several piece of proof mining from the data. It is found that the dirty data based on the flashing activities of the each firefly and their movements from one position to another. While there are no fireflies darker than another firefly, the firefly's moving arbitrarily[8]. It facilitates the fireflies to travel in the direction of preeminent location of duplicate records identification or replica records recognized and new attractive locations in order to obtain optimal record Deduplication. It does not have high accuracy**.**

**Xin Wang, et al.** presented an Onto Clean Framework for Ontology-Based Data Cleaning. If the data records hold errors such as missing values and mislay values, the system can use Ontology[9] to verify the domain constraints on the attributes. To check some other semantic errors domain,Ontology has been used[10]. The record duplication problem occurs if the same person is represented in a contact list with a little varying names or addresses. Based on the purpose of the cleaning and the domain, an appropriate cleaning algorithm is selected. Ontology-Based Data Cleaning is able to clean some classes of semantic errors. It cleans only the some classes of the semantic errors.

**Bilal Khan et al.** introduced a de-duplicator algorithm which is based on numeric conversion of entire data**.** The proposedsystem considers three phases. The phases,are conversion, clustering and matching.In conversion phase,theuniform format data are converted into string, numeric or date by using radix formula on the data. Those values are stored in the column[11]. The 'k'-mean clustering algorithm applied on the values which is stored in the column. Here, the matching records are stored in one cluster and mismatching records are stored in another cluster. Once match is found among the records,then the percentage of duplication is computed. This proposed technique detects with fully duplicated records and partially duplicated records.

**Moise's G et al.** presented a Genetic Programming technique to record deduplication. This technique merges different pieces of evidence extracted from the data content which is used to discover two or more entries in the data base are replicas or not. Reproduction is a process of copy of individuals without any modification[12]. Generally, this operator is used to carry out an exclusive strategy that is adopted to keep the genetic code of the fittest individuals athwart the changes in the generations. In that mutations procedure, each piece of evidence E is a couple <attribute; similarity function> that symbolizes the requirements of exact resemblance function over the values of a specific attribute found in the data being calculated. At the final stage, entire number of correct and incorrect replicas is determined.

# 3. Methodologies

## 3.1 Joint Duplicate Record Detection using Multiple Hidden Markov Model

Record deduplication is a mission of discovering duplicate records which holding variety of writing styles, misspelling and repetitive words. Discovering duplicates in individuals records from the data collected at various sources are most important mission. In different records, such as relational databases, a record of one type is dependent on the other record types. Perfect deduplication for records of one type is frequently dependent on the resolution made for records of other types. To identify the duplicates in such records, the system proposed a Joint duplicate record identification by using Multiple Hidden Markov Models (MHMM).

The Hidden Markov Model has a fixed set of states. Transitions between these states are direct by a set of probabilities which is called as transition probabilities. Here, the records with variety of attributes are called states and a resemblance among the two records is characterized as transition probability. The attribute information in the data records contains author name, published year, implemented title, venue and pages. An individual state outcome or observation can be produced according to the connected probability distribution.

The number of states is denoted as N. The set of states is $S = \{S_1,S_2,...S_N\}$

Where,

$S_i$, i=1,2,...,N is an individual state.

$q_t$ - The state at time instant t.

The number of separate observation symbols per state is M. The state transition probability matrix $A = [a_{ij}]$.

The observation symbol probability matrix is denoted as $B = [b_j(k)]$. The observation sequence $O = O_1, O_2, O_3...O_R$, where each observation $O_t$ is one of the symbols from V, and R is the number of observations in the sequence. The multipleobservation probability is represented as grouping of individual observation probabilities without losing generalization in the Hidden Markov Model scheme. Multiple observation sequences are associated with the hidden state sequence, and these observations may not be synchronized to each other. The states are not visible to the external user.

The state sequence is represented as $\{S_t\}$ and $O_t$ represents the observable output at time t coupled with state $s_t$ and let $b_m (S_t)$ be the probability of observing $O_t$. It is assumed that, Two sequences $\{O_t\}$ and $\{q_t\}$ are outputs of an HMM state sequence. If some random delayed τ among the two output sequences, these two sequences are no longer synchronized. The symbol $\varphi_t$ represents the missed observation (i.e., null observation) of the output at time t.

The multiple observation probability is represented as a grouping of individual observation probabilities without reducing the generality in the HMM (Hidden Markov Model) method. These observations are combined to capture relational dependencies between each collected records.

These multiple observation sequences hold their observation intervals, initial points, etc. Here, propose a new relational partitioning method for conclusion which allows the decisions from one record type to update the decisions for another record type. To this end, it is described that a group of binary random variables representing whether or not two records are duplicates. $A_i^a$ and $A_j^b$ observed records means $R_{ij}^{ab}$ indicates whether some relation R holds between record mentions $A_i^a$ and $A_j^b$.

For example, in a research paper database, $A_i^a$ represents the set of paper records, $A_j^b$ represents the set venue records.To capture the dependence among observed records, factorize the feasible functions to consider them jointly duplicate the records.

If multiple columns of records exist in that time, it can perform deduplication process for the all columns.

**Algorithm 1**

```
Step 1: Initialize S
Step 2: Compute $\pi$
        // $\pi$ -Start probability
Step 3: Determine A
// A- transition probability
Step 4: Compute B
// B- emission probability
Step 5: Compute $\{o_t\}$
Step 6: For each observation → delay $\tau$
Step 7:  Compute $\{q_t\}$
Step 8: Compute multiple observation
Step 9: calculate $R_{ij}^{ab}$
// R holds between record $A_i^a$ and $A_j^b$.
Step 10: Eliminate duplicates
Step 11:end process
```

## 3.2 Improved MHMM Fuzzy Clustering Approach

Let assume λ=(A,B,π) be a given model and series of observations $O=(O_0,O_1,\ldots O_{T-1})$. Similarly, let $o_t$ represent the observable output at time linked with state $s_t$. The multiple observation probability is denoted as a group of individual observation probabilities without ruining generality in the HMM (Hidden Markov Model) technique. These observations are combined to capture relational dependencies between each collected records.

It is assumed that many sequences are available as the outputs of an HMM state sequence. If some random delay τ is established between the output sequences, these sequences are no longer synchronized. The symbol $\varphi_t$ represents the missed observation (i.e., null observation) of the output at time t. Here ,fuzzy c means clustering is used for group the observation based on the membership function.

In fuzzy clustering, the sequence were clustered according to membership value. In that system each and every cluster having cluster center that is also known as Cluster Head (CH). More the data is near to the cluster head more is its membership towards the particular cluster head. That cluster head has information about all records in that cluster. Thus, records on the edge of a cluster may be in the cluster to a lesser degree than the records in the center of cluster.

In this model, each and every output sequence is related with every cluster by means of a membership value. That clustering approaches the duplicate data are grouped into one cluster. The duplicate records in the data base were detected. It should improve the performance of Multiple Hidden Markov Model based deduplication.

**Algorithm 2**

```
Step 1: Initialize (A,B,π)
Step 2: Compute similarly among S
Step 3: S ← multiple observation sequence
Step 4:  for each sequence $C_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} - x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}}$
Step 5: Compute membership value
            $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left\{ \frac{\|x_i - c_i\|}{\|x_i - c_k\|} \right\}^{\frac{2}{m-1}}}$
        // $u_{ij}$ - Degree of membership of $x_i$ in the cluster
        $j$, $x_i$ – Output observation sequence, $C_j$- Center
        of the cluster//
Step 6:  if $\| U^{(k+1)} - U^{(k)} \| < \in$
            Group the output
Step 7: Otherwise
Go to step 3
Step 8: Remove duplicate record group
```

## 3.3 A Novel Approach for Record Deduplication using Fuzzy Ontology Model

To achieve semantic relatedness of various records Fuzzy Ontology method is used. Ontologies with a huge knowledge base suggested in various forms such as hierarchical trees and hyperbolic trees, etc. Ontology was used to express the meaning of user query terms by attains the synonyms of all the words that make up the user's query.

Fuzzy Ontologies are capable of dealing with fuzzy knowledge, and are efficient in determination of the precise meaning of a word as it relates to a record collection. It contains fuzzy theory and fuzzy membership functions.

Fuzzy Ontology structure includes a set of relations between concepts and each other. All relations are represented as membership degree. Fuzzy Ontology is represented as a pair (C, R).

Where,
C - Set of concepts,
R - Set of fuzzy relations between concepts.

It is assumed that many observations are available as the outputs of an MHMM state sequence. The constructed fuzzy ontology provides the semantic relatedness of various records obtained from these multiple HMM. The relationship between the record attribute is represented by a membership value in [0, 1].

In Fuzzy Ontology, each attributes is related to other attributes in the ontology and degree of membership μ (0 ≤ μ ≤ 1) is allocated to this relationship.

$$\sum_{i=1}^{i=n} \mu_i = 1$$

Where 0<μ<1 and μ corresponds to a fuzzy membership relation. Then, the membership function associated with fuzzy set F is defined as follows:

$$\mu_F : U \rightarrow [0, 1]$$

Where,
0 - no-membership

1 - Full membership

Table 1. Membership function of the Fuzzy Ontology record deduplication system

| Fuzzy output variable | Membership function |
|---|---|
| Relevant | High |
| | Medium |
| | Low |

In case of the user requests for information regarding a concept *a* (attribute a from MHMM) and an ontology models following fuzzy similarity relations: similar to attributes (a, b) = 0.8, similar to attributes (a, c) = 0.5 and similar to attributes (a, d) = 0.2. The attribute a has a high semantic relatedness with concept.

The Membership functions of the Fuzzy Ontology record deduplication system is represented in Table 1.

- First, if the degree of membership of one of the attributes is 0.8, then the attributes are highly relevant.
- Secondly, if it is 0.5, then the attribute is moderately relevant.
- Thirdly, if the membership function is 0.2, then the attribute is not relevant.

Low membership value represents an object does have a semantic similarity of attribute which is considered as duplicates and it has been eliminated. Here Fuzzy Ontology provides the semantic relatedness of various records obtained from the multiple HMM and finds the record replicas and duplications. The proposed Fuzzy Ontology approach increases the recall value, as more relevant results are considered, and also increase the precision.

**Algorithm 3**

```
Step 1: Initialise multiple $o_t$
            // $o_t$- Observation at t
Step 2: Construct Fuzzy Ontology
Step 3: For each attribute membership with other
                    $\mu$ (0 ≤ $\mu$ ≤ 1)
    // $\mu$-Membership function
Step 4: MHMM←Attribute
Step 5: Fuzzy Ontology ←query
Step 6: If $\mu$ >0.5
             High semantic similarity
Step 7: If $\mu$ <0.5
            Low semantic similarity
Step 8: Eliminate → low $\mu$ attribute
Step 9: end process
```

## 4. Experimental Results

### 4.1 Data Set Description

In this experiments, two real data sets are known as Bibliographic data set and Restaurants data set are used for evaluation. They are commonly employed which are based on real data gathered from the web.

First real data set is Cora Bibliographic data set. That data set is collection of 1,295 distinct citations to papers of 122 taken in computer science from the Cora research paper search engine. These citations were split into multiple attributes (author names, year, title, venue, and pages and other info) by an information extraction method.

Second real data set is Restaurants data set; it contains 864 entries of restaurant names and additional information, including 112 duplicates that were obtained by integrating records from Fodor and Zagat's guidebooks. The following attributes used from this data set: (restaurant) name, address, city, and specialty.

**Performance metric**

The performance measures was used to evaluate the proposed MHMM, Improved MHMM Fuzzy approach and Fuzzy Ontology.

- Precision value.
- Recall value.
- F-Measure value.
- Execution time.
- Accuracy.

## 5. Performance Comparison

### 5.1 Precision

Precision is defined as the percentage of correct predicted results from the set of input terms. The precision value should be more in the proposed methodology than the existing approaches for the better system performance. Precision is calculated by using following equation

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

The graphical representation is given in the following Figure 1.

The performance offered by various methods for record deduplication was analyzed and compared. In this graph, numbers of records are predicted in the x axis and the precision value is predicted in the y axis.

Here if the no of records are increased, the precision may also be increased linearly while deduplication process. The following graph shows Fuzzy Ontology based record deduplication has highest precision compared to all other system.

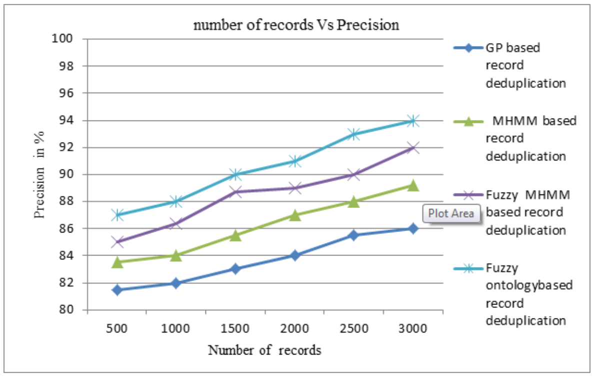

**Figure 1.** Precision Comparison.

## 5.2 Recall

The recall is the proportion of positive cases that are accurately identified, as computed using the equation:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{True Negative}}$$

The graphical representation of recall value is plotted in the following Figure 2.

The performance offered by various methods for deduplication was analyzed and compared. In this graph numbers of records are predicted in the x axis and the recall value is predicted in the y axis. Here, if the no of records are increased the recall may also be increased linearly while deduplication process. The following precision graph shows Fuzzy Ontology based record deduplication has highest recall over all other system.

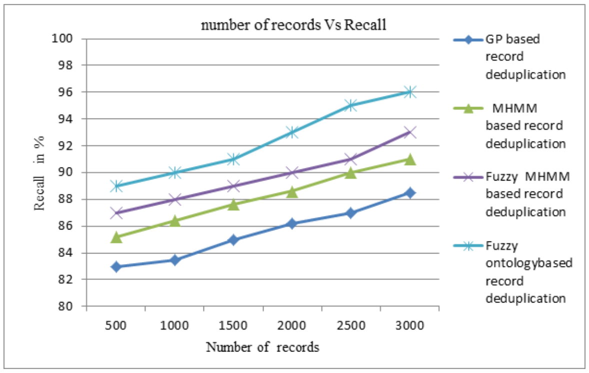

**Figure 2.** Recall Comparison.

## 5.3 F-Measure

The F-Measure computes some average of the information retrieval precision and recall metrics

$$\text{F - measure} = \frac{2 * precision.recall}{precision + recall}$$

From the above graph, it can be proved that the proposed Fuzzy Ontology based record deduplication method provides better result than other two approaches. In this Figure 3, x axis plots the number of records and y axis plots the F-measure value. Here,if the no of records are increased the F-measure may also be increased linearly while deduplication process.

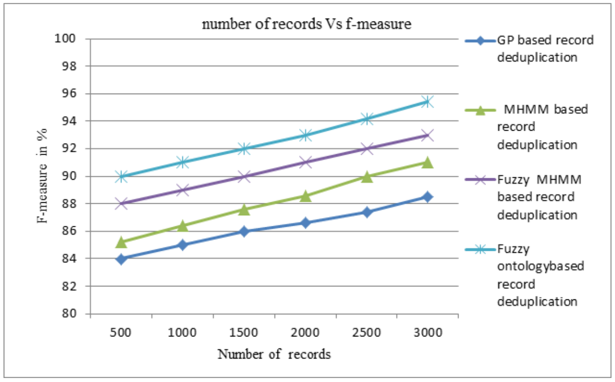

**Figure 3.** F-measure Comparison.

## 5.4 Execution Time

The time taken to perform deduplication process is called Execution time.The performance offered by various methods for deduplication was analyzed and compared. Here if the no of records are increased the Execution time can also be decreased linearly while deduplication process. In this Figure 4, the x axis plots the number of records and the y axis plots the execution time. The following Execution time graph shows that Fuzzy Ontology based record deduplication has lower Execution time compared to all other system.

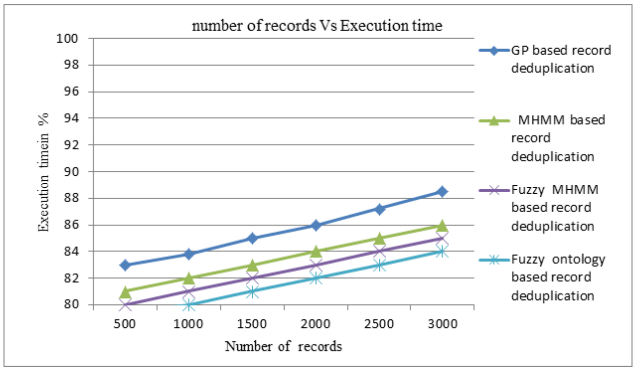

**Figure 4.** Execution time Comparison.

### 5.5 Accuracy

Accuracy is evaluated as,

$$Accuracy = \frac{(Truepositive + Truenegative)}{(Truepositive + Truenegative + Falsepositive + Falsenegative)}$$

In this Figure 5, the x axis plots the number of records and the y axis plots the accuracy value. Here, if the no of records are increased the accuracy may also be increased, linearly while deduplication process. The following accuracy graph shows Fuzzy Ontology based record deduplication has highest accuracy compared to all other system.

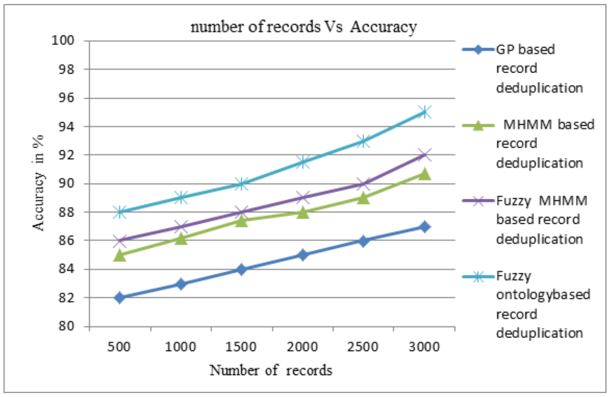

**Figure 5.** Accuracy Comparison.

## 6. Conclusion

Joint duplicate record identification is done by using Multiple Hidden Markov Models (MHMM) which is used to detect the duplicate records in relational data base. For accurate detection of duplicates, improved MHMM Clusters by using fuzzy approach is used. It is trouble to discover semantic similarity among the records and clustering of records from the heterogeneous sources is often difficult. To solve this Fuzzy Ontology based record deduplication is used which improves the record deduplication result by relating the semantic relatedness between records through the construction of Fuzzy Ontology. From the constructed ontology, the deduplication function efficiently identifies the records of replicas and duplications. And also, it improves the accuracy result on the record deduplication by without affecting the quality of the final solution.

## 7. References

1. Karthigha M, Anand SK. A survey on removal of duplicate records in database. Indian Journal of Science and Technology. 2013 Apr; 6 (4):4307–11.
2. Gujar PP. A survey of record deduplication techniques. IJLTET. 2013 Jul; 2(4):246–50.
3. Karunya MR, Lalitha S. Evolutionary Innovations in record deduplication. 2013 Nov; 2(11):766–70.
4. Khan B, Rauf A, Javed H, Khusro S, Javed H. Removing fully and partially duplicated records through K-Means clustering. IACSIT. 2012 Dec; 4(6):750–54.
5. Subi S, Thangam P. An optimized approach for record deduplication using mbat algorithm. International Journal of Engineering and Computer Science. 2013 Jun; 2(6):1874–78. ISSN: 2319-7242.
6. Devi1 LC, Hansa SM, Babu GNKS. A genetic programming approach for record deduplication. International Journal of Innovative Research in Computer and Communication Engineering. 2013 Jun; 1(4):766–70.
7. Yamini W, Mohanpurkar A. Review on record LINKAGE and deduplication based on suffix array indexing. International Journal of Computer Applications 2014 Dec; 108(6):28.
8. Gayathri R, Malathi A. Exploration of data mining techniques in record deduplication. IJSR. 2013 Nov; 2(11):216–19. ISSN (Online): 2319-7064.
9. Jeong H, Jeong H. Ontology-based Integration and refinement of evaluation-committee data from heterogeneous data sources. Indian Journal of Science and Technology. 2015 Sep; 8(23):207–21.
10. Wang X, Hamilton HJ, Bither Y. An Ontology-Based Approach to Data Cleaning. 2005 Jul; 137–52. ISSN: 0828-3494.
11. Khan B, Rauf A, Javed H, Khusro S, Javed H. Removing fully and partially duplicated records through K-means clustering. IACSIT International Journal of Engineering and Technology. 2012 Dec; 4(6):750–54.
12. Carvalho MGD, Laender AHF. Genetic programming approach to record deduplication. IEEE Transactions on Knowledge and Data Engineering. 2012 Mar; 24(3):399–412.