IMPROVE THE QUALITY OF STATISTICAL METHOD OF OBTAINING REPRESENTATIVE DATA SCHEME FOR DE-DUPLICATION USING FUZZY CLUSTERING AND GENETIC ALGORITHM

[1]RAVIKANTH.M, [2]DR.D.VASUMATHI

[1]Associate Professor, Department of CSE, CMR Technical Campus , Medchal, Telangana, India
[2]Professor, Department of CSE,JNTUCE, Hyderabad, Telangana, India

E-mail:ravikanthm.cse@gmail.com,vasukumar_devara@yahoo.co.in

ABSTRACT

Record De-duplication is the important task under merging different database records. We can provide tuning results to the users after implementation of de-duplication operation. Existing approaches are failing under tuning of web databases and removal of duplicate records. All existing approaches are not providing efficient and effective results [1] [2] [3] [4]. In this paper we are designing one new prototype discussion related to effective and enhanced de-duplication. Prototype design starts with fuzzy clustering and genetic algorithm. Its can control more number of duplicate records compare to other approaches. Its saves more storage and time compare to other approaches [12] [13]. In distributed databases the complexity of finding similarity factor is very high. The existing techniques are not accurate to minimize the duplication in the same data base. In the present work a new technique is proposed to improve the accuracy level [24]. In the proposed work a multi-level technical process implemented like tuning. The tuning technique finds all types of duplicated documents in the database. Here all duplicate files are searched with all attributes in sequential order in tree fashion. The results are further improved and reached to an optimized and acceptable range with new data duplication detection method with Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). It further removes unwanted residual files from the database. Bases on the view of previous ranking system problems a new manifold ranking is proposed in the current research work. In the proposed system the ranking is evaluated with new multimodality manifold ranking with sink points.

Keywords: *Web Databases, De-Duplication Operation, Un-Supervised Duplicate Recognition, Edit Distance Algorithm, Fuzzy Clustering Algorithm, Genetic Algorithm, Margin Relevance.*

1. INTRODUCTION

Data Sciences explores its branches in many directions. It has many wings which waves its benefits to many of real time applications. It has extended its topographies to many areas like Intelligent Data Analysis, Big Data Analytics, Data Mining, Information Securities, Data Base management Systems, Internet, Web Database, Internet of Things, and many more. They are facing many challenges in corporate fields, economic field, and in daily life. Researchers have much interest in the field of data analysis. The data analysis era starts long back when organizations start data sharing between computers in large networks. During data analysis people start realising about data duplication problem. As the network area increasing the process of data uploading and downloading also increased, but this increase the multiple data copies in the network systems. This creates large problems for the company databases. Data duplication in data base increased data storage needs also diminished data accuracy. The duplication in such personal organization data bases can be cleared at regular intervals by doing simple identification and deletion process to improve the database searching process. In this process the professions identify the files where they are the copies of the original files which are located somewhere in the database. The identification done by certain parameters like size of the file on the disk, number of terms in the file, sizes of the chunks, and by identifying key words which are similar from the original file. When such type of similar matching is found in any file the file can be deleted or removed from the system hard disk with permissions.

The requisite information can be retrieved from the large data base by combining unified heterogeneous and valid data base from various servers or local systems. While fetching from such data bases it is highly essential to see that fetching records without any duplication the phenomenon is called de-duplication.

But, when come to the storage device different disk places are available in computers. The clearance of duplication process is required both in hard disk and other secondary disks like compact disks. The duplication can always be identified in hard disks whenever a document is trying to copy init. It is a regular process always done in hard disk after a system been logged in, which decreases the system performance. This can be solved by selecting pre-duplication algorithm or post-duplication algorithm at appropriate instance. The selection procedure is different when come to secondary device disks. As the external disks size is small duplication identification process is easy and copy clearance is also very easy. The system softwares itself maintain simple data duplication identification algorithms which they give accurate results in identifying duplicated files. But they are failed in identifying duplication when storage size is increases. In the present work a new record linkage algorithm is proposed to put away the duplication. The proposed algorithm is analysed and compared with various existing and similar type of algorithms. There are few techniques already works to manage duplication. Eliminating duplication from large data bases like internet is very difficult. But, the de-duplication can be applied to network to increase the byte transmission in unit time. During search process the eliminating process must be done and same time it should be produce requested data to client system. This will come about with appropriate de-duplication algorithm. Hence designing power full record linkage algorithm is highly essential in getting optimized search engine. In the present work a new method is proposed along with the discussions made with the flaws of current methods. The demerits of existing methods are overcome with the new method. In the new method both texts based and image based duplication is identified and analysed for eliminating data duplication. So, in this work individual algorithms are proposed to eliminate duplication in both types of data. This type of data is called hyper-text, which contain different types of data.

2. RELATED WORK

The traditional methods of de-duplication conveys as set of independent decisions. In conventional method, for each pair of records, similarity factor is measured. A threshold value is set for the similarity factor. The records will be merged if the similar factor crossed certain threshold value. The similarity rank results a matrix. By considering adjacency matrix, the independent decisions are merged in unified field. In the previous work, McCallum, Wellner, Parag, Domingos have proposed that better results can be achieved than convention method by considering multiple de-duplication decisions in relational database. In relational database records have types and there exists relation between different record types, because, the identity of one entity every so often depend on identity of other record entity. These relations provide evidence to provide de-duplication decision.

The existing methods are called as Conditional Random Fields (CRFs). In CRF the predicted nodes are the de-duplication nodes and the observed nodes are references. The remarks are clustered in the conventional model as an instance of graph partitioning. They are clustered based on their distance from all other partitions. The de-duplication decisions are taken based on the dependent relation with each other. The multiple decision taking will help to overcome inconsistency, errors and noise in the similarity factor. In the present work a multiple decision de-duplication technique is proposed inorder to solve the existing problems in the conventional methods. A new concept is proposed with approximate query processing based on distance specified duplication detection mechanism. Most of the existing techniques worked based on predefined matching rules either hand-coded or matching rules which are learned offline by few learning methods from different training example [1].

3. THE PRESENT IMPLEMENTATION LEVELS OF DATA PROBLEM

The quality of the data is tested at different level of the record. The Data quality levels are categorised into two types according to the constraints present in the data arrived from different sources. They are representing in hierarchy level in figure 1.1. According to the class of the source the Data Quality

Problem is basically classified into two types. They are single source problem, and Multi Source problem. In Single source problem, for instance the web database searches a document from single source in response to the user query. The web database retrieves the document either from the server data base or from co-server. In this case duplication occurs if same file exists multiple times in the server data base. In online marketing the requested data fetched from singe source of server. Multiple entities are processed with same attributes like Id, name and category. The Second category of data quality is derived from Multi sources. The data quality problems are inherited from multi source problems. Here the server processes the user query and produces the records from multiple sources. The requested document will be matched with document present in the requested server and also it retrieves the document from other server data bases. The retrieved documents will be matched with the requested document. All the documents from the server and other sources are matched with the queried data set the sends to the client device. In Multi source system the request will be matched with documents present in different final end database present in multiple systems or sources. Where, in single source process the user query will be matched with the documents present in single final end database.

The Single source problems are again classified into two basic levels. According to several factors against the problem faced in single source record matching it is classified as Schema Level, and Instance Level. The problems provoked due poor schema design of the documents in the record. Due to minimum level of constraints present in the data integrity the server produces poor result after client request. Due to lack of controls there is multiple degradation occurs in retrieved data sets. The degradation is due to misspellings, redundant objects, and abbreviations in poorly framed statement. Due to lack of rules the text is framed with improper schema. The text contains improper representation of graphical symbols, equations, figures, and parameters. The schema also creates severe problems during matching process. There is poor reliability in uniqueness referential of data sets. In Schema level the poor data quality may occur due to improper reference of unique documents. When multiple documents are found, only one document will be referenced and all the remaining duplicate documents will be removed. Hence proper reference is required to remove the duplicated documents. Improper reference will lead poor quality of the data result. So reliable referencing of unique documents is highly essential to achieve good quality in the result.

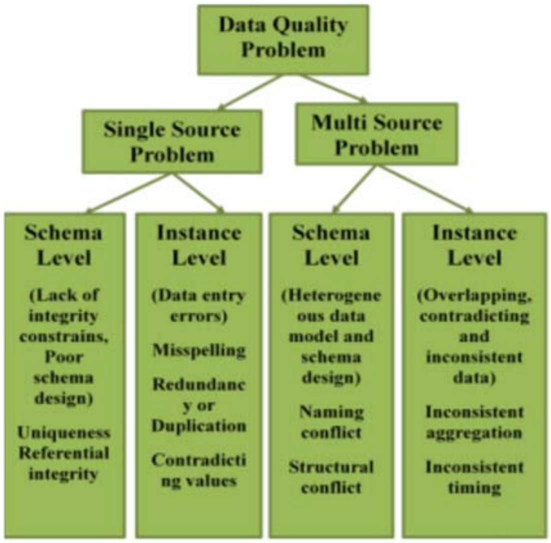


Fig 3.1: Data Quality Problem

Text Learning Techniques

There are many traditional methods present in search process. But, the edit distance and vector space measurement process detects the duplication better than the traditional methods. The edit distance and vector space measurement process are two text learning techniques [28][29].

Edit distance approach

It is a standard dynamic programming problem. For suppose S1 and S2 are two strings, then the minimum number of operations required to process one sting into another string is called edit distance between string S1 and S2.

The edit distance between two strings S1 and S2 is the minimum number of edit operations of single characters needed to transform the string S1 into S2

2. There are three types of edit operations:
   - insert a any word into the string.
   - delete a word from the string, and
   - modify one word with a different character.

3 To employ learnable text distance operations for each database field, and demonstrate that such measures are capable of adapting to the specific notion of similarity that is appropriate for the field's domain.

Algorithm discussion

```
Require: String A of length m, string B of length n
Ensure: Normalizes Levenshtein Edit-distance between A and B
Create 2D array $d_{0..m,0..n}$
    //for all i and j,$d_{i,j}$ holds the Levenshtein distance between the first I characters of A and the first j characters of B: note that d has (m+1)×(n+1) values
for i=0 to m do
    $D_{i,0}$←I //distance of null substring of B from $A_{1...j}$
end for
for j=0 to n do
    $D_{0,j}$←j //distance of null substring of A from $B_{1...j}$
end for
for j=1 to n do
    For i=1 to m do
        If $A_I$ = =$B_j$ then
            Di,j←$D_{i-1,j-1}$ / /no editing required
        Else
            Di,j←min($d_{i-1}$,j,$d_{i,j-1}$,$d_{i-1,j-1}$)+1 // deletion, insertion, substitution
        End if
    End for
End for
$N_{ED}$(A<B)= $d_{m,n}$/max(│A│,│B│)
Return $N_{ED}$(A,B)//Normalized edit distance
```

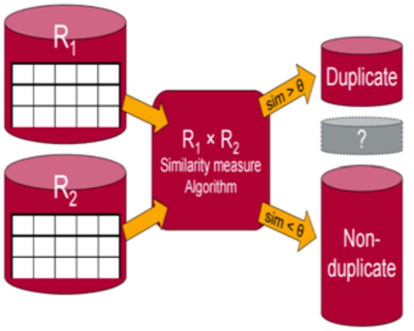


Figure2.1: Similarity Measure Based Duplication Detection Approach

Levenshtein distance It is used to measure differences between two sequences. The differences are measured in a metric Levenshtein. The minimum operations required to change one transaction to other transaction is performed on single character. The operation can be insertion, deletion or substitutions. The number of single character editions is the Leventshtein distance between two words [31] [32][33].

Brute Force Approach

It is also Known as generate and test method. It is a general problem solving technique that contains all enumerating possible entities for solution. Brute force technique is simple technique and finds all possible solutions. Brute force algorithm is a benchmarking algorithm for all other algorithm. The occurrence of brute force algorithm searches for all keys until reuired key is found [34]. The longer key is exponentially more difficult to fissure than the small key.

If the characters to be compared are equal i.e. S1[m] = S2[n], then compare (m+1)th character of S1 to (n+1)th character of S2.
If one character of string S1 is replaced then we will compare (m+1)th character of S1 to (n+1)th character of S2.

If one character is inserted to string S1 then compare m_{th} character of S1 to (n+1)th character of S2.

If one character is deleted from string S1 then compare (m+1)th character of S1 to nth character of S2.

4. PROPOSED MODEL FOR DUPLICATE DETECTION

/

The present work is motivated with several examples. For example, consider research papers in a data base, where records can be of type paper, venue, or author as shown in figure 1.2. For instance, the venue records will be labelled as duplicate if the corresponding records labelled as duplicate. The reverse is slightly true: if two venues are duplicates, then this may slightly increase the probability that their corresponding papers are duplicates. For instance the “author matching” problem exemplifies the scope of the present research work. The “author matching” problem describes the relationships between entities which can improve the quality of reference disambiguation in publication search in a journal or a book. In an independent systematic domain the methods of the analysers are described schematically in the present work. For instance consider a data base which contains both authors and publication entities listed to represent documents. Authors and papers in the journal are represented with special attributes as shown in figure 1.2. For example attribute_id, affiliation_ college, author Name, are used to represent authors of the paper publication. The form_id, title, authorref1, authorrefN are used to represent publication papers. For example consider a toy data base with following author and publication records.

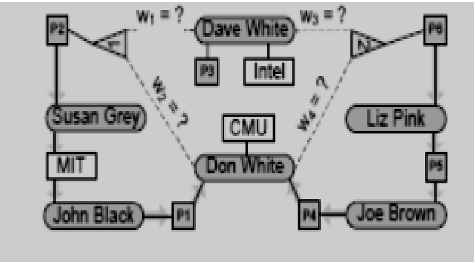


Figure 1.2 Graph For The Publication Example

1. _A1, 'Dave White', 'Intel'_,
2. _A2, 'Don White', 'CMU'_,
3. _A3, 'Susan Grey', 'MIT'_,
4. _A4, 'John Black', 'MIT'_,
5. _A5, 'Joe Brown', unknown_,
6. _A6, 'Liz Pink', unknown_.
1. _P1, 'Databases . . . ', 'John Black', 'Don White'_,
2. _P2, 'Multimedia . . . ', 'Sue Grey', 'D. White'_,
3. _P3, 'Title3 . . . ', 'Dave White'_,
4. _P4, 'Title5 . . . ', 'Don White', 'Joe Brown'_,
5. _P5, 'Title6 . . . ', 'Joe Brown', 'Liz Pink'_,
6. _P6, 'Title7 . . . ', 'Liz Pink', 'D. White'_.

To search a document it is highly required the key word that is either author name or paper title or both. In this problem the main aim is identify the correct author reference in each paper to refer the correct author. An effective existing technique feature-based similarity (FBS) is used to compare the description given in the author filed of the paper. In each paper the author ref values are comparing with author name attribute in authors field. In same manner all author ref values can be resolved for author reference attributes in the above problem. For illustration, the above type of methods would identify that 'Sue Grey' reference in P2 refers to A3 ('Susan Grey'). The only exception will be 'D. Either A1 ('Dave White') or A2 ('Don White') could match references in P2 and P6: 'D. White'. Exploiting additional attributes may disambiguate the reference 'D. White' in P2 and P6. For suppose, the paper p2 and p3 title is not equal and paper p1 and p2 titles are equal, suggesting that 'D. White' of P2 is indeed 'Don White' of paper P1. If we are unable to disambiguate the references using title (or other attributes) there is still possibility of disambiguate in referencing 'D. White in P2' and P6 by analysing relationships among entities. In this journal publication example the 'Don White' has co-authored in a paper with 'John Black' who is an author from MIT. For suppose the author 'dave White' does not have any publication as co-author with MIT authors. These instances can be used to disambiguate between two authors. There is higher possibility of that 'Don White' is the author in paper P2 in the place of 'D.White', since the co-author is 'Susan Grey' in paper P2 named as 'D.White'. In an absence of 'Dave white' and MIT the data suggests a connection between MIT and 'Don White'. There is another instance that, 'Don white' is co-author in Paper (P4) with 'Joe Brown' and also co-author in a paper with 'Liz Pink'. Similarly author 'Dave White' is not co-authored any paper with 'Liz Pink' or 'Joe Brown'. Since 'Liz Pink' is a co-author of P6, there is a higher likelihood that 'D. White' in P6 refers to author 'Don White' compared to author 'Dave White'. The reason is that often all co-author collectively in a network custom several groups/clusters of authors. This group of authors has research relation and may publish papers with each other. The data suggests that cluster contain the authors 'Don White', 'Joe Brown' and 'Liz Pink'.

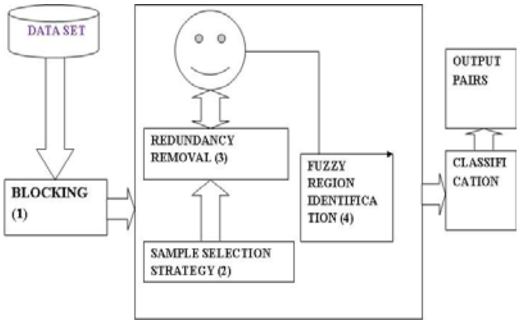


Fig 3: Proposed System Architecture

5. MAXIMAL MARGIN RELEVANCE (MMR) ALGORITHM

Maximal Margin Relevance (MMR)

This technique is used to remove redundancy objects with the help of similarity function. The similarity value is verified in two methods. Firstly, evaluate similarity value between document features. Another way is, evaluate between document and query [1]. These two techniques are used to value similarity value. Earlier diversification technique is correlation.

In Most of the existing methods the documents are sampled from a Euclidean space model [38] [39] [40] [41] [42] [43] [44] [45]. In the current scenario the documents are sampled from a nonlinear low-dimensional manifold which is embedded in the high-dimensional ambient space [46] [47]. In paper [45], MMC is shown to be more effective than LDA. It learns a sub-space, in which the sample is close to the documents in the same class but far from the documents in the different classes

Local Relevance Weighted Maximum Margin Criterion

In the paper [48], proposed a method, called Local Relevance Weighted Maximum Margin Criterion (LRWMMC) for text classification. The LRWMMC inherits all the properties of Maximum Margin Criterion [45] and it does not suffer from SSS. The objective of LRWMMC is to study a sub-space in which the documents are present as near as possible in the same class are while the documents are present as far as possible in the different classes in the local region of each document. Local class information is more discriminative than global class discriminative. In the high dimensional ambient space LRWMMC is able to catch the low dimensional manifold. It is robust when compared with other previous methods.

Algorithm:

Local Relevance Weighted Maximum Margin Criterion [48]

Input: Training set $\{x_i , y_i\}^n_{i=1}$, Local Relevant Region size k, desired dimensionality m;

Output: A ∈ $R^{d\times m}$;

1. Construct the with-class local relevant region and between class local relevant region for each x_i ;

2. Calculate the adjacent matrix W by following equation,

$$W_{ij} = \begin{cases} r(\mathbf{x}_i,\mathbf{x}_j), & \text{if } \mathbf{x}_j \in N_b(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_b(\mathbf{x}_j) \\ r(\mathbf{x}_i,\mathbf{x}_j)-1, & \text{if } \mathbf{x}_j \in N_w(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_w(\mathbf{x}_j) \\ 0, & \text{otherwise.} \end{cases} ,$$

and L = D − W;

3. Calculate the projection A as the m eigenvectors corresponding to the largest m eigenvalues of XLX^T .

Generally the linear method cannot map the documents to a sub space because of non-linearity of the ambient space like the documents are present as near as possible in the same class are while the documents are present far enough in the different classes.

The kernel Latent Semantic Kernel is proposed in paper [50]. Kernel method [49] eases this problem by mapping the input space to a high dimensional space. Then find the subspace of the feature space.

Algorithm 2

Kernel Local Relevance Weighted Maximum Margin Criterion[48]

Input: Training set $\{x_i , y_i\}^n{}_{i=1}$, Local Relevant Region size k, desired dimensionality m, Kernel type, Kernel Parameters;

Output: A ∈ $R^{d\times m}$;

1. Construct the kernel matrix K on the training set;

2. Construct the with-class local relevant region and between class local relevant region for each $\varphi(x_i)$;

3. Calculate the adjacent matrix W by following equation

$$W_{ij} = \begin{cases} r(\mathbf{x}_i,\mathbf{x}_j), & \text{if } \mathbf{x}_j \in N_b(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_b(\mathbf{x}_j) \\ r(\mathbf{x}_i,\mathbf{x}_j)-1, & \text{if } \mathbf{x}_j \in N_w(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_w(\mathbf{x}_j) \\ 0, & \text{otherwise.} \end{cases}$$

and L = D − W;

4. Calculate the projection α_i , $1 \le i \le m$ as the m eigenvectors corresponding to the largest m eigenvalues of KLK.Tensor LRWMMC.

The previous techniques are working based on Vector Space Model (VSM). In recent days, a new technique is proposed, that is Tensor Space Model (TSM) [51]. TSM achieves highest order correlation between words. The Tensor LRWMMC inherits properties of both TSM and MMC.

Algorithm 3

Tensor Local Relevance Weighted Maximum Margin Criterion [48]

Input: Training set $\{X_i , y_i\}^n{}_{i=1}$, Local Relevant Region size k, desired dimensionality J1, J2 . . . , J_N ;

Output: A sequence of projections $U_k \in R^{I_k\times J_k}$;

1. Construct the within class local relevant region and between class local relevant region for each X_i ;

2. Calculate the adjacent matrix W by following equation

$$W_{ij} = \begin{cases} r(\mathbf{x}_i,\mathbf{x}_j), & \text{if } \mathbf{x}_j \in N_b(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_b(\mathbf{x}_j) \\ r(\mathbf{x}_i,\mathbf{x}_j)-1, & \text{if } \mathbf{x}_j \in N_w(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_w(\mathbf{x}_j) \\ 0, & \text{otherwise.} \end{cases}$$

3. Initialize $U_k = I_k$ where I_k is any $I_k \times J_k$ orthogonal matrix;

```
for t = 1 to $T_{max}$ do
  for k = 1 to N do
     (a)Calculate Y \k i by following equation
```

$\mathbf{U}_1, \ldots, \mathbf{U}_{k-1}, \mathbf{U}_{k+1}, \ldots, \mathbf{U}_N$, denote $\mathcal{Y}_i^{\backslash k}$

$$\mathcal{Y}_i^{\backslash k} = \mathcal{X}_i \times_1 \mathbf{U}_1 \ldots \times_{k-1} \mathbf{U}_{k-1} \times_{k+1} \mathbf{U}_{k+1} \ldots \mathbf{U}_N$$

(b)Calculate Y \k i(k) by mode-k flattening of Y \k i ;

(c)Calculate Tk = $\sum_{i,j} W_{ij} (Y_{i(k)}^{\backslash k} - Y_{j(k)}^{\backslash k})(Y_{i(k)}^{\backslash k} - Y_{j(k)}^{\backslash k})^T$

(d)Calculate the projection U_k as the J_k eigenvectors corresponding to the largest Jk eigenvalues of Tk;

end for

end for

6 SIMULATION ENVIRONMENT

The flaws in the above two techniques correlation and similarity function are overcome in Learning Algorithm [3]. Learning algorithm can reduce more noisy objects when compared with correlation technique. It is succeeded in producing required results as user requested a query. It produces accurate results in response to the requested query, when compared with the other techniques. In learning algorithm it allocates rank to the requested documents after going through multiple iteration verification for the requested document. The noisy documents will be eliminated depending on the threshold value. The rank with top-k will be sent in response to the user request. Using Recommendation setting the top-k values are set to avoid noisy objects. After thorough iteration documents are recommended. The recommended documents are interested documents and remaining documents are removed. Based on the comparisons of previous technique results the recommendation technique is proposed to produce more accurate results. When compared with other methods the recommendation technique is reducing some more noisy objects [4][5][6].

Naive Bayesian Classifier

This technique is classified based on Bayesian Classifier. Naïve Bayesian model is very simple and easy to develop and understand. It is very useful for large data set of data bases. It has no dependency on predictor objects. It is used to define real time predicts. Naive Bayesian Classifier derives that a feature declared in the class does not have any relation with the other features. It is widely used in spam e-mail filtering [14][15].

Steps involved in naïve Bayesian algorithm

Step 1: Fetch the data set an map into a frequency table

Step 2: Create similar type of table by finding Overcast probability and probability of playing.

Step 3: Now use Navies Bayesian algorithm to find the best rank. The document with highest rank is the user satisfied resultant document.

Advantages

It need less training data

- It is easy and fast to process the test data
- It perform well with Unconditional

Disadvantages

- The unconditional dataset not observed in training data set, and will unable to make prediction
- The output is not accurate
- It cannot get completely independent set of predictors

Random Decision Tree

It is one of the best supervised models used in learning system. Random decision trees are used to map only no-linear entities. It works both for definite and constant input and output variable. The predictive models produce high accuracy results with tree structured models. Tree based algorithms are easy to learn, they are scalable, robust [18][19][20].

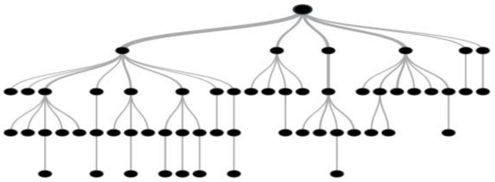


Figure 2.2 Dividing Populations Into Homogeneous Data Sets

There are other methods like forest, and grading methods actively involved along with Random decision tree in big data problems, de-duplication, and etc. In random decision tree model total population is divided into two or more homogeneous data sets as shown in figure 2.2.

Un-Supervised Duplicate Recognition

Query based Multiple Web Databases causes more record duplication. It an Un-supervised Data Duplication detection (UDD)approach which is used to reduce redundant and multiple matching records from query base multiple web data base. In UDD the field weights are computed based on the relative record distance. The relative distance it the major distance among the record in the negative training set. In the first step, the classifier will use the record weights to complement the records fetched from various databases. In the second step, the classifier again finds the further duplicate records. The second classifier uses matching records to identify duplicates from positive and negative set records. Finally all duplicated records and no duplicate records are identified and labelled. From these data sets information adjust the relative area distance. Start the classifier for several iterations to recognise new duplicate data sets. Repeat the procedure with classifier until all duplicate records are detected and no duplicated file can further detected in the web database. In this process there are more chances for quality degradation because of misspelling, abbreviations, redundant entities and conflicting data sets [2].

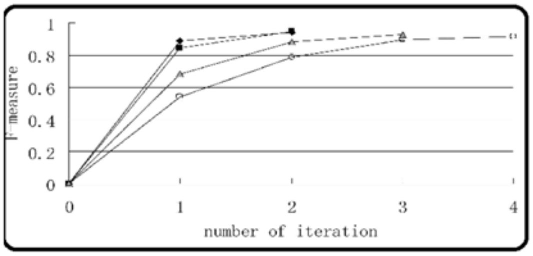


Figure 2.2 Performance Of Proposed UDD Methods

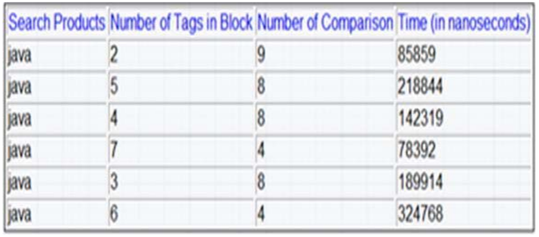
Semi Supervised Clustering. In other de-duplication environment efficient information can be retrieved in semi supervised clustering approach. In this technique the de-duplication achieves minimum object redundancy and recovers efficient datasets from multiple sources [9]. Semi-supervised learning exploits.

7 SIMULATION RESULTS AND PERFORMANCE

Datasets from the Riddle data repository was chosen for the experiment and the datasets used is Restaurant dataset. The datasets, which are used in our proposed approach, is detailed below.

Restaurant Dataset: This dataset consists of four files of 50000 records (400 originals and 100 duplicates), with a maximum of five duplicates based on one original record, and with a maximum limit of two changes in a single attribute in the full record. Cora Dataset: This dataset consists of four files of 40000 records (300 originals and 100 duplicates), with a maximum of five duplicates based on one original record, and with a maximum limit of two changes in a single attribute in the full record.

Table 5: Search Time Results For Different Queries



Search Products	Number of Tags in Block	Number of Comparison	Time (in nanoseconds)
java	2	9	85859
java	5	8	218844
java	4	8	142319
java	7	4	78392
java	3	8	189914
java	6	4	324768

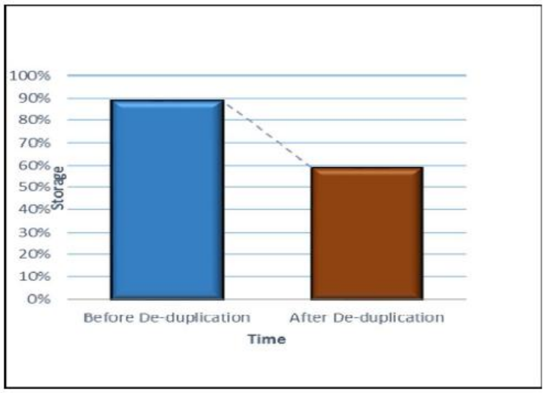
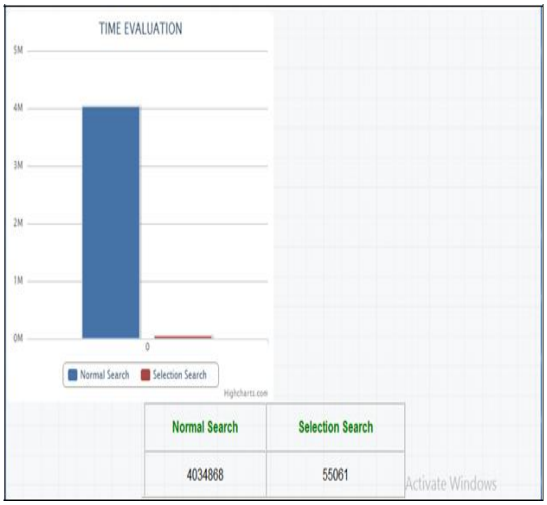


Fig 4. Storage Comparison Graph



Normal Search	Selection Search
4034868	55061

Fig 5. Time Comparison Graph

7 CONCLUSIONS AND FUTURE WORK

Apply the De-duplication matching algorithm to Efficient De-duplication computing model. Combining Fuzzy clustering and genetic algorithms design effective de-duplication computing model. Here in this paper we compare the performance results in between existing and proposed approaches. It's save high amount of data storage and time effectively compare to previous de-duplication methods. In Future we can design to identify the Efficient De-duplication computing model

REFERENCES

[1] A.Arasu, M. Gotz, and R. Kaushik, “On active learning of record matching packages,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2010, pp. 783–794.

[2] A. Arasu, C. R_e, and D. Suciu, “Large-scale deduplication with constraints using dedupalog,” inProc. IEEE Int. Conf. Data Eng., 2009, pp. 952–963.

[3] R. J. Bayardo, Y. Ma, and R. Srikant, “Scaling up all pairs similarity search,” in Proc. 16th Int. Conf. World Wide Web, pp. 131–140, 2007.

[4] K. Bellare, S. Iyengar, A. G. Parameswaran, and V. Rastogi, “Active sampling for entity matching,” inProc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 1131–1139.

[5] Beygelzimer, S. Dasgupta, and J. Langford, “Importance weighted active learning,” in Proc. 26th Annu. Int. Conf. Mach. Learn., pp. 49–56, 2009.

[6] M. Bilenko and R. J. Mooney, “On evaluation and training-set construction for duplicate detection,” inProc. Workshop KDD, 2003, pp. 7–12.

[7] S. Chaudhuri, V. Ganti, and R. Kaushik, “A primitive operator for similarity joins in data cleaning,” in Proc. 22nd Int. Conf. Data Eng., p. 5, Apr. 2006.

[8] P. Christen, “Automatic record linkage using seeded nearest neighbour and support vector machine classification,” in Proc. 14th ACM SIGKDD Int.Conf. Knowl. Discovery Data Mining, 2008, pp. 151–159.