International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

A Semi-Automated Record De-Duplication Technique for a Data Warehouse Environment

Vaishali C. Wangikar, Sachin N. Deshmukh, Sunil G. Bhirud

Abstract: ***Quality of Record de-duplication is a key factor in decision making process. Correctness in the identification of duplicates from a dataset provides a strong foundation for inference. Blocking is a popular technique in de-duplication. In the traditional de-duplication process blocking key is decided by the domain expert. In real time systems, automation of blocking key generation is a primary requirement. Blocking key generation without any human intervention is the objective of this paper. The proposed Automated Token Formation (ATF) algorithm is a fully automated way for blocking key generation. The attributes shortlisted by ATF are almost similar to that of the manual method for all datasets experimented. Datasets like Cora, Restaurant, and FEBRL are used. It is observed that the token provided by ATF has shown 20 % poor results over manual tokens for Cora dataset while for the other two datasets results are matching with manual tokens. A modification is made to ATF to improve the quality of the result by Semi-Automated Token Formation (SATF) algorithm. SATF is a semi-automated approach where training data is needed. SATF has shown better performance over all the manual tokens as well as tokens by ATF.***

Keywords: ***Automated blocking key formation, Record de-duplication, Record Linkage, Semi-automated blocking key generation, automated record linkage, Unsupervised record linkage token formation.***

I. INTRODUCTION

Identification of repeated or duplicate records from the repository of data warehouse is known as Record De-duplication. Data are collected from several sources in the data warehouse. The formats, the conventions, unique keys, primary keys of all these sources do not match. A single record is composed of several different attributes, each attribute belongs to a specific data type, size, and constraint. Two or more records which actually belong to the same entity may not match each other as corresponding type and size of the few attributes do not match because of the different sources. Such records though belong to one single entity are treated as different and segregated.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Vaishali Wangikar*, Research Scholar, Department of Computer Engineering and Technology, MIT Academy of Engineering, Alandi, Pune, India. E-mail: vaishali.wangikar@gmail.com
Sachin Deshmukh, Professor, Department of Computer Science and Information Technology, Dr. BabasahebAmbedkarMarathwada University, Aurangabad, India. E-mail: sndeshmukh@hotmail.com
Sunil Bhirud, Professor, Department of Computer Engineering and Information Technology, VeermataJeejabai Technological University, Matunga, Mumbai, India. E-mail: sgbhirud@vjti.org.in

This false segregation leads to wrong analysis and conclusion which finally affects the quality of decision making. In another case two or more records of different entities are grouped together and labeled as 'duplicates' due to similarity among few attribute values and treated as a single entity which also leads to the false conclusion and affects the quality of the decision. To avoid such false positive and false negative grouping de-duplication or record linkage is used.

The need for de-duplication can be well explained by the example of the healthcare field. In the case of any disease, outbreak government wants to know the number of patients reported with the disease, appropriateness in the count of patients is required for decision making. There are always chances of false positive as well as false negative segregation of patients as one patient might have gone to several doctors and clinics and gets a different patient id. There are chances that a patient might provide inconsistent information. If the patient details are gathered from different hospitals, clinics, doctors it may be possible that information collected have several different formats, conventions for the same record which may lead to false de-duplication.
Following three important steps are in the process of de-duplication

1. Collection of data from all sources.

2. Finding similarity among records on the basis of some key or attributes.

3. Grouping them on the basis of similarity and labeling as duplicates or distinct.

Research work related to De-duplication and Linkage till date is discussed here.

Elmagarmi et al.[1]discuss various similarity match techniques to identify de-duplicates. The similarity is calculated based on the matching of fields or attributes, similarity of tokens, phonetic similarity, and numeric similarity. It is observed that the need for similarity measures varies from data to data. If numeric data are available numeric similarity is preferred, spelling mistakes can be recognized by phonetic measures while characters of strings are treated by tokens.

The similar records are grouped for identification of duplicates. Optimization in the grouping is done by various blocking and indexing methods [2]. A Sorted Neighborhood Method (SNM) proposed by Hernandez and Stolfo[3],[4] is a popular fixed sized sliding window based blocking method which reduces the complexity of de-duplication from $O(n^2)$ to O(n log n).

An adaptive window sized sliding window approach [5],[6] is a step further for SNM approach where the size of the window is automatically set as per the need. It further improves the complexity and number of comparisons for de-duplication.

A lot of research is done on various parameters of de-duplication such as accuracy of de-duplication, reduction in the number of comparisons and flexibility in de-duplication parameter settings. A parallel de-duplication algorithm "FERAPARDA" [7] is used in the anthill parallel environment to attain scalability for huge datasets. Online record de-duplication and de-duplication in the distributed and parallel environment are explored further by Dey et al.[8]. They propose a solution to network bottleneck for record linkage in the distributed environment. A genetic approach for identification of de-duplication predicate is used by Carvalho et al. [9]. The approach works well for specific domain datasets only.The efficiency of de-duplication is always affected by the presence of missing values and spelling mistakes. The issue of missing values is handled by the three techniques such as "Weight Redistribution, Distance Imputation, and Linkage Expansion". [10].Use of Machine learning approach for finding blocking key predicates is proposed by Giang, Phan H. [11].Wangikar et al. [12]provide an optimum method of de-duplication where instead of using all blocking keys values only unique candidates values are selected for de-duplication. This improves this response time of the process. Use of temporal information such as time stamp of record creation or alteration for identification of duplicates is proposed by Hu et al. [13]. It is termed as Temporal record linkage. In the case of online web databases, a traditional approach of supervised record de-duplication is not suitable. A novel method of unsupervised de-duplication using Support Vector Machine (SVM) and Weighted Component Similarity Summing (WCSS) classifiers is proposed to capture on the fly queries and to find online duplicates[14].

In all the above de-duplication algorithms the blocking key/token is set by human experts. Full automation in the process of identification of token for de-duplication is the need of real-time environment.

This research tries to accomplish automation in blocking key generation i.e. token formation.

A recursive Feature selection approach is used for optimization of the token formation. The real datasets like Cora and Restaurant and gold standard data set FEBRL are used for experimentation. Duplicate Count Strategy algorithm (DCS++) proposed by Felix Neumann et.al [6]is used for de-duplication identification in the experiments.

In the naive approach, blocking key is formed with the help of domain experts. For the selection of appropriate tokens, experts need to scan the entire dataset and observe the most suitable attributes of the dataset with their expertise. It is easier for the experts if the dataset contains lesser number attributes and small number of records to find the appropriate blocking key (token) correctly, but in case of huge datasets where the attributes are more with thousands of records, it becomes burdensome to choose a perfect token. The improper selection blocking key leads to poor quality of duplicate identification.

The proposed approach will assist domain experts to select the correct blocking key and assure quality de-duplication for better decision making.

Following is the review of the work in the unsupervised de-duplication process.

II. LITERATURE SURVEY

Unsupervised De-duplication using Cross-Field Dependencies is given by Robert Hall and et al.[15]. Dependencies between two attributes values are the basis of de-duplication. Title and Venue dependency are modeled for the experimentation. Dirichlet-multinomial model" is used over titles and on the exchangeable string-edit model over venues. Domain-specific de-duplication is presented.

Automatic Blocking Key Selection for Duplicate Detection based on Unigram Combinations is contributed by Vogel and Felix Neumann[16]. Automatic blocking key formation techniques are applied for gold standard datasets and further, the same blocking key is re-used for non-gold standard datasets of the same domain. Unigram indexing approach is used for automatic key generation. Almost all attributes are used in blocking key formation. Huge set of blocking keys are generated, each one is tested for blocking quality. Also, a trained supervised dataset is required. Thus fully automated, unsupervised approach is not achieved.

An Unsupervised Algorithm for Learning Blocking Schemes is proposed by Kejrijwal et al.[17]. Indexing functions are provided manually to form specific indexing functions (SIF) and to construct specific blocking predicates (SBP). Windowing approach is used instead of approximate similarity functions, tokens are compared one by one and stop words are found and removed for similarity match. Term frequency and record frequency are used for similarity checking. DNF blocking scheme, a two-step process is used where the pseudo training set is generated followed by feature selection. Term frequency of the matching records is calculated initially, then the total term frequency of both the matching records is calculated. A block is made on each frequent token and window of fixed pre-decided size slides over the records and TF-IDF scores are recorded. TF-IDF scores are compared with the scores in the block if the score is less than the lower threshold then it is marked as non-duplicate else it is marked as duplicate. To improve the efficiency fisherman's algorithm with a greedy approach is used. This approach is useful for the unlabeled dataset. The setting of blocking function by human expert makes it unsuitable for real time environment where fully automated blocking keys are expected.

Unsupervised blocking key generation for real-time entity resolution is recommended by Ramdan et.al.[18]use multiple keys with multi-pass sorted neighborhood algorithm. Ramdan et al. follow Kejriwal, Fisher Discrimination algorithm is used for solving real-time queries of de-duplication. The researchers have ignored automated key formation rather focused on real-time de-duplication.

A novel ensemble approach is used for unsupervised record linkage by Anna Jurek et al.[19]. An ensemble of several self-learning models is proposed by using string similarity metrics.

This algorithm assisted to select the most appropriate similarity match algorithm. Automated blocking key formation is not focused by researchers.
For Arabic record de-duplication, Alian et al. (2018) use dynamic aware Inverted Index approach of Banda Ramdan et al. (2015).Telephone directory data is used for experimentation.
The result shows true positive coverage of 71.13 % for unsupervised tokens provided that there is only one attribute is having inconsistent data but shows only 26.99 % of true positive result when 5 attributes are corrupted in the data. Thus the results do not show consistent results.

After review of the existing research work the research gap found is that in the de-duplication process the token key formation is a manual step. Domain experts are required to set a correct token for the dataset. In the real-time environment availability of domain experts may delay the process real-time de-duplication so an unsupervised automated token formation becomes necessary.

The proposed research focuses on automated blocking key formation for structured dataset irrespective of domain. Following is the detailed explanation of the approach.

III. AUTOMATED TOKEN FORMATION

Hypothesis for ATF

The hypothesis for ATF is, In any dataset, attributes which have more unique values and less null (not known) values have better candidature for blocking key than rest of the attributes for the de-duplication process.
The following algorithm depicts step by step process of proposed ATF.

Algorithm 1. Automated-Token Formation (ATF)
Input:
D= $\{a_1,a_2,\ldots a_n\}$. Where a_1 ... a_n are the attributes of the Dataset D.
T = $\{t_1,t_2\ldots t_n\}$ where $t_1\ldots t_n$ Types of attributes such that $\{a_1\ldots. a_n\} \in \{t_1,t_2\ldots t_n\}$
M ← number of tuples in the dataset
Output:
Automated tokens for a dataset D is T $\{a_i\ldots a_n)$ where T is a set of attributes which forms a token.
Method:
Dcount ←0 //Dcount is Distinct count of attributes
Ncount←0 // Ncount is Null count of attributes
1. For i=1 to n do

1.1 Distinct_Count (a[i],M)
//distinct count function forith attribute
1.2 Null_Count(a[i],M)
// null count function forith attribute.
2. For i=1 to n do

2.1 Max_D_← countmax (Dcount[i];
//maximum distinct count of each attribute.
2.2 Min_N_count← min(Ncount[i];
//minimum distinct count of each attribute.
/* for selecting attributes having more distinct count lesser null count*/
3. For i=1 to n do

3.1 If ((Dcount[i] > α) &&(Ncount[i]< β))
Selected_column[j]=>a[i];
//where α is Acceptable threshold for distinct count and β is Acceptable threshold for null count
3.2 j=j+1;
4. Sort (selected_coulmn[] by Dcount ,Dsc , Ncount , Asc)
5. Create token ;
6. Return;
ATF algorithm provides the attributes having maximum distinctness and less missing values.

IV. EXPERIMETNAL EVALUATION OF ATF

The datasets and the essential characteristics of them are shown in the table 1.

Table1Datasets used for Experiments

Sr. no	Dataset	Specifications	Number of Records
1	Cora	Real, bibliographic dataset	1296 records
2	Restaurant	The real dataset, Gold Standard	865 record, Uniform Distribution
3	FEBRL	Gold standard dataset	1000 records of Normal distribution

ATF is evaluated for CORA dataset which is a bibliographic real dataset, Restaurant, a gold standard real dataset, and FEBRL a gold standard dataset.

For the experimentation of ATF random sampling is used. The size of samples is selected based on the size of the population, the margin of errors and confidence level. 90-95% of confidence level with 5-7 % of margin of errors is used for experiments.To validate sampling strategy Pair Completeness values of three samples and the entire dataset are compared. Fig. 1 shows the Comparison of Pair completeness values of Sample Versus Entire Dataset. It is clear from Fig. 1 that sample represents the population.
The Pair Completeness (PC) is given by the formula

$$PC=\frac{Number\ of\ identified\ \ duplicate}{Total\ Number\ of true\ duplicates}$$

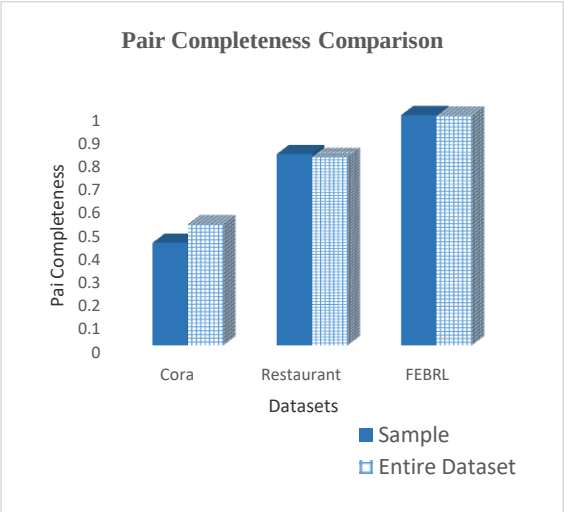

Fig.1Comparison of Pair completeness values of Sample Versus. Entire Dataset

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

For Cora dataset Venue, Author and Title are the candidate attributes selected by ATF to form a token. ATF gives first preference to 'Venue' being a highest distinct attribute, second highest value for distinct count is for 'Author' with zero null count, so selected as a second preference and the third shortlisted candidate is 'Title' with third highest distinct value and the third least null value.

Token formed by ATF is concatenated values of all three shortlisted attributes. It is observed that pair completeness (PC) for Cora dataset by ATF Token is 0.33 (i.e. 33 % true positive coverage) while PC value for the manual token is 0.55. In manual de-duplication token formation process, the token used is concatenation of first four char of author name and year.[5] [6].

Table 2 shows attributes shortlisted as tokens by ATF and attributes shortlisted by manual tokens. It can be observed from the table 2 that ATF has provided similar attributes as that of manual tokens. Thus ATF shows success in giving similar options to that of a domain expert. The comparison of pair completeness values by ATF token versus manual token for Cora, Restaurant and FEBRL dataset is shown in Fig. 2.

Table 2 ATF and Manual tokens for datasets

Dataset	ATF Token attributes	Manual Token attributed
Cora	Venue, Title, Author	Author (4 char), year
Restaurant	Name, Address	4 char of name, 4 char of address
FEBRL	Soc_sec_no, phone_number, Surname, name	Name, Surname

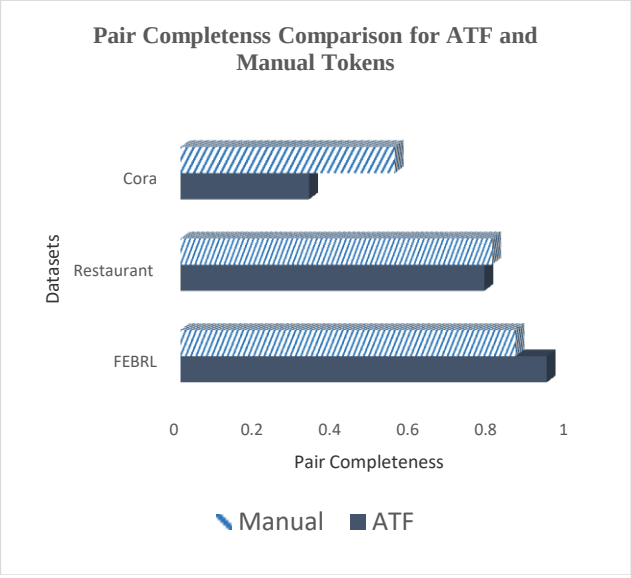

Fig.2Comparison of Pair Completeness by ATF and Manual Tokens

For Cora dataset pair completeness given by ATF is 20% less than manual token formation. For Restaurant dataset, ATF has given Name and Address fields as candidates for token formation, the result shows manual tokens slightly outperforms the ATF tokens. ATF has performed better than a manual token for FEBRL dataset.

To understand the differences in the results it is necessary to know the characteristics of the datasets.

It is observed that Cora dataset has more than two duplicates per original record. Also, it has more than one modifications per attribute thus it is more error-prone. Cora dataset has Zipf distribution where, there are a few clusters with more than 100 records, while many clusters have less than 50 records. There is a large variation in duplicate clusters. The reason for the poor results for Cora is, maximum inconsistent values present in the attribute 'Venue' due to this it shows most distinctness and get selected by ATF as a first choice. The incorrect choice of token leads to poor results for pair completeness.

Restaurant and FEBRL datasets have performed better with ATF as compared to Cora as both have a normal distribution for data duplication as well as the error rate. For Restaurant dataset, at most one duplicate record is present per original record and more than one modifications done to each record. While for FEBRL only one duplicate record is added with a maximum of one modification done each attribute.

After experimentation ATF shows following few notable observations:

- ATF finds out the best candidate attributes for token formation and provides better directions for token formation.
- All the attributes selected for token formation may or may not be required for the token (superfluous attributes). Few attributes are unnecessary which increases unnecessary comparisons.
- The proper sequence and combination of attributes remain unattended.
- No assurance for the quality of tokens(Few tokens of the dataset may give best tokens but few may not)

V. SEMI-AUTOMATED TOKEN FORMATION(SATF)

To overcome the problems of ATF, Wrapper Feature Selection (WFS) is used to produce more accurate tokens. It is based on the relevance feedback mechanism[20]. WFS is used for deciding the number of features to be retained or eliminated. This method uses a classifier to evaluate the subsets by their statistical accuracy. It trains a model using each feature from the feature set and measures the performance of each attribute. It provides the model by giving a set of optimum features and eliminating the irrelevant one. The principle of WFS is used for optimum token formation process. It is a semi-automated approach for token formation. Automated token formation algorithm is a baseline for further experimentation. Attributes shortlisted by ATF are used one by one as a blocking key for record de-duplication. DCS++ is used for de-duplication[6].Random groups of samples from the datasets are used for training. PC values are calculated for each sample. The mean of PC values of samples is recorded for each attribute. The attributes with the values greater than the mean of PC value are selected as a candidate for the token.

In the experiments, the window size of 5 and 80 % threshold is taken for record de-duplication process.

The detailed stepwise description of the SATF is depicted in Algorithm 2.

Algorithm 2 Semi-Automated Token Formation –SATF ($a_1 \ldots a_n$)

// (A semi-automated approach)

Input:

ATF-Token (attribute list) formed by the ATF algorithm.

Output:

The optimal token with a minimal number and proper sequence of attributes.

Method:

R ← no. of attributes in the ATF token

S← Sample from the dataset

n← no. of samples taken

θ ← Threshold for similarity match for DCS++

w← window size for DCS++

PC[S[i], R] ←PC value for i^{th}sample for each attribute R

1. For i=1 to n do
2. For j=1 to R do

// find the de-duplicates using DCS++

2.1 **DCS++(S[i], θ, w, j); // Call to DCS++ function**

//Duplicate count strategy DCS++ algorithm

2.2

$$PC[i] = \frac{\text{(No.of true duplicates identified for sample I and attribute j)}}{\text{Total number of true duplicates in sample i}}$$

2.2 μ← Mean of PC[s[i], j];

2.3 SATF_token ← PC[s[i], j} > μ

3. Return

Function 1 Duplicate Count Strategy- **DCS++ (**S[i], θ, w ,bk**)**

Input: Sample S of size Si
bk← blocking key
w←5 //w is window size
ϕ ←Avg_no. of_comparisons
θ ← 85% // a Similarity threshold

Output: De-Duplicate Groups

Method:

1. Select the key for sorting record and sort the records
2. Create a window with initial window size *w*
3. Compare the first record with all other records in the window within the threshold
4. While $((\frac{\text{No.of identified duplicates}}{\text{No.of Comparisons}}) >= \phi) \&\ (<= 1/w-1))$
   4.1 Increase the size of the window
   4.2 Slide over the window
4.3 Calculate the transitive closure
5. Return de-duplicate groups for blocking key bk

A. SATF on Restaurant dataset

For Restaurant, dataset ATF algorithm shortlist Name and Address attributes for the token formation based on more uniqueness and less null values.

SATF removes the superfluous attributes of ATF. Three different random group samples of size 100 are taken for the experiment.

Mean of Pair completeness values for each attribute for all three samples is shown in Fig. 3 Thus mean of PC value is 0.755

The tokens whose mean PC values are greater than 0.75 are selected for token formation. So here 'Name' is selected as a token for record de-duplication.

When 'Name' as a token for duplicate detection is used for entire restaurant dataset, 92 duplicates identified out of 112 true duplicates from the dataset. The pair completeness achieved is 0.821.

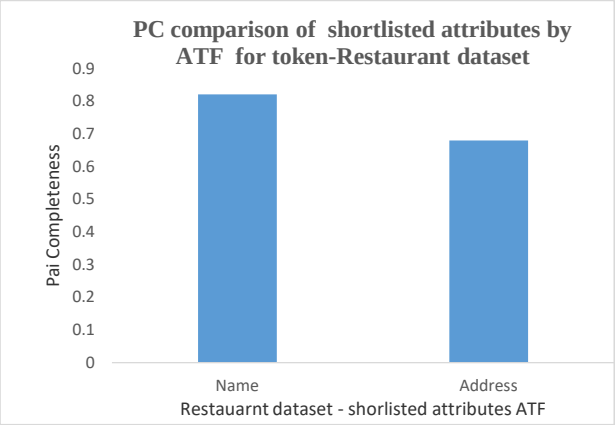

Fig.3Pair completeness comparisons for attributes shortlisted by ATF token for Restaurant dataset

B. SATF on Cora Dataset

The ATF has given Venue, Author, and Title as a shortlisted token for Cora dataset. Mean Pair Completeness value of Cora samples is 0.435, attributes whose pair completeness is more than 0.435 is selected as blocking tokens by SATF algorithm. Thus Title is selected as a shortlisted blocking token. Fig. 4 shows the comparisons of Pair completeness values for the tokens created by ATF.

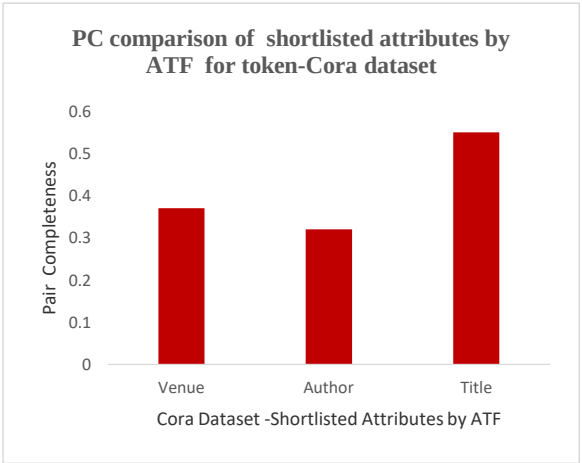

Fig.4Pair completeness comparisons for attributes shortlisted by ATF token for Cora Dataset

C. SATF on FEBRL dataset

The experiment is conducted on FEBRL dataset with 1000 records with duplicates. ATF has shortlisted 4 attributes for token formation as soc_sce-id, phone_number, DOB, Surname.

Fig. 5 shows comparisons of PC values of shortlisted token attributes of FEBRL dataset. The attributes less than mean of pair completeness values are to be discarded rest all the remaining are accepted for token formation. Thus Soc_sec_no,phone_number, and surname are shortlisted for token formation. The same token is been used to find out true duplicates. It finds 455 duplicates out of 469 in the given dataset. The pair completeness achieved by the token is 0.97. Manual tokens used in the comparison are already used by many researchers during their record de-duplication research. Cora has first four characters of author and year as a manual token. Restaurant dataset has the concatenation of first four characters of name and address, while FEBRL has given name, surname as manual token. From the experimental evaluation, it is observed that SATF gives good quality tokens for all the three datasets depicted in Fig. 6.

The attributes shortlisted for the token by ATF and SATF algorithms are shown in Table 3.

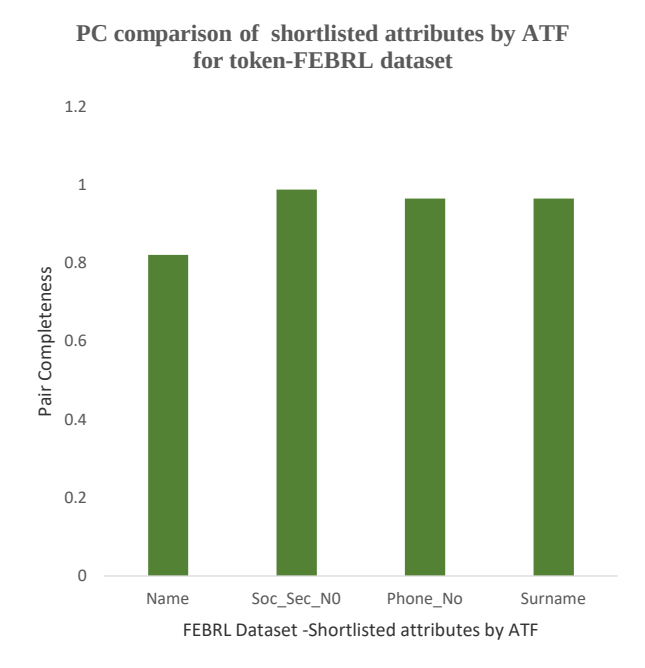

Fig.5Pair completeness comparisons for attributes shortlisted by ATF token for FEBRL Dataset.

Table 3 ATF and SATF token for datasets

Dataset	ATF Token attributes	SATF attributes
Cora	Venue, Title, Author	Title
Restaurant	Name, Address	Name
FEBRL	Soc_sec_no, phone_number , Surname, Name	Soc_sec_no, phone_number, surname

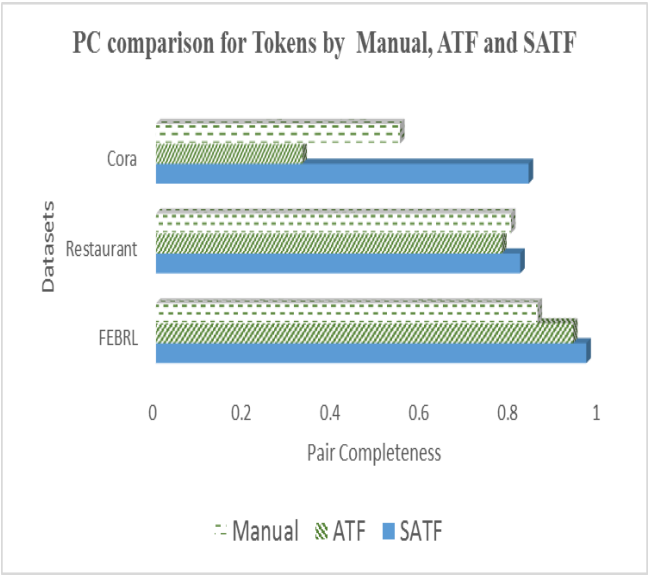

Fig.6 Comparison of Manual, ATF and SATF token for Cora, Restaurant and FEBRL datasets

VI. CONCLUSION

The automated token formation is always a need for real-time record de-duplication. In the initial experimental evaluation, it is observed that ATF a fully automated approach shows poor results compared to the conventional manual method. Rather it deteriorates the true positive coverage. There are many issues in ATF like the number of attributes as tokens, the exact sequence of attributes, superfluous attributes in a token. To overcome these problems, a semi-automated feature selection based approach SATF with random sampling is proposed. Though SATF is semi-automated semi-supervised approach as compared to ATF it gives better results than manual tokens.

The limitation of the SATF approach is it is a semi-supervised approach, needs supervised training dataset thus not appropriate for fully automated on-line de-duplication process.

The proposed semi-automated method assists user or domain expert in the selection of an appropriate list of attributes for token formation though not suitable in a real-time environment where no human intervention is expected.

SATF provides optimum attributes which assure better true positive coverage for record de-duplication. Random sampling approach also helped to reduce the comparison space.

Making an unsupervised fully automated token useful in real time environment will be the next direction of research.

REFERENCES

1. A. K. Elmagarmid and S. Member, “Duplicate Record Detection : A Survey (shorter version),” vol. 19, no. 1, pp. 1–16, 2007.
2. P. Christen, “A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication,” pp. 1–20, 2011.
3. M. A. Hernández and S. J. Stolfo, “Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem,” *Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 9–37, Jan. 1998.

4. S. J. Stolfo and M. A. Hernandez, "Real-World Data is Dirty: Data Cleansing and The Merge/Purge Problem," *Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 9–37, 1998.
5. S. Yan, D. Lee, M.-Y. Kan, and L. C. Giles, "Adaptive sorted neighborhood methods for efficient record linkage," *Proc. 2007 Conf. Digit. Libr. - JCDL '07*, p. 185, 2007.
6. U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "Adaptive windows for duplicate detection," *Proc. - Int. Conf. Data Eng.*, pp. 1073–1083, 2012.
7. G. P. Pezzi, M. C. Cera, E. Mathias, N. Maillard, and P. O. A. Navaux, "On-line scheduling of MPI-2 programs with hierarchical work stealing," *Proc. - Symp. Comput. Archit. High Perform. Comput.*, pp. 247–254, 2007.
8. D. Dey, V. S. Mookerjee, and D. Liu, "Efficient techniques for online record linkage," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 373–387, 2011.
9. R. Isele, "Learning Expressive Linkage Rules for Entity Matching using Genetic Programming," vol. 5, no. 11, p. 224, 2013.
10. T. C. Ong, M. V. Mannino, L. M. Schilling, and M. G. Kahn, "Improving record linkage performance in the presence of missing linkage data," *J. Biomed. Inform.*, vol. 52, pp. 43–54, 2014.
11. P. H. Giang, "A machine learning approach to create blocking criteria for record linkage," *Health Care Manag. Sci.*, vol. 18, no. 1, pp. 93–105, 2015.
12. S. G. B. Vaishali C. Wangikar Sachin , Deshmukh, "Study and Implementation of Record De-duplication Algorithms," in *In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies. (ICTCS '16) ACM*, 2016, p. Article 8, 6 pages.
13. Y. Hu, Q. Wang, D. Vatsalan, and P. Christen, "Regression classifier for Improved Temporal Record Linkage," 2012.
14. M. Ravikanth and D. Vasumathi, "Record matching over query results from multiple web databases with duplicate detection," *J. Adv. Res. Dyn. Control Syst.*, vol. 10, no. 4 Special Issue, pp. 2040–2049, 2018.
15. R. Hall, C. Sutton, and A. McCallum, "Unsupervised deduplication using cross-field dependencies," *Proceeding 14th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD 08*, p. 310, 2008.
16. T. Vogel and F. Naumann, "Automatic blocking key selection for duplicate detection based on unigram combinations," *Int. Work. Qual. ...*, 2012.
17. M. Kejriwal and D. P. Miranker, "An unsupervised algorithm for learning blocking schemes," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 340–349, 2013.
18. B. Ramadan, "Indexing Techniques for Real-Time Entity Resolution," no. March, 2016.
19. A. Jurek, J. Hong, Y. Chi, and W. Liu, "A novel ensemble learning approach to unsupervised record linkage," *Inf. Syst.*, vol. 71, pp. 40–54, 2017.
20. B. Ramadan, P. Christen, and H. Liang, *Databases Theory and Applications*, vol. 10837, no. July. 2018.

AUTHORS PROFILE

Vaishali Wangikar, is a Research Scholar at Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. She is working as a Senior Assistant Professor at MIT Academy of Engineering, Pune. She is having19 years of experience in Academics and Administration. Her area of interests are data cleansing, data warehousing, data Analytics and data mining .

Dr. Sachin N. Deshmukh, is a professor of at Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. He is B.E. in Computer Science and Engineering, M. Tech and Ph. d in Computer Science and Engineering from Dr. BabasahebAmbedkarMarathwada University, Aurangabad. He has been awarded his doctorate in August 2010. He has been 23 years of teaching and research and administrative experience. He has more than 150 publications in national and international journals. His Research interests are text mining, sentiment analysis, Artificial Neural networks.

Dr. Sunil G. Bhirud, is working as a Professor at Computer Engineering and Information Technology department, VeermataJijabai Technological Institute (VJTI) Mumbai. He is having vast experience of 29 years in Academics, Research and Administration. His specialization is in mainly Digital signal processing & Artificial neural networks. He has authored more than 100 research articles, journals and guided more than 20 research scholars. He has also handled additional charge as a Registrar, Mumbai University. He has also worked as an advisor for All India Council for Technical Education (AICTE),Delhi.