Exploring Research Pathways in Record Deduplication and Record Linkage

Vaishali Wangikar[1(✉)], Sachin Deshmukh[2], and Sunil Bhirud[3]

[1] School of Computer Engineering and Technology, MIT Academy of Engineering, Pune, India
vaishali.wangikar@gmail.com
[2] Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India
sndeshmukh@hotmail.com
[3] Department of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, Mumbai, India
sgbhirud@vjti.org.in

Abstract. This paper provides a detailed introduction, significance and research progression of record de-duplication (RDD) as well as record linkage (RL) process. The basic study starts with the experimental analysis of various Blocking and Indexing techniques for Record de-duplication process, where Sorted Neighborhood Method (SNM) is found to be the best choice among all the methods. SNM is further improved using Adaptive variants of SNM. The advancements in record de-duplication are further explored and various methods for it are reviewed and implemented. The major two contributions in the unsupervised record de-duplication, FDJ and OATF are implemented and compared where it is observed that OATF which is a completely automated and unsupervised approach performs equally well as compared to unsupervised FDJ approach, where limited automation is achieved.

Keywords: Record linkage · Record de-duplication blocking and indexing · Sorted neighborhood method · Unsupervised blocking key formation · Real time record linkage · Real time record de-duplication · Automated record linkage

1 Introduction

For the correct decision-making process, data need to be collected from several internal as well as external sources. The collection of data, transformation and loading, cleansing of data, detailed analysis and pattern recognition and visualization takes place systematically in a data warehouse framework. As in the data-warehouse repository data are collected from heterogeneous sources having different schema formats and conventions, different data types, different terminologies and also different primary keys. The very first step is to clean the data to assure quality decision making. It is necessary to transform the incoming data into a unified, consistent format for analysis purpose. Proper treatment to the present dirt such as noise, spelling mistakes, missing values, irregular formats,

K. C. Santosh and B. Gawali (Eds.): RTIP2R 2020, CCIS 1380, pp. 360–372, 2021.
https://doi.org/10.1007/978-981-16-0507-9_31

redundant values is necessary. Each type of dirt is removed using a unique data cleansing technique [1, 2]. The removal of redundant, duplicate entries from the dataset is an essential task to make correct inferences from the data. The process of removal of duplicate record entries from the data is termed as De-duplication [3, 4]. The duplicate identification and removal take place by two different ways, one with Record De-duplication and other with Record Linkage. In the Record De-duplication process, duplicate record entries are identified and re- moved from the same dataset, while in Record Linkage, two or more datasets having the similar record types are checked to identify whether different records are pointing to the same real-world entity. Record de-duplication and record linkage are having slight difference in processing, in record linkage, duplicates are identified from more than one datasets while in record de-duplication the duplicates from the single datasets are identified. Record linkage is also termed interchangeably as Entity Resolution. The de-duplication and linkage need same techniques to identify duplicates. Thus the technique used for record de-duplication can be used for record linkage; the only difference is in the number of datasets. Both the tasks would have easy if a common unique identifier (unique key) present for the exact matching process, but in practice, as the data is extracted from different heterogeneous sources availability of universal, unique identifier is mostly not possible. Even though the unique identifiers are present in the different data sources, they have different conventions, so the identification of repeated record entries become difficult. So more than one attributes are needed with near similarity to identify duplicates. This paper focuses primarily on the Record De-duplication process [5, 6] though the results and conclusion by this study are equally applicable to the record linkage process.

Applications of Record Linkage and Record De-duplication

In Record linkage, more than one records are linked together to find whether they relate to the identical real-life entity. The entity mentioned here could be some individual, or some family or some business or any identifiable object. Record linking involves linking and merging of more than one datasets into a single file without duplicates. The applications of record linkage or Record de-duplication can be broadly categorized in two different ways. The first application, where two or more databases are combined to produce a single database and removes the repetitive or duplicate entries from it to assure the uniqueness, for example to identify the same patient counted many times for the same disease. The second application, two or more datasets are combined together to assure the more correct search entries (quality data) from different sources, for example, to find the double beneficiary for the same scheme. Many unique entries are reflected in dataset which are not actually unique and refer to the same entity in the dataset, identification of such entries and removing them for improvement in the data quality is the third application of record linkage [7].

The record de-duplication a well as the linkage quality is dependent upon the selection of correct attribute or attributes for matching, and it also depends upon the type of matching, exact matching or approximate matching selection.

2 Literature Survey

This section presents the work done so far in the field of record de-duplication and linkage.

2.1 Deterministic Record De-duplication

Deterministic Record de-duplication [8] is based on the threshold-based similarity match. It is based on the rules of matching similarity so also termed as rule-based de- duplication. While deciding these rules, types of dirt present in the dataset must be considered. The matching rules need a correct selection of attribute or combination of attributes. It also requires the information about the type of data to be matched such as integer, string, date, image or video. The selection of similarity match function, the threshold of acceptance and rejection for similarity match decide the quality of de- duplication.

The deterministic approach is the most straightforward approach of de-duplication. Although to keep the quality of de-duplication, it is necessary to monitor the feasibility of the rules throughout the system, especially when the new data entries are made to the system. If the continuous data entry changes the characteristics of the dataset, then the rebuilding of record de-duplication rules set is required. Thus it may become an expensive and time-consuming task.

2.2 Probabilistic Record Linkage

Several different attributes are taken into consideration to find the distance similarity of the records. Based on the quality of the match, the weights are allocated to these attributes. The probability of match and non-match is calculated by making use of the associated weights. A group records above a certain matching probability threshold are considered as a match and below a threshold are considered as non-match [9].

Probabilistic algorithms allocate similarity/non-similarity weights to the attributes of the dataset by calculating the averages of the two probabilities called as *U* probability and *M* probability. *U* probability can be defined as the probability that an attribute in two non-similar records are found similar. For example, u probability for a 'day' field in Date of Birth attribute is 1/30. The attributes having the non-uniformly distributed values may have different *U* probabilities, i.e. the attributes with unknown values or missing values. The *M* probability can be defined as the probability of matching entities in a true duplicate pairs, i.e. for near similar strings, where the distance between the strings is low Jaro-Winkler or Levenshtein distance. This value would be 1.0 in the case of a seamless match. The agreement and dis-agreement weights are calculated based on u and m probability.

2.3 Research Progression in Record De-duplication

The geneticist Howard Newcombe [10] has introduced the concept of record link- age and de-duplication. He used odds ratios of frequencies and the decision rules for describing

similar and non-similar record pairs. It is used in many epidemiological applications in health care domain.

$$log2(pL) - log2(pF) \tag{1}$$

Where, pL is the relative frequency of matches and pF is the relative frequency of non- matches. Further, an approximation is provided to the above odds ratio by the following ratio, as the true matching status is not known.

$$log2(pR) - log2(pR)2 \tag{2}$$

Where, pR is the frequency of a particular string (first name, surname DOB, address, etc.).

Fellegi and Sunter [3] follow Newcombe and a formal mathematical foundations of record linkage is provided through their research. The optimality of the decision rules is demonstrated by the researchers. Further the datasets under considerations are classified the datasets into matches, M and non-matches U. the ratios of these probabilities is given by the Eq. (3).

$$R = P(\gamma \varepsilon \Gamma | M)\ P(\gamma \varepsilon \Gamma | U) \tag{3}$$

Where, γ is an arbitrary agreement pattern in a comparison space Γ, $\gamma \varepsilon \Gamma$ is a relative frequency of specific attribute. The ratio R is a matching score (or weight).The decision rule is given by:

If $R > \theta$, then mark the pair as a match.

If $R < \Theta$, then mark the pair as a non-match.

Where, Θ and Θ are the upper and lower cutoff threshold respectively, are determined by a priori error bounds on false matches and false non-matches.

$$\text{If}(\Theta \leq R \leq \Theta) \tag{4}$$

Then, the given pairs are treated as a probable match and marked for manual review.

In case of three matching fields and only simple agree/disagree weights are considered, then a conditional independence assumption is shown as

$$\begin{aligned} &P(agree\ first,\ agree\ last,\ agree\ age \mid M) = \\ &P(agree\ first \mid M)\ P(agree\ last \mid M)\ P(agree\ age \mid M) \end{aligned} \tag{5}$$

Similarly,

$$\begin{aligned} &P(agree\ first,\ agree\ last,\ agree\ age\ = \mid U) = \\ &P(agree\ first \mid U)\ P(agree\ last \mid U)\ P(agree\ age \mid U) \end{aligned} \tag{6}$$

Such conditional independence assumption must hold on all combinations of attributes that are used in matching.$P(agree\ first|U)$, $P(agree\ last|U)$, and $P(agree\ age|U)$ are called as Marginal Probabilities while $P(|M)\&P(|U)$ is called the M and U-Probabilities, respectively. A total agreement weight can be defined as the natural logarithm of the ratio R of the probabilities. The logarithms of the ratios of probabilities

associated with individual attributes are called the Individual Agreement Weights. The M and U probabilities are also referred to as matching fields. In the conditional independence circumstances, the parameters are calculated through the simple Expectation maximization (EM) algorithm [10]. The EM algorithm which finds the maximum likelihood estimates of parameters is termed a 'Frequentist Approach'. It is mainly used in handling unknown or missing data.

EM is a probabilistic approach which needs availability of the training dataset. EM may not be effective if the dataset has typographical errors and missing values in huge quantity.

On the other hand, for the deterministic approach, there is no need of training dataset, although the intervention of human expert is required to provide appropriate rules for matching pairs. If the deterministic approach is used distance-based algorithms such as Jaro distance, Edit distance, Levenshtein distance algorithms must be chosen with the correct threshold. Availability of domain expert and need of distance-based algorithm with a correct threshold is necessary for the deterministic algorithm. Poor choice of a threshold may lead to poor de-duplication.

Apart from the quality of de-duplication, the process of record de-duplication is further refined for optimization in comparison space. As finding the similarity among the different records is the main task of record de-duplication, the naive approach requires similarity to be measured for all pairs in the entire dataset. The process of matching similarity in a pairwise fashion raises the computational complexity quadratically with the size of the input dataset. Therefore scaling is much more needed for similarity matching especially for large datasets. Also, at many a times, the similarity checking tasks become unnecessary, because many of the record pairs are not similar and just add comparison time unnecessarily.

Blocking methods [11] are introduced to improve the efficiency of similarity match process. Blocking assimilates similar records in a group or blocks using certain criteria. Blocking can be based on pre-specified similarity threshold which groups the similar pairs together based on their similarity distances, for example, Jaro-Winkler distance, Levenshtein distance, Jaccard similarity distance [5] etc. Another way of blocking is to sort the records lexicographically on pre-specified blocking key or token and group them for similarity match. Human intervention is needed for both of these tasks in blocking, one for parameter setting and other for the selection of correct blocking key. A detailed overview of several blocking and indexing techniques is presented by Peter Christen [12]. Sorted Neighbourhood method, Q-gram based indexing, suffix array-based indexing, canopy clustering, string map based indexing are few indexing and blocking techniques discussed by Christen.

The researchers further explored the techniques required to handle de-duplication for web-based systems [13]. An unsupervised online record matching techniques are used to handle web-based De-duplication. Network bottleneck caused during online de-duplication is reduced using the decision tree approach [14]. A progressive sorted neighbourhood method and progressive blocking techniques are used to improve the speed of de-duplication in large datasets [15]. A Map-reduce distributed framework is used for implementing parallel Sorted neighbourhood method for improving speed and scalability of de-duplication, especially for large datasets [16]. A Temporal record

linkage approach is used for de-duplication for the records collected over a period of time. For example, in the Customer Care databases, often temporal information is present, such as the time of instance creation or modification. Temporal De-duplication keeps track of changes that happened to a record over the period of time during de-duplication. A regression-based approach is used for temporal De-duplication over the traditional model for temporal datasets [17]. Karapiperis use Bloom filter space and Hamming Locality–Sensitivity hashing for online record linkage. The Bloom filter-based blocking technique improves response time as well as recall which a requirement of online systems [18].

Ma et al. [19] make use of domain type and subtype information of attributes for blocking key formation. A genetic algorithm is used for selecting a correct predicate for de- duplication [20]. A semi-supervised de-duplication is provided using ensemble learning [21]. Unigram based blocking key generation technique is used for automatic blocking key generation by Vogal and Felix Naumann [22]. Fisherman discrimination based blocking scheme is used for unsupervised blocking key generation [23]. A Fisher Dis-Junctive Dynamic Sorted Neighborhood Indexing method is used for unsupervised blocking key formation and extended use of it for real time record de- duplication [24]. Two different semi-supervised approaches are proposed based on recursive feature elimination and rough set approach by Wangikar *et al.* [25, 26]. A fully automated blocking key formation is proposed by Wangikar *et al.* using the relevance feedback mechanism. A real time de-duplication framework is also proposed by the researchers [25].

Thus the record de-duplication research work addresses several issues such as quality of de-duplication, optimization of comparison space, speed and scalability of de-duplication, de-duplication in large datasets, de-duplication in temporal and web-based datasets, Online de-duplication, semi-supervised as well as unsupervised de-, real-time de-duplication. Thus from basic to advanced techniques in de-duplication are discussed by several researchers.

3 Blocking and Indexing Techniques

Peter christen studied and reviewed various blocking and indexing methods. The methods like sorted neighborhood, q gram indexing, canopy clustering, suffix array indexing, string map based indexing are discussed and implemented in this section.

In Sorted Neighborhood Method (SNM), a correct blocking key chosen by the domain expert is used to sort the records and the fixed size widow is used for identifying duplicates within it. Due to sorting and windowing, similar records are grouped together, and comparison space is reduced to the window size, which provides faster de-duplication and reduces complexity [27]. With fixed-sized window, there is always a possibility of missing duplicates, if the size of blocking window is smaller than the number of actual duplicates. On the other hand, there will be unnecessary comparisons if the blocking-window size is larger than the duplicates present. An adaptive approach of SNM uses flexible window size to overcome the issue of fixed window size and provides better efficiency for de-duplication. Adaptive SNM has two approaches Accumulative Adaptive (AA-SNM) and Incrementally Adaptive (IA-SNM) [28]. Felix Naumann *et al.* put forth an alternate adaptive SNM approach which have improved SNM in the flexibility of window size called as Duplicate Count Strategy (DCS) [29].

The Canopy blocking is proposed by McCallum *et al.* [30]. The similarity match algorithm which allows efficient retrieval of all records within pre- defined distance threshold from a randomly selected record. Canopy centres are chosen randomly for forming Blocks which retrieves similar records within a pre-defined threshold.

Q gram based indexing [31] is another approach for blocking as well as indexing where Q grams are the substrings of length q. During similarity checking of two records, the q-grams of both the sets are matched with one another, and the intersecting q-grams are found out. These number of intersecting q-grams are converted into a similarity using coefficient methods.

In suffix array indexing [32] suffixes are used for blocking purpose. TF/ IDF similarity match techniques are used for making clusters of record for de-duplication. String map based indexing uses the distance between the strings to form the group.

A comparison of all blocking and indexing methods with respect to response time is shown in Fig. 1. From the Fig. 1 it can be concluded that SNM outperforms all remaining blocking methods.

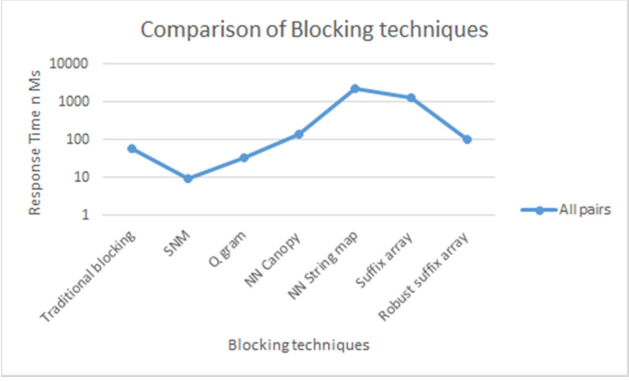

Fig. 1. Response time comparison (log scale) of different blocking and indexing techniques

It has been observed that SNM has gain popularity due to fast response time as compared to other methods. Though SNM has fewer limitations such as unnecessary comparisons when the window size is more than the potential duplicates and missing few duplicates when window size is less than the potential duplicates.

4 Adaptive Variants of SNM

The limitations of SNM are addressed through Adaptive SNM approach. Accumulative Adaptive SNM (AA-SNM), Incrementally Adaptive SNM (IA-SNM), Duplicate count strategy (DCS, DCS++) [28, 29] are all adaptive variants of SNM.

In IA-SNM the boundaries between the two adjacent windows are found out using distance threshold. Window enlargement and retrenchment are performed according to the threshold criteria, and the non-overlapping windows of different sizes are made for comparison [28].

In AA-SNM minimum window size is set initially and the window is enlarged till the records follow the similarity distance threshold criteria. Thus many windows are accumulated to form an enlarged window. The Retrenchment is done to fit only desirable records in the window. Thus a flexible window is created for comparison [28].

For DCS, the blocking-window size is determined on the basis of the already marked duplicates for the window. If more duplicates found, it enlarges the blocking-window to accommodate them. If no duplicates found in neighborhood records, it is assumed that there are no duplicates present and window boundary is decided [29]. DCS++ is a refinement to DCS approach where instead of on adding every duplicate record in the window, the next (w-1) records are added where w is the window size. Transitive closure is calculated to save comparison time [29]. For experimentation two real datasets Cora and Restaurant are used.

A comparative analysis of SNM with all variants of Adaptive SNM is depicted in Fig. 2.

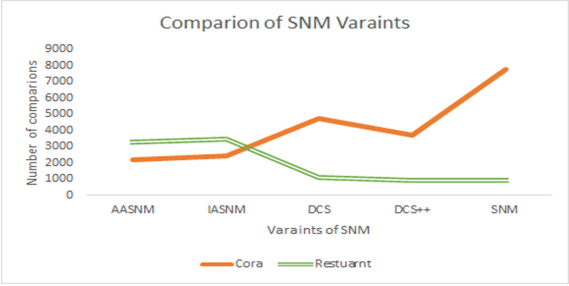

Fig. 2. Comparison of all variants of Sorted neighborhood method

It is observed that all the Adaptive variants of SNM proven better than SNM, especially in terms of the number of matching steps. SNM needs much more comparisons as compared to adaptive methods. It is seen that DCS++ method outperforms in the restaurant while in Cora dataset IA and AA SNM methods perform well.

The evaluation criteria for the de-duplication process is given by the three parameters Pair Completeness (PC), Reduction Ratio (RR) and F-Score.

PC is a measure of true positive coverage, RR is measure how efficiently the blocking schemes the comparison space. High values of RR shows a reduction in comparison space. F score is a harmonic mean of PC and RR values. The Pair completeness, Reduction Ration and F-score comparison of all adaptive SNM approaches for Restaurant dataset are shown in Fig. 3 while Fig. 4 shows it for Cora Dataset.

It is observed that the performance of all adaptive methods is nearly equal for Restaurant dataset. IASNM has shown better performance over the rest of the methods for Cora data. Due to better performance of Adaptive variants of SNM, many researchers prefer ASNM for the advanced research in Record de-duplication.

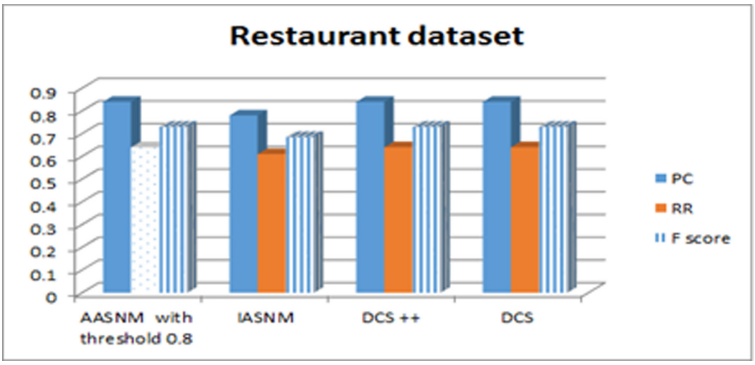

Fig. 3. Comparison of ASNM for restaurant

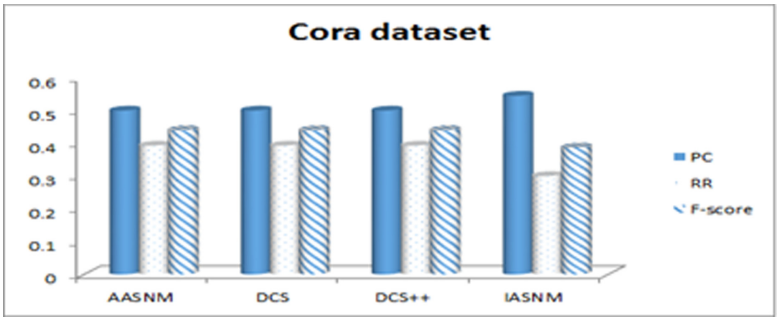

Fig. 4. Comparison of ASNM for CORA

5 Advanced Record De-duplication Techniques

In the current business scenario, maximum businesses are on either on b-based or real times systems. The data is updated and feed to the systems continuously. For making decisions, the fresh real-time data is needed as stale data may hamper decision making. Thus to cater to today's real time systems a scalable, as well as real time de-duplication, is required. Following is the overview of some unsupervised RDD methods which can fulfil the requirements of Real-time environments.

Vogal and Felix Naumann propose an approach of automatic blocking key selection based on unigram indexing [22] which is a step toward automatic de-duplication process. A two-step process is followed for blocking key formation, in the first step, all possible uni-keys are generated, with each of these keys de-duplication is per- formed on the training gold standard dataset. The comparison threshold set for acceptance and rejection of keys. If any of these keys exceeds the number of comparisons, it is discarded else the Blocking key quality is calculated. All the accepted keys are sorted according to the Blocking key quality. This is the first step called as training phase. In the second step, all the shortlisted blocking keys of the training phase are validated against new dataset using de-duplication algorithm. The keys do not follow the specified criteria are discarded while the remaining keys are accepted. Although the algorithm is claimed as

unsupervised, it needs a gold standard dataset to validate the similar type of new dataset, thus entire automation of the system is not achieved.

Further, Kejriwal *et al*. [23] explored unsupervised blocking key formation termed as FDJ. The approach is two-step, in the first step, a weakly labelled training set is generated using TF-IDF similarity. From the weakly labelled set, feature vector sets are derived using all pre-specified specific blocking predicates, and a Boolean vector set is made. In the second phase, Fisher discrimination formula is used to find the optimum feature set as a token or blocking key. The fisher based blocking keys performed better than the manual key approach. Though it is claimed as unsupervised but complete automation of the algorithm is not achieved. Human intervention is needed to set specific blocking key predicates.

The research work of Kejriwal is taken forward by Banda Ramdan *et al.* [24]. Ramdan *et al.* use Fisher Discrimination (FD) based record de-duplication termed as FDY-SN approach and used it for Real time Record linkage framework. They use a three-step approach. In the first step, the training dataset is identified. This dataset is used for learning blocking keys. In the second step, TF-IDF similarity match along with special blocking predicates are used to identify the duplicate and non-duplicate groups. In the third step, optimal blocking keys are identified using fisher score, block size and distribution of blocks. A threshold is set to decide the size of the block, the keys which generate block more than the threshold are discarded. Blocks of similar sizes are preferred to avoid the skew.

Thus, the optimal blocking keys made available for real-time record linkage framework. The dynamic similarity aware indexes are used for the real-time framework. Three types of indexes are used Block index (BI), Similarity index (SI), and Record index (RI). Whenever there is any new insertion or modification occurs to the dataset, the attribute values are inserted into RI, based on it, its block is identified, and similarities between the existing and new attributes are calculated and stored in SI. Thus online updations are handled by the real-time framework.

As FD approach which is used by Ramdan as well as Kejriwal needs human intervention for setting specific blocking predicates, also, in FDY-SN approach the optimal size of the blocks is needed to be decided prior by the domain experts. Therefore both these parameter settings make it unsuitable for fully automated real-time environment. While concluding one can say that fully automated blocking keys and fully automated record linkage are remain unattended by both the approaches discussed.

Wangikar *et al.* [25, 26] work on the same line, for unsupervised, automated blocking key formation. A fully automated way of blocking key formation, Optimized Automated Token Formation (OATF) is proposed. It is a two-step approach; in the first step, primary feature set is prepared using distinct and null feature count; in the second step, a recursive feature elimination is used to select the optimal features. The frequent duplicate coverage (FDC) is calculated for each feature, the features having FDC less than the mean FDC are discarded, and the optimal blocking key feature set is made ready. This approach does not need any human intervention for making tokens. However, as it is based on deterministic distance-based de-duplication, the rebuilding of Key formation logic is needed if the data characteristics of the dataset are changed drastically over the period of time.

Wangikar *et al.* use the automated token formation approach further to build Real time De-duplication framework. The Sorted dis-joint indexes (SID) of blocking key values (BKV) are maintained with the repeat count of each index entry. For the new entry of the record, the blocking key value is generated based of blocking key logic, the new BKV is matched with existing BKVs from SID, if the match is found then the repeat count is increased and de-duplication takes place, in case the match is not found in SID then a new entry of BKV is inserted at appropriate place in SID and repeat count is maintained, and de-duplication takes place using DCS++. Thus Automated Record de- duplication is used to build a framework for real-time de-duplication process.

The experimental evaluations of unsupervised tokens formation approaches such as FDJ and OATF with supervised manual tokens for Cora and Restaurant datasets are in shown in Fig. 5 and Fig. 6 respectively.

It is observed that both the unsupervised approaches FDJ by Kejriwal *et al.* and OATF by Wangikar *et al.* outperform supervised manual token. OATF, which is a completely automated approach, perform equally well as that of FDJ approach for Restaurant dataset. OATF shows little low pair completeness for Cora dataset, as the approach is completely automated in comparison with FDJ which is governed by human intervention for setting few parameters of blocking, thus it can be concluded that OATF works equally good and suitable for real time environments where no human intervention is expected.

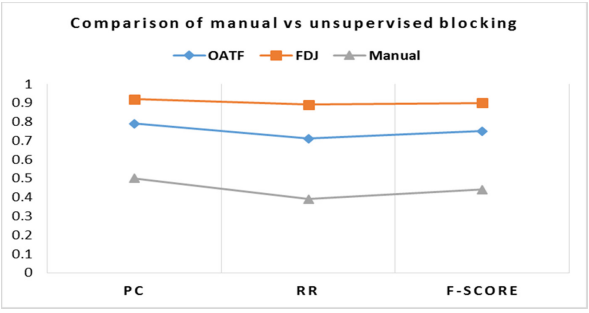

Fig. 5. Comparison of manual and unsupervised, automated blocking for restaurant

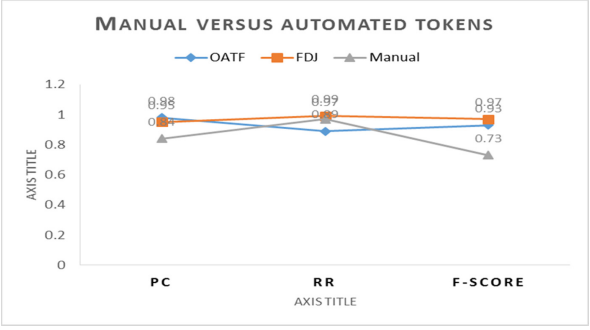

Fig. 6. Comparison of manual and unsupervised automated blocking for cora

6 Conclusions

This paper reviews the entire progress of research in the field of record linkage as well as record de-duplication. The process of identification of duplicates started with the initial concern of identification of correct duplicate and non-duplicate groups from the dataset. The research is further focused on optimizing the task of similarity comparison using blocking and indexing methods. The Comparative analysis of various blocking and indexing methods guided the most suitable and preferred method of blocking, i.e. SNM. The popular blocking method, SNM is further improvised for the scalability, temporal nature of data, web-based data and real time data. The unsupervised and fully automated, real time record de-duplication is explored till date.

The paper presented all pathways of progress in Record linkage and de-duplication systematically and provided an experimental evaluation of many methods wherever needed.

References

1. Li, L.: Data quality and data cleaning in database applications, vol. 242 (2012)
2. Lorenzi, L.: Missing Data Problems in Record Linkage-How to Find Links and Non-. Links, vol. 2001, pp. 1–3 (2013)
3. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J. Am. Stat. Assoc. **64**(328), 1183–1210 (1969)
4. William, E.: Winkler: Matching and Record Linkage (1983)
5. Elmagarmid, K., Member, S.: Duplicate Record Detection : a Survey (shorter version). IEEE Trans. Knowl. Data Eng. **19**(1), 1–16 (2007)
6. Goiser, K., Christen, P.: Towards automated record linkage. In: Conference on Research and Practice in Information Technology Series, vol. 61, pp. 23–31 (2006)
7. Thomas, W.E.W., Herzog, N., Scheuren, F.J.: Applications of record Linkage Techniques. https://www.soa.org/library/newsletters/the-actuary-magazine/2007/february/link, February 2007
8. Muse, A.G., Mikl, J., Smith, P.F.: Evaluating the quality of anonymous record linkage using deterministic procedures with the New York state aids registry and a hospital discharge file. Stat. Med. **14**, 499–509 (1995)
9. Blakely, T., Salmond, C.: Probabilistic record linkage and a method to calculate the positive predictive value. Int. J. Epidemiol. **31**, 1246–1252 (2002)
10. Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P.: Automatic linkage of vital records. Science PUBMED **130**(1959), 954–959 (1959)
11. Kelley, R.P.: Blocking considerations for record linkage under conditions of uncertainty. In: Proceedings of Social Statistics Section, pp. 602–605 (1984)
12. Christen, P.: A survey of indexing techniques for scalable record linkage and de-duplication. EEE Trans. Knowl. Data Eng. **24**(9), 1–20 (2011)
13. Ravikanth, M., Vasumathi, D.: Record matching over query results from multiple web databases with duplicate detection. J. Adv. Res. Dyn. Control Syst. **10**(4), 2040–2049 (2018)
14. Dey, D., Mookerjee, V.S., Liu, D.: Efficient techniques for online record linkage. IEEE Trans. Knowl. Data Eng. **23**(3), 373–387 (2011)
15. Papenbrock, T., Heise, A., Naumann, F.: Progressive duplicate detection. IEEE Trans. Knowl. Data Eng. **27**(5), 1316–1329 (2015)
16. Kolb, L., Thor, A., Rahm, E.: Parallel Sorted Neighborhood Blocking with MapReduce (2010)

17. Kim, J., Shim, K.: General chairs, preface. Lecture Notes Computer Science (including Subser. Lecture Notes Artificial Intelligence Lecture Notes Bioinformatics), LNAI, vol. 10234, pp. 561–573 (2017)
18. Karapiperis, D.: Summarization algorithms for record linkage. In: Edbt, pp. 73–84 (2018)
19. Ma, Y., Tran, T.: TYPiMatch. In: Proceedings of Sixth ACM International Conference on Web search data Min. - WSDM 2013, p. 325 (2013)
20. De Carvalho, G., Laender, A.H.F., Andre, M., Silva, A.S.: A genetic programming approach to record deduplication. IEEE Trans. Knowl. Data Eng. **24**(3), 399–412 (2012)
21. Jurek, A., Hong, J., Chi, Y., Liu, W.: A novel ensemble learning approach to unsupervised record linkage. Inf. Syst. **71**, 40–54 (2017)
22. Vogel, T., Naumann, F.: Automatic blocking key selection for duplicate detection based on unigram combinations. In: International Work Quality (2012)
23. Kejriwal, M., Miranker, D.P.: An unsupervised algorithm for learning blocking schemes. In: Proceedings - IEEE International Conference on Data Mining, ICDM, pp. 340–349 (2013)
24. Ramadan, B.: Indexing Techniques for Real-Time Entity Resolution, vol. March (2016)
25. Wangikar, V.C., Deshmukh, S.N., Bhirud, S.G.: An Efficient approach for automated token formation for record de-duplication with special reference to real-time data-warehouse environment. Int. J. Eng. Adv. Technol. **4**, 151–159 (2019)
26. Wangikar, V.C., Deshmukh, S.N., Bhirud, S.G.: Rough set based approach for automated token formation in real-time record. J. Adv. Res. Dyn. Control Syst. **11**, 380–390 (2019)
27. Hernández, M.A., Stolfo, S.J., Hernandez, M.A.: Real-world data is dirty: data cleansing and the merge/purge problem. Data Min. Knowl. Disc. **2**(1), 9–37 (1998)
28. Yan, S., Lee, D., Kan, M.-Y., Giles, L.C.: Adaptive sorted neighbourhood methods for efficient record linkage. In: Proceedings of 2007 Conference Digital Library - JCDL 2007, p. 185 (2007)
29. Draisbach, U., Naumann, F., Szott, S., Wonneberg, O.: Adaptive windows for duplicate detection. In: Proceedings - International Conference on Data Engineering, pp. 1073–1083 (2012)
30. Mccallum, A., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 169–178 (2000)
31. Navarro, G.: Indexing text with approximate q –grams. J. Disc. Algorithms **3**(2–4), 157–175 (2005)
32. De Vries, T., Ke, H., Chawla, S., Christen, P.: Robust record linkage blocking using suffix arrays and bloom filters, ACM Trans. Knowl. Disc. Data **5**(2), 1–27 (2011)