ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse361942020.pdf
https://doi.org/10.30534/ijatcse/2020/361942020

Redundant Data Normalization using the Novel Data Mining Algorithms

Palash Chaudhari
BE (E&TC), Sinhgad College of Engineering, Vadgaon, Pune.
Savithribai Phule Pune University,
palashchaudhari11@gmail.com

ABSTRACT

Redundant detection is a process of identifying set of words which relates to corresponding real world object in database documents. A database implies a group constituent parts linked data which is controlled & recovered efficiently. The thought of the more extensive database is high system database which resolution include of a great number of records can be collected within the database that know as great quantity regards data may collected under database scheme. In unusual statements, similar data warehousing, information digging or information integration data necessity be refined as a pre-processing tread to assure the variety of data & the performance of uses. An essential task in information purification is redundant data normalization. Existing LCS (Least Cost Methods) refer to various data paradigms & record types. Trustworthy problems along those inquiries are however to be defeated. This writing introduces a (GFS) Greedy Forward Search based approach to redundant record detection to find duplicate words in large dataset. The outcomes show that the system has a significant improvement in design prediction accuracy moreover robustness.

Key words: Record normalization, data analytics system, data integration, entity matching, pre-processing

1. INTRODUCTION

Information preprocessing includes the alteration of the crude dataset into a justifiable configuration. Preprocessing information is a major stage in information digging to develop information productivity. The information preprocessing techniques legitimately influence the results of any diagnostic algorithm; be that as it may, the strategies for preprocessing can change regards territory of utilization. Information preprocessing has been noteworthy stage under data tapping method. Correspond to a statement through Aberdeen Gathering, information readiness alludes to any activity designed to enhance the feature, ease of use, convenience, or transportability of information. A definitive goal of information planning is to permit diagnostic frameworks along spotless & consumable information to be changed into significant insights. Information preprocessing hold onto various practices, for example, purification, union, alteration & modification. The preprocessing stage may expend a generous measure of time however the result is a final informational collection, which is foreseen right & helpful for additional information mining algorithms. The crude information accessible on information stockroom, information shops, database documents (Jiawei, Micheline & Jian, 2012) was generally not composed regards examination like might remain inadequate, incompatible concerning that might be appropriated into a different table or spoke to in an alternate arrangement, to put it plainly, it is messy. The way toward creating data from the large establshed data origins is described like Knowledge Discovery in Databases (KDD) either Information Mining (Malley, Ramazzotti & Wu, 2016; Gupta & Gurpreet, 2009). It has been time regards large information & each area of topic remains to create information at an extraordinary level. Very critical assignment has been to increase correct data through current information sources. Assignment regards restructuring information has been identified like report readiness. It was utilized to find foreseen knowledge. It was consisting as understanding domain based issue viable & afterwards an assortment regards focused information to accomplish foreseen objectives (Gülser, Inci & Murat, 2011). Forrester appraises still 80% of information investigator time has been devoured under creating information (Goetz, 2015). Chose information would after able to be preprocessed has been data mining. Information preprocessing has been good answer for enhancing information quality. Data preprocessing involves cleansing regards information, normalization regards information, transformation, highlight removal & determination, & so on. The prepared information remains the practice set to the machine learning algorithm.

Steps of Data Pre–Processing

Cleaning of Data: Initial phase regards information pre-processing is information purification which perceives halfway, inaccurate, loose or unsuitable pieces of the information of datasets (Tamraparni & Theodore, 2003).

Information purification may reduce typographical mistakes. It might overlook tuple includes absent qualities or adjust values contrasted along an identified rundown of substances. The information at that point gets predictable along other informational indexes accessible in the framework.

Handling missing qualities is troublesome as inappropriately dealt along the missing qualities may prompt helpless knowledge separated (Hai & Shouhong, 2009). Different types of information cleansing arrangements using approved informational index upper filthy information under request for cleaning it. A few apparatuses relate information improvement strategies which perform unfinished report index total through expansion of associated data. Binning strategies may utilize to expel uproarious information. Clustering procedure has been utilized to distinguish anomalies (Jiawei, et al., 2012). Information may likewise denote grade escape through applying it within a relapse work. Various relapse methodology, for example, linear, different or calculated relapse are utilized to control relapse work.

Data Integration: Information Integration is the technique for merging information got from various wellsprings of information under predictable dataset. Information on web has been increasing under scale & multifaceted nature & has been or unstructured either semi structures. Collaborated regards information has been incredibly bulky & iterative procedure. Contemplations at the time of integration procedure was for the most part identified along gauges of heterogeneous information sources. Also, the way toward collaborating recent information resources for current dataset has been tedious, at last outcomes in the inappropriate utilization of important information. ELT (Concentrate Change Burden) devices are utilized to deal along bigger volume regards information; it integrates various resources under single physical area, gives uniform calculated patterns & gives querying capacities.

Data Transformation: Crude information is normally changed into an organization appropriate for investigation. Information may standardize regards instant transformation regards numerical variable for typical range. Information normalization may accomplish utilizing scale normalization strategy either Z score technique. Clear cut information can likewise be changed using collection which consolidates at least two traits under single property. Speculation may utilize on low level credits that was changed to more elevated level.

Reduction of Data: Multi faceted investigation of enormous information resources can expend significant time either also infeasible. At point when the quantity of indicator factors or the quantity of instances turns out to be huge, mining algorithms experience the ill effects of dimensionality handling issues (Jiawei, et al., 2012). Final phase of information pre-processing has been data decrease. Information reduction makes input information increasingly successful in portrayal except loosening its integrity. Information reduction could possibly be lossless. The end database may contain all the information of the original database is efficient organization (Bellatreche & Chakravarthy, 2017). Encoding procedures, chain of importance dissemination information shape total can be utilized to lessen the size of the dataset. Information reduction fits include choice procedure. Instance determination (Vijayarani, Ilamathi & Nithya, 2015) & Instance age are two methodologies utilized through information mining algorithm to decrease information size.

2.LITERATURE SURVEY

In [13], the author proposed a plan described proxy re-encryption which presents large security moreover efficiency of subsistence. The word secured encryption algorithm signifies the elimination of redundant data in storage. Although this system breaks to implement protection.

In [14], the author stated, Duplicate Detection also Fragment Placement (DDFP) a deduplication operation that reduces redundant data including fragments placing that allocates individual occurrences of an information record on storage connections. For duplicated data, source pointer was practised moreover unique information is collected on the storage joint. This improves the section of duplicate records discovery. A particle position algorithm implies practised for placing particles on separate warehouse joints. To choose nodes T-coloring obtained practised, Set T obtained applied, which contains the nodes that are at length T from one added.

In [15] the proposed scheme was completed on Hadoop which retail more comprehensive database. It consists of the discovery of duplicate data based on various qualities. In our practice, both applied data preprocessing to an information mining system that includes of converted row data in an acceptable composition. It employed the Parallel Progressive Sorted Neighbourhood Method & Map Reduce algorithm on this information to make a complete database. It produces a larger adaptable season & practical approach of data.

In [16], the Levenshtein distance algorithm utilizes a Bayesian Network to decide the likelihood of two XML objects being copies. The Bayesian Network model was made out of the structure of the articles being thought about, along these lines probabilities of all items are figured considering the information the items contain, yet in addition how such information was organized. Levenshtein distance algorithm did deftly, XMLDup requires little client intercession client just needs to give the dataset & a comparability edge.

In [17], the author analyzed the issue of record normalization over a lot of coordinating records that allude to a similar genuine element. They introduced three degrees of normalization granularities (record-level, field-level & worth segment level) & two forms of normalization. For each type of normalization, they proposed a computational system that incorporates both single-technique & multi-procedure approaches including four single-methodology draws near: recurrence, length, centroid, & feature-based to choose the normalized record or the normalized field esteem. The thorough system for registering the normalized record incorporates a set-up of record normalization methods, from credulous ones, which utilize just the information assembled from records themselves to complex methodologies, which all around mine a gathering of copy records before choosing an incentive for a trait of a normalized record.

In [18], the author executed on improving query execution in information investigation frameworks that straightforwardly store & query JSON information. Through examining a

genuine hint of creation remaining task at hand, they distinguish worldly & spatial connections among various questions regarding JSONPath get to. Such relationships would cause excess parsing of JSON information to accomplish similar information esteems. They proposed Maxson, a JSON Way reserving framework that was executed as a segment in SparkSQL. Maxson directs every day forecasts of JSON way access through client inquiries & performs pre-parsing & pre-reserving when the framework asset was under-used, normally during mid-night.

In [19], the creator introduced a novel way to deal along the issue of ordering occasion reports from a disseminated occasion recognition framework through abusing confinement procedures. The proposed classifier was surviving information collection procedures & was powerful at diminishing parcel load inside a network proposed framework a novel methodology that uses restriction strategies to rapidly distinguish & dispose of copies among occasion reports. Utilizing only the areas of hubs announcing occasions, the proposed classifier can settle on a choice on which occasion reports are to be disposed of as copies.

In [20], the creator introduced the way toward distinguishing copy incorporates three modules, for example, selector, preprocessor, & copy identifier which utilize XML records & competitor definition as info & produces copy objects as yield. This exploration planned to build up a productive algorithm for recognizing copy in complex XML archives & to diminish the quantity of bogus positives through utilizing the MD5 algorithm.

In [21], the creator actualized three distinctive normalization methods. While applying information mining to this present reality, gaining from information that fall inside an enormous explicit range is an evitable circumstance. Attempting to standardize information was an undeniable arrangement where information were scaled to fall inside a little explicit range. Strategies that were utilized to standardize information must not present commotion.

In [22], the author proposed air conditioning (N-K implies). N-K implies bunching algorithm applied normalization before grouping on the accessible information just as this methodology computes beginning centroids dependent on loads. A proficient algorithm where they have first preprocessed our dataset dependent on normalization procedure & afterward created successful bunches. This was finished through allocating loads to each ascribe an incentive to accomplish normalization.

In [23], the author proposed a disseminated way to deal along versatile normalization for Large information stream. Utilizing sliding windows of fixed size, it gives a straightforward system to adjust the measurements for normalizing changing information in every window. Executed on Apache Tempest, a disseminated continuous stream information structure, our methodology abuses conveyed information preparing for proficient normalization.

In [24], the creator actualized to standardize the mutual information utilized in the method along the goal that the mastery of the significance or the excess could be wiped out. They get some ordinarily utilized acknowledgment models including Support Vector Machine (SVM), k-Closest Neighbor (kNN), & Linear Discriminant Analysis (LDA) to contrast the algorithm & the first (mRMR) and an as of late improved adaptation of the mRMR, the Normalized Mutual Information Feature Selection (NMIFS) algorithm.

In [25], creator the proposed methodology was to be used as a pre-handling method that transforms the fine granular time-scaled dataset (that has visit spans) into a likelihood dataset in a period, subsequently better arrangement model preparing. The TBSS pre-handling method had adequately tackled the issue of repeatability & commotion that exist in the sensor information. TBSS has easily joined along irregularity identification & customary arrangement Sensors algorithm.

In [26], the creator talked about the neighborhood filtering algorithm to proficiently channel nearby copies, & afterward extend it to worldwide copies filtering. To adjust to various extra correspondences overhead in worldwide copies filtering, they introduced energetic & apathetic methodologies for Blossom channel sharing. In nearby filtering, the algorithm can simply channel neighborhood copies as equivalent to DTFilter approach & worldwide filtering algorithms, which presented distinctive extra correspondence cost.

In [27], the creator proposed a novel methodology for record linkage & combination in a web based setting. This depended on iterative reserving: a lot of regularly mentioned records (acquired from the distinctive Web databases through inspecting) is cleaned offline & stored for future references. Recently showing up records in light of a query are cleaned together along the records in the store, introduced to clients & fittingly affixed to the reserve. It introduced a general system for the web based setting dependent on an iterative record-based storing strategy. A lot of every now & again mentioned records is reduplicated off-line & stored for future reference. Recently showing up records in light of a query are reduplicated mutually along the records in the reserve, introduced to the client & annexed to the store.

In [28], the creator introduced a way to deal along explore & confirm the no-information misfortune property of semistructured information normalization. This model encodes the confirmation measures in the SemanticWeb Rule Language (SWRL) & utilizes its cosmology thinking motor to offered robotized help for the checking procedure. In these methodology researches the information safeguarding part of semistructured information normalization, & furthermore gives a versatile & robotized arrangement towards.

In [29], a complex saving band-significance metric was advanced to quantify the bendwise centrality. This guarantees the maintenance of groups including plentiful inborn structures helpful for order. In particular, planned for acquiring the introduced band-significance metric, a fulfillment algorithm was introduced, which for the most part depends on the implanting learning & linear relapse, trailed through the presentation of multi-normalization blend. Additionally, concerning the enormous repetition brought about through profoundly connected groups, MPWR further builds up an obliged band-weight streamlining model. At that point, both bendwise complex saving ability & intraband connection are completely coordinated into the band selection process. To talk about the issue, a comparing algorithm inside

the structure of the alternating direction method of multipliers (ADMM) was additionally evolved.

In [30], the creator investigates the utilization of alter distance measures to build an authoritative portrayal that is "focal" as in it is generally like every one of the divergent records. This methodology lessens the effect of loud records on the authoritative portrayal. Moreover, in light of the fact that the client may lean toward various styles of canonicalization, they show how extraordinary alter distance expenses can bring about various forms of canonicalization. For instance, lessening the expense of character erasures can bring about portrayals that favor condensed forms over extended forms (for example KDD versus Meeting on Information Disclosure & Information Mining). They portray how to take in these expenses from a modest quantity of physically commented on information utilizing stochastic slope climbing. Additionally, they acquainted feature-based methods along pick up positioning inclinations over canonicalizations. These methodologies can consolidate self-assertive printed proof to choose an authoritative record.

In [31] author built up an ITS to encourage normalization process. The client can utilize the database normalization ideas, specifically, competitor key, second ordinary structure & third typical structure. The framework chooses any issue haphazardly from the database & presents it to the client. Simultaneously, it creates its answer for check the rightness of the client's answer. The framework additionally gave proper indications to the client & an exchange instrument to show the client to discover conclusion & standardize the diagram to the second & third typical structure.

In [32], the author introduced a solo, online methodology, UDD, for identifying copies over the query consequences of different Web databases. Two classifiers, WCSS & SVM, were utilized agreeably in the combination venture of record coordinating to recognize the copy sets from all potential copy matches iteratively. There are some other works we have analyzed related to the IoT domain where we can apply the data reduction strategies [33-35]. In [36], the novel data mining techniques introduced for data clustering after data normalization. In [37], another data mining-based approach introduced for data pre-processing with objective of language features estimation. In [38], similar study recent introduced for medical data mining.

3.METHODOLOGY

Assume the operator produces a design practice collection $C = \{(R^1, l_1) \dots . (R^n, n)\}$ wherever every fixed of redundant statements R^n= {R1 . . . Rk} is explained along a description li ∈ {1 . . . k}, showing which of the redundant should be preferred as the canonical transcription (i.e., Rli ∈ R is the actual canonical work). We need to appropriate C to determine the contents of S. There has been a moderate significance of work on methods to automatically acquire to change distance values, for the most part applied to record de-duplication. Nonetheless, we do not conscious of any transaction that gets to modify distance costs for canonicalization. We recommend two straightforward despite viable approaches to learning change range costs from practice data: least-cost search & Greedy Forward Search advance. Figure 1 and 2 shows the proposed two models for data normalization using Least Cost Search (LCS) and Greedy Forward Search (GFS) respectively.

A) Least Cost Search

The most straightforward technique continues to exhaustively enumerate frames of each value & maximize unique canonicalization appearance on the training collection. Make L(a, c) be the end function for an appointment to s. Toward model, L may obtain the balance of reports in c for which Cd(R) returns a non-canonical record; i.e. We want to optimize c as regards:

$$s * = argmin\ s\ L(s, C). \qquad (1)$$

Because we requirements discretize the command frames to send an exhaustive examination, the information is the corresponding: min: The smallest cost value, max: The maximum cost value & step: The number to confound each value to achieve a new frame Search returns through cycling within the specific context of s & turning the best-found frame s∗. The specifications are provided in Algorithm 1. The scheme NextCosts creates the subsequent cost frame as defined through the step contents.

B) Greedy Forward Search

Processing L(s, C) requires figuring Cd(R) for all Ri ∈ C. This computational cost constrains the quantity of settings we can list utilizing comprehensive hunt. Rather, we propose a basic stochastic slope climbing algorithm to advance Condition 1. Given an underlying setting for c, the algorithm proposes an alteration to c & acknowledges the change if L(s, C) diminishes. This can be comprehended as recreated strengthening except the glow boundary. The subtleties of this method are given in Algorithm 2. The method SampleCostElement tests an expense consistently from the cost vector. The method RandomUpdate consistently picks between augmenting or decrementing c through step.

Algorithm 1: Least Cost Search Algorithm

1. Input: Training Dataset D
   Initial cost $s = \{s_i, s_d, s_r, s_r \neq\}$
   Min - The minimum cost value
   Max - this highest value assessment
   Step - These measures to cross every value to get a distinct series**.**
2. for I < Number_of_iteration
3. s ← Sample Cost Element (s)
4. s ← Radom_Update(s, Step, Max, Min)
5. s ← Next_Costs (s, Max, Min, Step)
6. if (s, D) < best_loss then
7. best_loss ← (s, D)
8. s*← s
9. Finish if
10. i = i+1
11. Finish for

sd= edit distance between two records.
sr= the replacement cost for swapping character
si= is the insertion cost, sd=deletion cost

Algorithm 2: Advanced scheme to normalize form the Greedy Forward Search:

Input: F- Complete quantity from features
D- Complete quantity from data samples
K- Amount of features to be decided
$C_{i,j}$ Feature cost,
To which i=1, 2... F along along j = 1, 2,…, D
A_j= Group label concerning that reports samples for which j 1,2,..D
x - Record picked measure
Output: Ck - The selected feature index where k=1, 2… D
Forward:
$C = \emptyset$
// Feature Normalization
for f==1 to F do
μ_f = Mean cost of C_f
σ= C_f stand Standard derivation
C_f=C_f-μ_f
C_f=C_f/σ_f
// Transform characteristics into discrete from practising direct quantization
$\dot{C}$= approximate (C)
//Begin picking features
for k=1 up to K do
for i=1 up to C do
Measure f^a ($\dot{C}_i$)
c=argMax$i \in c$ (f^a ($\dot{C}_i$))
$C = C \cup c$
End

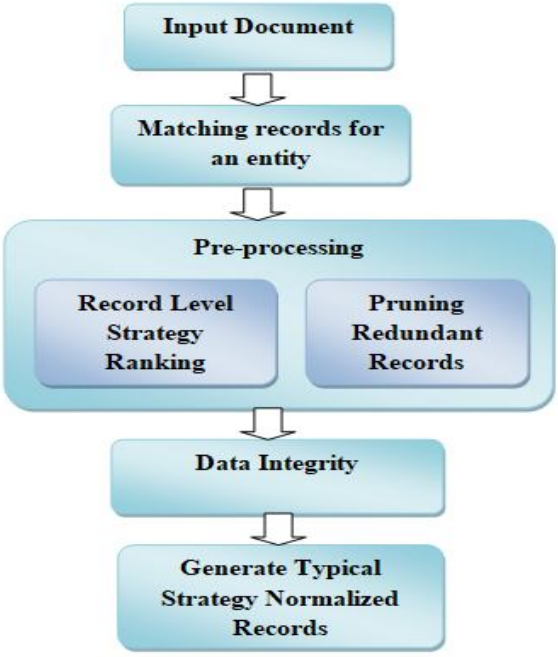

Figure 1: Redundant Data Normalization

Prior to broadly expounding of our technique, we initially think about the a higher bound of the common data of irregular factors. Because any nonstop factor container obtains approximately into the discrete structure, we expect that pair discrete irregular factors X including Y are presented alongside their minor & joint circulations. Henceforth, the common data of X & Y is processed.

To reduce the issue of inconsistent normalizing loads, we recommend utilizing the element free upper attached to standardize the mutual data. In this way, our standardized component mutual data is determined through

$$NI(X,Y) = \frac{I(X,Y)}{log_{\dot{C}}S} \qquad (2)$$

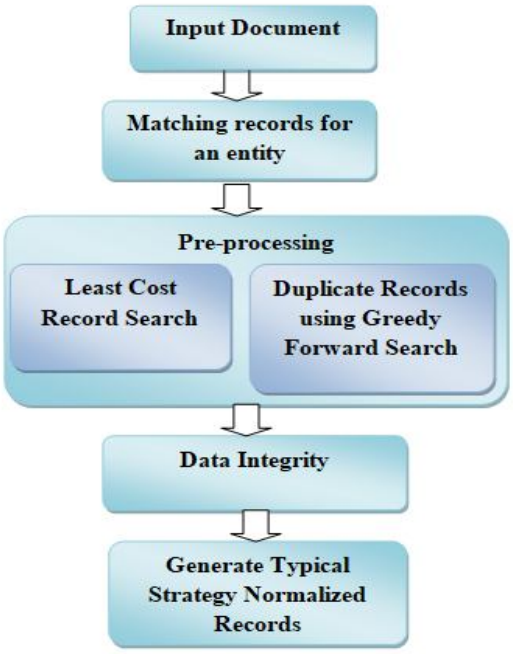

Figure 2: Proposed Redundant Data Normalization

The normalized include highlight mutual data is consistently inside the range [0, 1]. Along these lines, to accomplish a harmony between the pertinence & the excess, we separate the class-include mutual data through log2 |ΩC|. The normalized class-feature common records is immediately established as

$$N\ I(S,X) = \frac{I(C,X)}{log_2(|\Omega_S|)} \qquad (3)$$

Doing the normalized mutual data roles described, we estimate the possibility of a characteristic as

$$f^1(Xi) = NI(S,Xi) - \frac{1}{|Ci-1|}\sum_{Xc \in Ci-1} NI(Xc,Xi) \qquad (4)$$

To explain our development, we associate along the other two computations in the duration of the analysis accuracy, we express them as f^2 & f^3respectively. Furthermore, to verify

the outcome of the disproportion among the importance including the repetition that we look explanation preceding, we join normalized form peculiarity common data including the corresponding form peculiarity common data. While this process, each morality of a feature is contained through

$$f^4(Xi) = NI(S, Xi) - \frac{1}{|Ci-1|}\sum_{Xc \in Ci-1} \frac{I(Xc,Xi)}{\min(H(Xc),H(Xi))} \quad (5)$$

Moreover, measures to analyze our system along other traditional methods such as FBR [23], WBorda [23], UDD [25] & PEBL [25] are shown in the appendix division to withdraw a combination of the practice. This following in program design in Algorithm 2 shows the collection method employing a greedy forward-searching approach.

4.RESULTS AND DISCUSSION

In the following figure shows the accuracy precision & recall for GFS compared along four different normalization approaches, FBR, PEBL, UDD & Wborda method, on the appropriate the dataset PVCD [26]. The dataset contains information about the distribution scene canonicalization [27]. PVCD has 3,683 distribution scene esteems for 100 unmistakable certifiable distribution records. It is just worried about the field setting, which is apparently the most troublesome field to standardize, in view of the nearness of abbreviations, shortened forms, & incorrect spellings. We utilize this dataset to contrast our methodologies along alludes along a typical PVCD dependent on 1, 00 copy, & 1, 00 no copy vectors, which are haphazardly chosen from the beginning. While both LCS & GFS utilize two standardization strategies, the two strategies in UDD on the other hand coordinate within the cycles, while those powerless classifiers inside PEBL simply operate ere that emphasis. Therefore, GFS beats LCS since in GFS either technique can distinguish occurrences that can't be recognized through different strategies, which coordinated information utilizing expel excess records. It can likewise be seen that LCS is slower than GFS in light of the fact that GFS needs two emphases to recognize the copies. Notwithstanding, GFS is quicker than LCS & UDD strategy, which require a larger number of emphases than LCS to recognize all repetitive standardize procedures, which require preparing information. In the database situation, where records to coordinate are enormously inquiry needy, a pertained approach isn't relevant as the arrangement of records in each question's outcomes is a one-sided subset of the full informational collection. Figure 3 shows the performance of accuracy of various data normalization techniques. Similarly, the results are estimated in terms of precision and recall rates in figures 4 and 5 respectively.

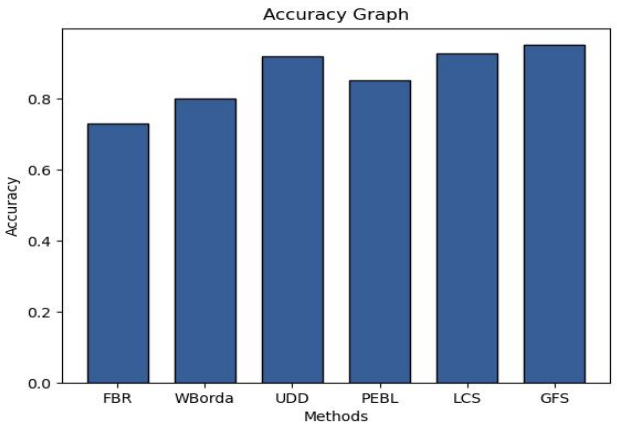

Figure 3: Measurement of the accuracy of the recommended method as associates various methods

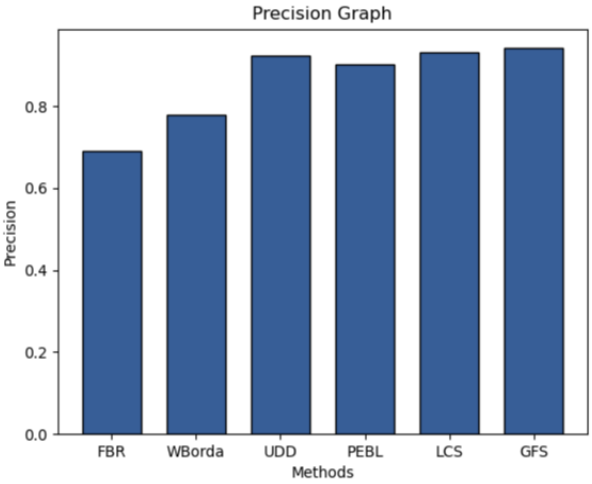

Figure 4: Measurement of the precision of a recommended method as associates various methods

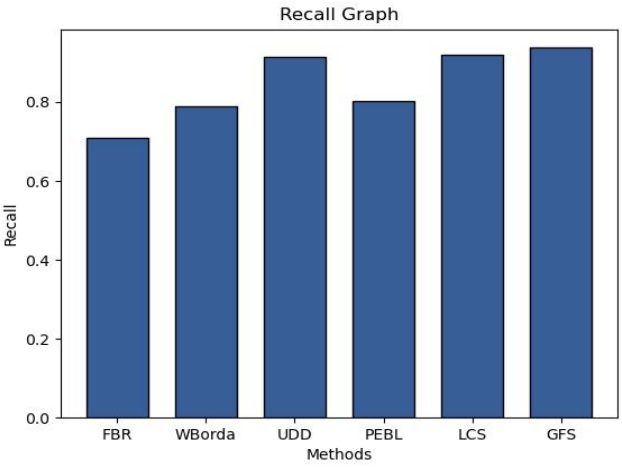

Figure 5: Measurement of the recall of the recommended method as associates various methods

To conquer this issue, we introduced a solo, proposed approach, GFS, for identifying copies over the question brings about a dataset. Two standardization strategies, LCS & GFS, are utilized agreeably in the intermingling venture of record coordinating to distinguish the copy sets from all potential copy matches iteratively. Exploratory outcomes show that our methodology is similar to past work that requires preparing models for recognizing redundant form the question brings about databases of different techniques. The work in [26] is a case of ordinary standardization since it chooses one of the copy records or one of the fields esteems as the standardized record or field esteem, separately. It doesn't endeavor to make new handle esteems or new records as normalized records.

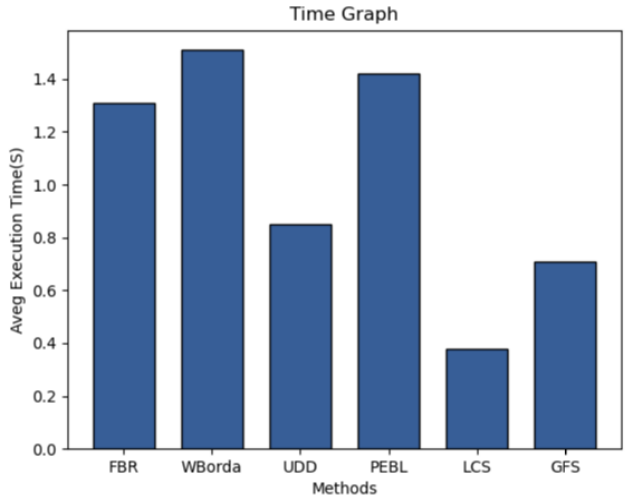

Figure 6: Calculating the execution time of the recommended scheme

Figure 6 showing the outcome of execution time required to normalize the input data. All results proves the proposed algorithms LCS and GFS delivered the improved data normalization performance as compared to existing methods The implementation of the proposed system is better accuracy, precision, recall & execution time that show in figure 6. Our Proposed method LCS & GFS are shown higher performance as compared to other approaches.

5.CONCLUSION

In the contemporary method, the corresponding work records in the database, resolution improve the extension of the database. The redundant item performs not intimate that there are two final models of identical items. Preferably, redundant records are data that is insignificantly separate but which connected to the same character as remarkable additional data. A, consequently, a greater quantity of thought will be needed to collect the data & the complexity of the resolution of the part repairs. To obtain a document of the database remains a challenging job. In the proposed scheme, we applied two algorithms that are LCS & GFS that will eliminate repetitive of the history. Also, we utilized the pre-processing procedure that inclination remodels it into a suitable arrangement. The mechanism is performed to produce consistent data. The operative provisions increased manageable space furthermore effective treatment of data. Additionally, it will enhance the availability of data including offering more durable access preceding the database. An indispensable work in data purge is redundant information normalization. The outcomes exhibit that the system has significant development in representation prediction accuracy also the robustness

REFERENCES

[1]. Alasadi, S. & Bhaya, W. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering & Applied Sciences, 12(16), pp. 4102–4102.

[2]. Andrew, K. (2015). The research of text preprocessing effect on text documents classification efficiency. International Conference Stability & Control Processes IEEE, St. Petersburg, Russia.

[3]. Bellatreche, L. & Chakravarthy, S. (2017). Big Data Analytics & Knowledge Discovery. Proceeding of 19th International Conference DAWak Lyon France. Expectation maximization algorithm, Wikipedia, Retrieved February 10, 2019

https://doi.org/10.1007/978-3-319-64283-3

[4]. Goetz, M. (2015). Three ways data preparation tools help you get ahead of Big Data.

[5]. B. & Murat, C. (2011). A review of data mining applications for quality improvement in the manufacturing industry. Expert System along an application, 38(10), pp. 13448–13467.

[6]. Gupta, V. & Gurpreet, S. (2009). A Survey of Text Mining Techniques & Applications. Journal of Emerging Technologies in Web Intelligence, 1(1), pp. 60–76.

[7]. Hai, W. & Shouhong, W. (2009). Mining incomplete survey data through classification. Knowledge & Information Systems Springer, 24(2), pp. 221–233.

[8]. Jiawei, H., Micheline, K. & Jian, P. (2012). Data Mining Concepts & Techniques. (3rd ed.), USA: Morgan Kaufmann. 221 Edición Especial Special Issue Mayo 2019

[9]. Malley, B., Ramazzotti, D. & Wu, J. (2016). Data Preprocessing; Secondary Analysis of Electronic Health Records. Springer.

https://doi.org/10.1007/978-3-319-43742-2_12

[10]. Tamraparni, D. & Theodore, J. (2003). Exploratory data mining & data cleaning. New York, USA.

[11]. John Wiley & Sons. Vijayarani, S., Ilamathi, M., & Nithya, M. (2015). Preprocessing Techniques for Text Mining – An Overview. International Journal of Computer Science & Communication Networks, 5(1), pp. 7–16.

[12]. Xindong, W., Xingquan, Z., Gong–Qing, W. & Ding, W. (2014). Data Mining along Big Data. IEEE

transactions on knowledge & data engineering, 26(1), pp. 97–107.

[13]. Maragatharajan, M., & Prequiet, L. (2017). Removal of duplicate data from encrypted cloud storage. 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization & Signal Processing (INCOS).

[14]. Patil, J., & Barve, S. S. (2017). DDFP: Duplicate detection & fragment placement in the deduplication system for security & storage space. 2017 1st International Conference on Intelligent Systems & Information Management (ICISIM).

[15]. Bhoi, B., Vyawahare, P., Avhad, P., & Patil, N. (2017). Data duplication avoidance in a larger database. 2017 International Conference on Innovations in Information, Embedded & Communication Systems.
https://doi.org/10.1109/ICIIECS.2017.8276031

[16]. Gaikwad, S., & Bogiri, N. (2015). Levenshtein distance algorithm for efficient & effective XML duplicate detection. 2015 International Conference on Computer, Communication & Control (IC4).

[17]. Dong, Y., Dragut, E. C., & Meng, W. (2018). Normalization of Duplicate Records from Multiple Sources. IEEE Transactions on Knowledge & Data Engineering, 1–1.

[18]. Shi, X., Zhang, Y., Huang, H., Hu, Z., Jin, H., Shen, H., … Zhou, K. (2020). Maxson: Reduce Duplicate Parsing Overhead on Raw Data. 2020 IEEE 36th International Conference on Data Engineering (ICDE).

[19]. J. Pfender & W. K. G. Seah, "Leveraging Localisation Techniques for In-Network Duplicate Event Data Detection & Filtering," 2017 IEEE 42nd Conference on Local Computer Networks (LCN), Singapore, 2017, pp. 163-166.

[20]. Lwin, T., & Nyunt, T. T. S. (2010). An Efficient Duplicate Detection System for XML Documents. 2010 Second International Conference on Computer Engineering & Applications.

[21]. Shalabi, L. A., Shaaban, Z., & Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine. Journal of Computer Science, 2(9), 735–739.
https://doi.org/10.3844/jcssp.2006.735.739

[22]. "Normalization based K means Clustering Algorithm " Deepali Virmani, Shweta Taneja, Geetika Malhotra in International Journal of Advanced Engineering Research & Science (IJAERS)-Feb 2015

[23]. Vinh, L. T., Lee, S., Park, Y.-T., & d' Auriol, B. J. (2011). A novel feature selection method based on normalized mutual information. Applied Intelligence, 37(1), 100–120.

[24]. Wang, X., Zhang, Q., & Jia, Y. (2008). Efficiently Filtering Duplicates over Distributed Data Streams. 2008 International Conference on Computer Science & Software Engineering.

[25]. Su, W., Wang, J., & Lochovsky, F. H. (2010). Record Matching over Query Results from Multiple Web Databases. IEEE Transactions on Knowledge & Data Engineering, 22(4), 578–589.

[26]. A. Culotta, M. Wick, R. Hall, M. Marzilli, & A. McCallum, "Canonicalization of database records using adaptive similarity measures," in SIGKDD, 2007, pp. 201–209.

[27]. Rezig, E. K., Dragut, E. C., Ouzzani, M., & Elmagarmid, A. K. (2015). Query-time record linkage & fusion over Web databases. 2015 IEEE 31st International Conference on Data Engineering.
https://doi.org/10.1109/ICDE.2015.7113271

[28]. Li, Y. F., Sun, J., Dobbie, G., Lee, S., & Wang, H. H. (2009). Verifying Semistructured Data Normalization Using SWRL. 2009 Third IEEE International Symposium on Theoretical Aspects of Software Engineering.

[29]. Sui, C., Li, C., Feng, J., & Mei, X. (2019). Unsupervised Manifold-Preserving & Weakly Redundant Band Selection Method for Hyperspectral Imagery. IEEE Transactions on Geoscience & Remote Sensing, 1–15.

[30]. Culotta, A., Wick, M., Hall, R., Marzilli, M., & McCallum, A. (2007). Canonicalization of database records using adaptive similarity measures. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '07.

[31]. Mendjoge, N., Joshi, A. R., & Narvekar, M. (2016). Intelligent tutoring system for Database Normalization. 2016 International Conference on Computing Communication Control & Automation (ICCUBEA).

[32]. Su, W., Wang, J., & Lochovsky, F. H. (2010). Record Matching over Query Results from Multiple Web Databases. IEEE Transactions on Knowledge & Data Engineering, 22(4), 578–589.

[33]. Mahajan, H. B., & Badarla, A. (2018). Application of Internet of Things for Smart Precision Farming: Solutions and Challenges. *International Journal of Advanced Science and Technology*, Vol. Dec. 2018, PP. 37-45.

[34]. Mahajan, H. B., & Badarla, A. (2019). Experimental Analysis of Recent Clustering Algorithms for Wireless Sensor Network: Application of IoT based Smart Precision Farming. *Jour of Adv Research in Dynamical & Control Systems*, Vol. 11, No. 9.
10.5373/JARDCS/V11I9/20193162.

[35]. Mahajan, H. B., & Badarla, A. (2020). Detecting HTTP Vulnerabilities in IoT-based Precision Farming Connected with Cloud Environment using Artificial Intelligence. *International Journal of*

Advanced Science and Technology, Vol. 29, No. 3, pp. 214 - 226.

[36]. Mohamed Shenif. (2020). Understanding User's Behavior by Social Media Data Clustering. International Journal of Advanced Trends in Computer Science and Engineering, Vol.9, No.1, https://doi.org/10.30534/ijatcse/2020/25912020.

[37]. Noorli Khamis, & Nurul Farahin Musa. (2020). Corpus-based Data for Determining Specialised Language Features. International Journal of Advanced Trends in Computer Science and Engineering, Vol.9, No.1 https://doi.org/10.30534/ijatcse/2020/07912020.

[38]. C. Narmatha, Dr. M. Thangamani, & S. Jafar Ali Ibrahim. (2020). Research Scenario of Medical Data Mining Using Fuzzy and Graph theory. International Journal of Advanced Trends in Computer Science and Engineering, Vol.9, No.1 https://doi.org/10.30534/ijatcse/2020/52912020.