Juni Khyat
(UGC Care Group I Listed Journal)
ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

A Review on Data Quality Enhancement using Data Cleaning Algorithm

Amit Kumar Jha
Research Scholar, Poornima University,
Jaipur, India
Email: amitjha@poornima.org

Dr. Madan Lal Saini
Associate Professor, Poornima University, Jaipur, India
Email: madan.saini@poornima.org

Abstract:
In this paper, we have reviewed the different algorithms used to enhance the data quality. We have studied and analyzed different data cleaning approaches andit's found that if quality of data is not improved before the data analysis, then result will not come as expected. So, it is concluded in our analysis that data quality is a critical factor in any database application. Since data is integrated from heterogeneous data source the redundancy in data is significantly increased. in this paper we have shown the differences in between different data cleansing algorithms. In our study it is found the algorithms which are simple behave fairly well in order to deal with duplicate data.

Keywords: Data cleaning; Duplicate removal; Redundant Data

1. INTRODUCTION

In order to remove the duplicate, irrelevant or inaccurate data from different heterogeneous data sources, the data cleaning or data cleansing methods are used. The inconsistencies and errors in data is caused during data entry, or during data transmission or by other type of modification. In other words data cleaning is a process of removing typographical error or correcting the wrong values against the known values. Some data cleaning process use cross checking method. The process of data cleaning involves activities like harmonization of data or making standard measure of data i.e. changing the data set to a new predefined protocol. The problem of duplicate data removal has been considered a critical aspect of data cleaning. When data from distributed data sources is integrated, the same data in different data sources leads to non-redundant data. This requires the identification and removal of duplicate data that leads to redundant data.

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

2. DATA CLEANING TECHNIQUES

Many cutting-edge commercial tools help the extraction, transformation and loading (ETL) of (likely unclean) facts right into a trustworthy (cleansed) database. ProbClean efficaciously helps relational queries, and lets in new forms of queries against a hard and fast of possible upkeep and treats reproduction detection strategies as data processing tasks with uncertain outcomes. Based on the prevailing replica records, identification algorithm SNM and MPN, turned into proposed as an progressed algorithm which analyses attributes and sort the dataset multiple instances to make reproduction statistics greater clustered. The algorithm also gives a unique weight to every characteristic and introduces the idea of effective weight so that to make the evaluation greater accurate. A filtering mechanism changed into delivered to enhance the efficiency of detection. Following are different strategies for records cleaning.[21]

A. Parsing

Using parsing techniques data cleansing is performed for the finding the syntax errors. There is probability of lexical and domain errors in the data. Use of simple set of values is done to find the structure of domain. Beside this, discrepancy detector is used for checking and detecting the anomaly in the data. [21]

B. Data Transformation

Using mapping the data transformation is performed which changes the given format into format expected by means of application. The facts from different resources are mapped into a general uniform schema that is the needs of the intended application. The process of Standardization along with normalization is renovation on the instance level applied with the goal of doing away with deviations in information. This includes simple value conversion or translating features in addition to normalizing numeric values to lie in a hard and fast interval given by means of the highest and lowest values. [21]

C. Integrity Constraint Enforcement

Integrity constraint guarantees that integrity must be maintained after transactions like editing a date via inserting, deleting, or updating tuples. Integrity constraint checking drops transactions

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

that, if applied, could defy a few integrity constraints. Integrity constraint upkeep is worried with figuring out extra updates to be introduced to the original transaction to assure that the resulting facts series does no longer breach any integrity constraint. [21]

D. Duplicate Elimination

There are many processes used for duplicate removal or file linkage, which is a component of facts cleansing. The duplication detection technique proposed calls for an algorithm for deciding whether two or extra tuples are reproducing representations of the same entity. For efficient reproduction and detection each tuple has to be evaluated to every different tuple using this replica detection technique.[21]

3. PROPOSED DUPLICATE REMOVAL TECHNIQUE

A simple and new approach has been used to design the data cleaning system especially for duplicate detection and removal.

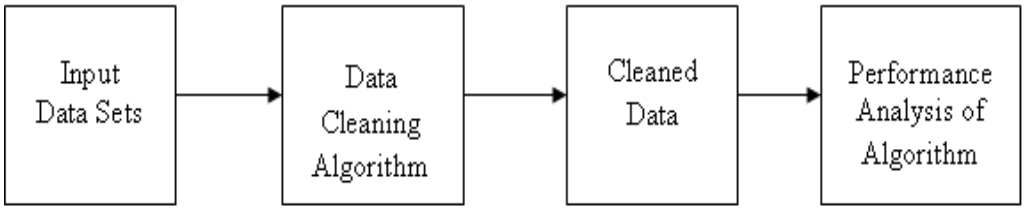

Fig1. Block Diagram of the System

4. REVIEW OF LITERATURE

The issue of managing blend of enormous information sources utilizing Resource Description structure and ontologies was managed by Salima Benbernou. Author proposed a quantifiable methodology which guarantees great quality information for comprehending inquiries over huge RDF information in an appropriated manner on a Spark environment. Following advances are utilized in building the RDF information approach for cleaning conflicting enormous data.(1)modeling consistency rules to recognize the irregularity significantly increases in spite of the fact that it is certainly covered up including deduction and conflicting guidelines (2) through rule assessment which depends on Apache Spark structure irregularities are distinguished and insignificantly sub-set of conflicting triples are found. (3) Inconsistencies are cleared by finding

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

the best fix for reliable question replying. Creators presented an examination on conflicting reasoning in large RDF information base. Information significantly increases are the main spot where information irregularities occur. At first, irregularity is identified by the methodology and conflicting subset of RDF significantly increases is found by utilizing intelligent guidelines. The irregularity distinguishing proof was called attention to on Spark structure with the goal that huge information of RDF is taken care of. The following stage is to clean the irregularity by using the subset of irregularities. The best pair is looked among the fix sets displayed in a hyper diagram.[1]

Joeri Rammelaere proposed data cleaning methods for extra information of the data; for example, constraints provided by the user where the dirty data exists. Restrictions of domains or illegal combination of values are also important examples of the above mentioned concept. Authors presented a type of repairing which is different and which prevents introduction of novel constraint violation as per the discovery algorithm. This algorithm repaired ensures that all errors identified through discovered constraints on the dirty data are fixed. The constraint discovering process would not identify violation of constraints. Authors did this for novel constraints which are called forbidden item set (FBI's) that captures unlikely co-occurrences of values. They showed that high precision data is detected by FBI. Similarities are computed in a flexible algorithm. In the context of future work various likeliness functions are intended to be experimented for the forbidden item set. In fact, any likeliness function could be used for any single object constraints. As far as the impact of a fixed number of changes on this function could be bounded, that means the applicability of approach for larger classes of constraints. Further research is warranted by the repair algorithm. Could a better reparability be achieved on high dimensional data is the question here. Conclusively, varied patterns of constraints and patterns are open for revisiting due to viewing the data quality dynamically.[2]

Morteza Alipour Langouri elevated the prerequisite to remember setting for information cleaning keeping in see the emotional idea of information quality. Late work which depends on Functional Dependencies and remembering ontologies for to it contended that ontologies are an incredible wellspring of setting. These are likewise a viable device for displaying area ideas and connections for information cleaning. Utilizing datasets which are genuine, they represent how ontologies can make information cleaning work processes better. Issues and bearings are

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

likewise sketched out for future work. Contrast with existing conditions and data cleaning work processes significant data cleaning models wealthier associations inside the data and keeps up a vital good ways from misclassifying consistent data as off-base. Creators noticed that ontologies were by all account not the only method for determining setting and different systems for including setting in information cleaning ought to likewise be explored. Moreover, relevant conditions are not prone to delete the requirement for human consideration in the information cleaning process. They could diminish the weight of human affirmation. [3]

Virender Kumar focused on pollution data set and tried to explain the importance of data cleaning. The prime focus of the paper was on two things: One was to analyses and visualizes data on Air pollution which is unstructured. Secondly, was to comprehensively survey methods to clean data which is unrequired, dirty and doubtful. The scenario of industrialization that creates an effective impact on the economy of a nation but also need to take care of our environment as the money is of no value if we do not have a fresh air to breath. The visualization of data presents the rise in hazardous gases which create harmful impact on the health of human life. It also assists us to identify the locations that are conspicuously affected by the harmful gases and also the areas where these gases are raising flambiountly. Numerous data cleaning techniques and tools can be utilized to prepare a structured framework for this. This results in more exact, accurate and trustworthy dataset for arriving at valid conclusion and fulfilling this paper's scope the scope of this paper. [4]

Otmane Azeroual proposed few novel techniques that help organizations to upgrade the quality of their information pertaining to research. In the present paper the issue of those data quality problems which can occur in research information systems is addressed. And various ways to fix and upgrade them with new techniques or methods is also mentioned. The improvement of data quality is always targeted. The present concept can be used as a basis for the using facilities. It offers an appropriate procedure and that enables user to evaluate the data quality in a better way in Research information systems. This also enables users to prioritize problems in a better way and prevent their recurrence in future. Furthermore, these data errors must be rectified and updated with data cleansing. Data cleansing tools are basically commercial. These are available for both small application contexts as well as large data integration suites. Currently a market for data cleansing is developing as a service. As far as future work is conserved, interviews of

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

experts or quantitative surveys are determined in the universities and many research institutions in order to find out how high the data quality is in their research information system and discuss what methods and measures are required to improve and upgrade quality of data are used to research information..[5]

YinglongDiao proposed an set of rules for huge statistics on-line cleaning. This is based totally on dynamic outlier detection for correct cleaning of the large-scale, blended and erroneous collective information. This reduces the value of information cache and also ensures the regular deviation detection on timing of each records cycle. The records cleansing technique is progressed by neighborhood outliner detection on density, sampling cluster uniformly dilution Euclidean distance matrix retaining some rectification into next cycle of cleansing, which avoids a sampling causing overall cleansing deviation and reduces quantity of calculation within information cleansing solid time, enhancing the rate greatly. The disbursed solutions upon on line cleansing set of rules are based on Hardtop platform. This algorithm has expanded the real-time information pre-processing performance. Especially the performance for far flung sensor records source for instance electricity control, surroundings monitoring, machinery manufacturing that realizes the massive data supply on line cleansing has increased. Simultaneously, the algorithm can also examine on-line real-time facts. Data processing and the cleaned historical information also keep a few memory (such as distribution, density distribution and bizarre deviations), excluding massive range error. Data cleansing deviation resulted from the sensor system malfunction or natural environmental factors communique interference.[6]

Nan Tang proposed achievable answers for address information quality issue. In nutshell, Data cleaning has assumed a crucial job throughout the entire existence of the executives and investigation of information. Taking care of top notch information has been demonstrated to be significant for organizations for basic leadership. Particularly in the ongoing data driven age and the period of enormous information Asset Description Framework (RDF) is a standard model for information trade on the semantic web. Notwithstanding, it is realized that RDF information is messy, since a large number of them are naturally extricated from the web. In this paper, we will initially return to information quality issues showed up in RDF information. Albeit numerous

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

endeavors have been put to clean RDF information, sadly, the greater part of them depend on relentless manual assessment. They likewise depict potential arrangements that shed lights on (semi-)naturally cleaning (enormous) RDF information. [7]

Samir Al-janabi introduced an equivalent system and calculations to coordinate information duplication with conflicting information fixing and finding of the precise qualities in information. Information cleaning is a critical piece of the information change organize in information warehousing where in the removed information from social databases are typically unclean. This may influence basic errands in various associations, for example, information investigation and basic leadership. Current procedures of information cleaning by and large manage a couple of value viewpoints. The procedures guess the accessibility of ace information and that clients are associated with information cleaning, for example, physically setting certainty scores which speak to the rightness of the estimations of information. Creators utilized the implanted thickness data in information to fix mistakes dependent on information thickness where significantly increases that are near one another are pressed as one. They exhibited a weight model to dole out certainty scores which depend on the thickness of information. The assignments are mechanized and no client is engaged with the procedure. They thought about the conflicting information as far as infringement as for a lot of utilitarian conditions (FDs) on the grounds that these infringements are regular by and by. They introduced a cost model for information fixing which depended on the weight model. In future work, creators are attempting to extend the calculations to deal with different sorts of respectability limitations including contingent utilitarian conditions [8].

Liu Hong has established an information cleaning system for a large amount of information, which will improve the quality of the information. Information cleaning is a method to repair or evacuate corrupted or incorrect information. For big data, this process is fundamental and crucial because the underlying information can lead to inadequate investigations and inevitably produce unacceptable results. In this article, they have: I) generate auxiliary metadata and expressive metadata for informational things and datasets; ii) track the use of information; iii) use these two information settings and use examples to discover from numerous sources Comparable

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

informational things and data sets, and iv) construct an information dependent report that can benefit subsequent information cleaning processes and enhance information data in our proposed information cleaning system. The test results show that the calculation is feasible for improving the quality of information. The discovery in principle was that the proposed method could choose to identify informational things and related data sets that could benefit from the subsequent information cleaning process. As a future work, they intend to further establish their proposed information cleaning structure by establishing a positioning framework, which may generate demand based on data and estimates of cleaning costs. In addition, they will build information-based models and intelligent methods for information cleaning by combining AI calculations and artificial prayers.[9]

A calculation method is proposed by S Swapan to make a decision for each attribute. The pronoun in the leaf center will replace the missing quality so that they can get more quality information and can use the relevant information mining program for quality inspection. Information is quickly generated through the Web, so choosing the right option has become a daunting task. They are surrounded by information, but they long for information. Getting more information from business information is a key task for those in charge. Information warehouses provide workable answers for processing information. For proper inspections and basic leadership, information should be informative and accurate. The information collected from different sources may contain dirty information. Before stacking information in a distribution center, it should be cleaned to ensure quality information. Once the information is cleared, it will produce accurate results when applying information mining problems. Therefore, stable information is essential for basic leadership. Identify and eliminate loud qualities for future use. [10]

R. Krishnamoorthy proposed a new method called effective data removal (EDC) technology, which can show the relevant and unrelated instances in a large data set by the degree of missing values, and reconstruct them by the closest instance within the instance Missing values in related instances. Instance set. The EDC system combines two technologies: identifying related instances (IRI) and reconstructing missing values (RMV). IRI technology distinguishes between applicable cases and redundant cases by the missing estimation level of each case in the example set. This case has a place in huge occasions. The RMV strategy can reproduce the missing

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

incentives in major events through the closest case depending on the separation degree. The experimental results show that the proposed EDC strategy is direct and powerful in distinguishing applicable occasions from unimportant occasions, and can reproduce the missing quality in important examples through the closest case. The future work of this method is improvement. In addition, it includes testing a large set of constant examples and comparing the results with the current information cleaning strategy [11]

J. Hossen proposed an AI calculation that predicts missing qualities and changes existing systems for enormous information examination by improving significant strides in preprocessing (that is, information cleaning). They tried two characterization calculations: SVM and Random Forest. At that point, two classifier calculations (i.e., irregular woodland classifier and straight SVM classifier) are utilized to prepare the dataset to arrange the information quality and build up a smart model. The outcomes show that the two classifiers are substantial, contingent upon the informational collection. The trial results show that the precision of irregular timberland and direct relapse consistently stays around 90%. They need to utilize this strategy to give a cleaned set of information for further preparing. Moreover, investigators can profit by the framework during the information examination period of the cleanup stage and infer that the information has been tidied up. For future work, extra information purifying guidelines will be included. On the off chance that the qualities in the informational collection aren't right, honesty requirements (IC, (for example, useful conditions (FD)) can be coordinated with AI [19] to order the kind of mistakes to be gotten. [12]

Sanjay Krishnan proposed a structure called AlphaClean, which reconsidered the parameter adjustment of the information cleaning pipeline. AlphaClean provides clients with a rich library to clarify information quality metrics through a weighted whole of SQL total queries. AlphaClean applies a post-production search system where each assembly line cleaning administrator contributes changes from competitors to a common pool. Non-parallel, in discrete strings, pursuit computations will inherit it in clean pipelines to magnify the quality metric of customer representation. This design allows AlphaClean to apply various improvements, including stable evaluation of quality metrics and learning dynamic pruning rules to reduce

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

query space. Experiments on real benchmarks and manufacturing benchmarks show that AlphaClean can find a caliber arrangement that is 9 times higher than innocently applying the most advanced parameter adjustment technology. In essence, it is useful for intricate information cleaning strategies and the redundant information has a solid foundation and can be used as a cleaning administrator in a class-like cleaning framework, such as HoloClean. They are eager to extend AlphaClean to an adaptive, visual and intelligent cleaning process. They intend to coordinate AlphaClean with the information representation framework so that customers can control the perception of information outwards, turning it into quality capabilities. [13]

Shen Xiaojun proposed an exception evacuation strategy and procedure dependent on the change point gathering calculation and the quartile calculation, and afterward hypothetically broke down its practicality. The contextual investigation and its examination with the quartile-change point gathering calculation and the nearby exception factor calculation show that the proposed change point gathering quartiles calculation can adequately recognize four sorts of anomalies and tidy up The impact is great, the productivity is high, and the adaptability is solid. As indicated by the spatial dispersion area and shape, the exceptions of the breeze vitality bend of the breeze turbine are separated into four sorts: anomalies stacked at the base, center, and upper, and spread exceptions around the bend. The exploratory outcomes show that the proposed gathering of quartiles of progress focuses can viably recognize the stacked anomalies and dissipated exceptions of the breeze control bend, and has great cleaning impact, high proficiency and solid adaptability. This article utilizes a cancellation technique to manage unusual information, which certainly affects the uprightness of the information. In future work, it is important to consider further improving information quality through information redress and information interjection, and create information amendment dependent on explicit information utilization.[14]

Salman Salloum proposed the RSP-Explore strategy to allow information researchers to iteratively view a large amount of information on a small graph stack. They tend to three main areas: measurable estimates, mislocalization and information cleanup. Because of the large number of irregular example information squares (called RSP obstacles) that are going to use the entire information, a random sample partition (RSP) circular information model is used to speak

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

to the information. Choose a square-level example of RSP squares to understand the information, distinguish between potentially significant values errors, and get clean information. They proposed a hypothesis test on the use of RSP obstacles for fact estimation and demonstrated the benefits of RSP-Explore technology exactly. The trial results of three real information databases show that the speculative results from RSP-Explore can be quickly added to the real quality. In addition, cleaning an RSP square example is sufficient to evaluate the factual properties of fuzzy cleaning information. Future work will use the RSP method for histogram estimation, information presentation and highlighting. [15]

Gaudence Uwamahoro proposes a calculation method based on word position, which can reduce the search range of competitors and expand their proficiency and survival against incomplete report correlations ability. Evidence that distinguishes duplicates or near-copy archives in a large number of records is one of the serious problems in data recovery. In this experiment, the results showed that during the scanning problem, the up and down sizes have been reduced by up to 12% of the size of the record set, which reduces the lookup time. The results also show higher accuracy, and thus help customers avoid sitting tightly in the chair and grabbing problems and getting bad records. With this technique, they proved that the word position plays a vital role in the importance of records, and that is the basis for archival competitors' selection. In the future, they plan to explore stress techniques in this strategy to improve query capabilities. [16]

Lavanya Pamulaparty came up with another idea, which was to find pages close to copying from a huge archive. Web Ming faces huge problems due to the duplication and near duplication of web pages. In a large amount of information such as the "network", identifying near-duplicates is very cumbersome. The proximity of these pages reduces the performance of the presentation while merging information from heterogeneous sources. These pages either add room to the list or increase service costs. Recognize that these pages have many potential applications, such as possibly proving literary theft or copyright infringement. This article involves identifying and revoking copies and methodological archives used to perform report summaries. Investigation sequel suggests that this calculation is related to similarity measures. Recognizable evidence of duplication and duplication records are close, leading to reduced memory in the warehouse. With

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

the premise of discovery, future work will be studied to find increasingly powerful and accurate strategies to approach replication identification and disposal. In order to expand the accuracy and sufficiency of the DD algorithm, a grouping tool can also be used for feasible report grouping [17]

Nancy Jasmine Goldena proposes a general-purpose computing technology that relies on fluffy bunches and dull cycles to jointly investigate the closest copy of the content report, but has yet to find a way to achieve it Opportunities are incorporated into separate plans and attempts are made to overcome the shortcomings of each calculation in this way. The fluffy beam calculation analyzes the similarity of different data in the report, which contains two pictures and content that can be accessed online through an application appointment. In the first stage, the fluffy grouping calculation checks the copied content of the picture by destroying the RGB broken by the Euclidean metric, and then in the second stage, the parametric view is used to check the content mining for comparable language structures according to the test schedule. . Then, each file is tested according to the content time through the parameter view, the similarity of content mining is checked, and the results are compared with different existing technologies. Related results show that compared with other strategies, the proposed calculation method can decompose replica records in less time and reduce misclassification. In the future, this fluffy grouping calculation will be implemented and the content based image decomposition will be used to find the compactness [18]

CihanVarol proposes a hybrid approach that implants Jaro separation and the fact that words use recurring facts to repair poorly characterized information. In a real book data set, the proposed "half and half" method generally improves the accuracy of shingles calculations by 27%, and can be done on more than 90% of basic shingles. Despite the negative effects of too much information, this duplication has some positive effects. For example, duplicate information may be helpful to you, and even critical to achieving management-level execution. In addition, copying information may be of interest to openness and accessibility. In addition, various descriptions of similar information in the information repository can help to obtain new, obscure data. Along these lines of thought, it is worthwhile to realize that if the copy data is appropriately merged, the information quality can be improved. The vast majority of current reduplication programs rely on the fully coordinated contribution of consolidation to a data set. The creator

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

created a framework that coordinates behavioral spelling based on misspelled content, and then ranks recommendations based on usage. The system naturally chooses the best accessibility for the less-characterized content and then matches the data in another report to check if the closely indistinguishable records are undoubtedly similar, but at least one contains error-prone data. The overlap calculation is suitable for mirroring the closeness of two files. By withdrawing the data that identifies misspellings from the framework to determine the applicant's word manual labor, it can be argued that due to the lack of human desire for changes, tightly indistinguishable coordination accuracy levels will be lower. Testing reflects the use of the crossover method as a powerful answer for large-scale short content copy recognition and extends the representation of strip chart calculations. As a future work, the semi-breed method will be used in different spaces (for example, in medical service information) to improve the location of close copy. In addition, other language alternatives will be added to the reuse of word information to comprehensively solve the problem of close copying [19]

MarzanaIfat Moly planned and performed the ETL process by using Microsoft SQL Server and Microsoft Visual Studio (C #). After making this ETL, you can discover the source of the information and, after distinguishing the information from its sole source, change it to the general structure of any information, and after the last change, stack it into the main data distribution center. A large amount of information is stored from various sources within the organization to guide management choices. Information warehouse is one of the basic vocabularies of the last ten to twenty years, and big data is the hot model of the previous 5 to 10 years. ETL stands for separation, change, and burden. This is another idea for data distribution centers. [20]

5. COMMON FINDING

- In order to identify the relevant and irrelevant instances, Effective Data Cleaning (EDC) techniques are applied.
- Extract transform and Load (ETL) is a technique which is used to collect the data from different heterogeneous source, thereafter uses different algorithms to transform the data into suitable format and the data is loaded into database for data analysis.
- Resource Description Framework (RDF), is used to improve the data quality

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

- The document processing efficiency by reducing its size is done with the help of Word Positional Based Approach for Document Selection (WPBADS).
- The fuzzy based cluster algorithm analyzes the duplicate document in less time with reduced misclassification.
- Use of the Decision Tree Induction (DCT) Algorithm, Handling the missing values.
- Use of Data Cleansing Techniques improved the data quality.
- In order to judge, how proposed model behave on new data set, the Random Forest and Linear support vector machine algorithms are applied.

6. STRENGTHS AND WEAKNESS

- In order to minimize the searching time in data word positional based technique for document selection (WPBADS) is applied.
- Using Duplicate Detection (DD) Algorithm, increase the efficiency of web crawling.
- Using Effective Data Cleaning (EDC) technique, reconstruct the missing with high accuracy.
- Using the fuzzy based cluster algorithm, analyze the duplicate document in less time with reduced misclassification.
- WPBADS method not applies compression methods for the query efficiency.
- The fuzzy based cluster algorithm not implemented and analyzed on text-based images to find the similarity

7. CONCLUSION

The process of data cleaning is used to remove the error in data, used to insert suitable value in the field where data is not available and used to check the semantics of data. The data is an integral part of data analysis and must be handled with care. In order to improve the data quality data interpolation and correction techniques are applied which essential before analytical process.

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

REFERENCES

[1] Benbernou, S. and Ouziri, M., 2017, December. Enhancing data quality by cleaning inconsistent big RDF data. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 74-79). IEEE.

[2] Rammelaere, J. and Geerts, F., 2019. Cleaning data with forbidden itemsets. *IEEE Transactions on Knowledge and Data Engineering*.

[3] Alipour-Langouri, M., Zheng, Z., Chiang, F., Golab, L. and Szlichta, J., 2018, April. Contextual Data Cleaning. In *2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW)* (pp. 21-24). IEEE.

[4] Kumar, V. and Khosla, C., 2018, January. Data Cleaning-A Thorough Analysis and Survey on Unstructured Data. In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 305-309). IEEE.

[5] Kumar, V. and Khosla, C., 2018, January. Data Cleaning-A Thorough Analysis and Survey on Unstructured Data. In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 305-309). IEEE.

[6] Diao, Y., Liu, K.Y., Meng, X., Ye, X. and He, K., 2015, September. A big data online cleaning algorithm based on dynamic outlier detection. In 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (pp. 230-234). IEEE.

[7] Tang, N., 2015, April. Big RDF data cleaning. In 2015 31st IEEE International Conference on Data Engineering Workshops (pp. 77-79). IEEE.

[8] Al-janabi, S. and Janicki, R., 2016, July. A density-based data cleaning approach for deduplication with data consistency and accuracy. In 2016 SAI Computing Conference (SAI) (pp. 492-501). IEEE.

[9] Liu, H., Tk, A.K., Thomas, J.P. and Hou, X., 2016, March. Cleaning framework for bigdata: An interactive approach for data cleaning. In 2016 IEEE +Second International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 174-181). IEEE.

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

[10] Swapna, S., Niranjan, P., Srinivas, B. and Swapna, R., 2016, March. Data cleaning for data quality. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 344-348). IEEE.

[11] Krishnamoorthy, R., Kumar, S.S. and Neelagund, B., 2014, May. A new approach for data cleaning process. In International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014) (pp. 1-5). IEEE.

[12] Hossen, J. and Sayeed, S., 2018, September. Modifying Cleaning Method in Big Data Analytics Process using Random Forest Classifier. In 2018 7th International Conference on Computer and Communication Engineering (ICCCE) (pp. 208-213). IEEE.

[13] Krishnan, S. and Wu, E., 2019. Alphaclean: Automatic generation of data cleaning pipelines. arXiv preprint arXiv:1904.11827.

[14] Shen, X., Fu, X. and Zhou, C., 2018. A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm. IEEE Transactions on Sustainable Energy, 10(1), pp.46-54.

[15] Salloum, S., Huang, J.Z. and He, Y., 2019. Exploring and cleaning big data with random sample data blocks. Journal of Big Data, 6(1), p.45.

[16] Uwamahoro, G. and Zuping, Z., 2013. Efficient Algorithm for Near Duplicate Documents Detection. International Journal of Computer Science Issues (IJCSI), 10(2 Part 2), p.12.

[17] Pamulaparty, L., Rao, C.G. and Rao, M.S., 2014. A near-duplicate detection algorithm to facilitate document clustering. International Journal of Data Mining & Knowledge Management Process, 4(6), p.39.

[18] Goldena, N.J. and Victor, S.P., 2014. Effective Analysis of Nearest Duplicate Web Document by using Fuzzy Clustering Method. International Journal of Advanced Research in Computer Science, 5(3).

[19] Varol, C. and Hari, S., 2015. Detecting near-duplicate text documents with a hybrid approach. Journal of Information Science, 41(4), pp.405-414.

[20] Moly, M.I., Roy, O. and Hossain, M.A., 2019. An Advanced ETL Technique for Error Free Data in Data Warehousing Environment.

[21] Bhargava, S., Hemrajani, N., Goyal, D. and Gander, S., 2011. DWH-Performance Tuning for Better Reporting. International Journal of Computer Applications, 32(1).

D.O.I : 10.46528/JK.2020.V10I06N08.07

ISSN: 2278-4632
Vol-10 Issue-6 No. 8 June 2020

[22] Mohamed, H.H., Kheng, T.L., Collin, C. and Lee, O.S., 2011, December. E-CLEAN: A data cleaning framework for patient data. In 2011 First International Conference on Informatics and Computational Intelligence (pp. 63-68). IEEE.

D.O.I : 10.46528/JK.2020.V10I06N08.07