Journal of Computer Science and Information Technology, Vol. 1, No 1, 2016, pp. 71–81

Detecting Duplicate Entry in Email Field using Alliance Rules-based Algorithm

Arif Hanafi[1*], Sulaiman Harun[2], Sofika Enggari[3], and Larissa Navia Rani[3]

[1]Facultyof Computer Systems & Software Engineering
Universiti Malaysia Pahang,26300 Kuatan,Pahang, Malaysia
[2]Asia Pacific University, Kuala Lumpur, Malaysia
[3]Universitas Putra Indonesia YPTK, Padang, Sumatera Barat, Indonesia

*sulaiman.harun@apu.edu.my

Received 02 February 2016; Revised 19 March 2016; Accepted 20 June 2016

Abstract

The way that email has extraordinary significance in present day business communication is certain. Consistently, a bulk of emails is sent from organizations to clients and suppliers, from representatives to their managers and starting with one colleague then onto the next. In this way there is vast of email in data warehouse. Data cleaning is an activity performed on the data sets of data warehouse to upgrade and keep up the quality and consistency of the data. This paper underlines the issues related with dirty data, detection of duplicatein email column. The paper identifies the strategy of data cleaning from adifferent point of view. It provides an algorithm to the discovery of error and duplicates entries in the data sets of existing data warehouse. The paper characterizes the alliance rules based on the concept of mathematical association rules to determine the duplicate entries in email column in data sets.

Keywords: Datacleaning, Algorithm, Alliance rule, Duplication.

1. Introduction

Email is one of the devices for communication through text. It is evaluated that a normal PC client gets 40 to 50 emails for each day. Numerous applications need take emails as inputs, for instance, email examination, email routing, email separating, email outline, data extraction from email, and newsgroup analysis [1]. Unfortunately, email data can be very noisy. Specifically, it may contain headers, signatures, quotations, and program codes. It also may contain extra line breaks, extra spaces, and special character tokens. It may have spaces and periods inaccurately removed and it may contain words badly cased or non-cased and words misspelled. In order to achieve high quality data mining, it is necessary to conduct data cleaning at the first step [2,3].

Data cleaning is the method of identifying and removing inaccurate records from a record set, table, or database [4]. It is mainly used in databases; the phrase indicates to identifying incomplete, incorrect, inaccurate, irrelevant, and etc. as parts of the data and then replacing, modifying, or deleting this dirty data or unclean data [5,6]. Data cleaning is also called data scrubbing; it is the method of changing or deleting data in data warehouse that is inaccurate, incomplete, inappropriately designed, or duplicated. Organization in a data environment field like insurance, retailing, banking, telecommunications, or transportation might use a data scrubbing tool to analytically study information error by implementing technique, algorithms, and certain data mining rules. Basically, a data cleaning tool includes a framework that were capable of correct a number of specific types of mistakes, such as missing values in database or finding duplicate records. Using a data cleaning technique appropriately will definitely save a database administrator a significant amount of time and can be less costly than fixing errors manually. Data cleaning technique task practice to load in missing values, unified date format, converting nominal to numeric, identify outliers and smooth out noisy data and also correcting the inconsistent data. Data information quality difficulties are often present in particular data groups, such as files and databases, example due to misspellings during data entry, missing information or other invalid data. When various data sources need to be integrated, example in data warehouses, associate database systems or global web-based information systems, the need for data cleaning increases significantly [7]. This scenario happens because the sources repeatedly consist of duplicate data in variety of demonstrations.

In order of providing access or admission to accurate and reliable data, combination of various information demonstrations and elimination of duplicate data become essential. Data warehouses require and provide extensive support for data cleaning. They load and continuously refresh huge amounts of data from a variety of sources so the probability that some of the sources contain"dirty data" is high. Furthermore, data warehouses are used for decision making, so that the correctness of their data is vital to avoid wrong conclusions. For instance, duplicated or missing information will produce incorrect or misleading statistics ("garbage in, garbage out"). Due to the wide range of possible data inconsistencies and the complete data volume, data cleaning is considered to be one of the biggest problems in data warehousing [7,8,9,10].

Many data warehouses have been equipped with email data cleaning features. However, the number of noise types that can be processed is restricted. No previous study has so far sufficiently examined the problem in the research community, to the best of our knowledge. Data cleaning work has been done mainly on structured tabular data, not unstructured text data. In natural language processing, sentence boundary detection, case restoration, spelling error correction, and word normalization have been studied, but usually as separated issues. The methodologies proposed in the previous work can be used in email data cleaning. However, they are not sufficient for removing all the noises.

This paper presents detecting duplicate entry in email field using alliance rules-based algorithm. The paper characterizes the alliance rules based on the concept of mathematical association rules to determine the duplicate entries in email column in data sets.

The rest of this paper is organized as follow. Section 2 presents related works. Section3 presents rudimentary on alliance and HADCLEAN algorithms. Section 4 presents proposed methodology. Section 5 presents results and following by discussion. Finally, the conclusion of this work is presented in Section 6,

2. Literature Review

Data mining is widely used nowadays by organization with powerful consumer focus such as in business, banking, telecommunication and retail organization [11]. It allows these organization to decide their connection among these crucial factors such as product price, product allocating or staff skill and also a factors like economic indicator, anticipation between competition of another organization and customer demographic that is include age, sex, class, address and many more about customer details that allow an organization to determine the impact on sales marketing, customer satisfaction and corporate incomes. Therefore it allow them go deeper into information summary to view any detail on transactional data [12]. With data mining the administrator could use sale records of customer purchases to send targeted promotion based on product sales history. In data mining demographic data collected from the information, the organization could analyse their data for their companies purpose such as put on the promotion, increase the sales, therefore in order to implement all this, the quality data from data mining is crucial for the organization or government that can increase the development of the country [13].

Wide ranging information technology is arising in developing separate transaction and analytical systems; data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Usually, any of four types of relationships are required [14,15].

a. **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials [16,17].

b. **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities [18,19].
c. **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining [20,21,22,23].
d. **Sequential patterns**: Data is mined to anticipate behavior patterns and trends [24]. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Based on previous study of the alliance rules, the algorithm is implemented in the name field where it not contain any headers, signatures and quotations, it also may not contain extra line breaks, extra spaces, and special character tokens. Therefore the email field is much more complex than name field because there will be more special character inside the field. The scope of the study is to focus of email field and stress on the duplicate detection.

3. Alliance and HADCLEAN Algorithms

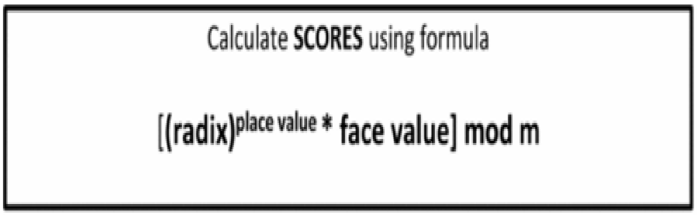
The essential of this algorithm important in the dataset of system of an association or corporate company which involve billing process. The storage of huge quantity of data about the customers suffers from the dirty data. A data warehouse if formed by integration data from different sources which can have different field formats. A data mart of a telecommunication system may involve section like generate the billing information, account section, personal info section and etc. which the integration of process can lead to the presence of dirty data [11].

3.1. Alliance Algorithm

This algorithm shows the duplicity error of string data type (name filed) and uses the algorithm of de-duplicity in the name field of the data warehouse [13]. The steps involved are:

3.1.1. Pre-processing

Here the strings in the name field are converted into a numerical value which is stored in another file called Score for reference [13]. The integer values are called scores of the name. The string is converted into numbers using relation (See Figure 1).



Calculate **SCORES** using formula

$$[(\text{radix})^{\text{place value}} * \text{face value}] \bmod m$$

Figure 1: The calculation scores

The Figure 1 above describes in the calculation scores. The paper refers the sum of number words value in a name defined as *N*.For example, name Sonal S.Porwal has *N*=3, then the total number of scores would be *N*+1.

a. Radix is 27 characters (26 alphabets and ‘.’),
b. b) Face value is the sequence of occurrence of characters in the world of alphabets starting with 0- a---25-z and 26-(.)
c. The place value is marked from right to left starting from 0.
d. M is any large prime number
e. letters are case-insensitive

3.1.2. Alliance rules application

Here the 2 data marts are considered such that a name from DM1 is to be checked and matched for duplicity with all the names in another data mart DM2. The steps involved are briefly introduced in paper [6] as alliance rules application and duplicity detection.

3.1.3. Detection of errors

The errors in the name are evaluated using the concept of q-grams testing. The q-grams are the substring of a given name string. The length of the substring can be of any value smaller than the length of the name string itself. For example, ARIF NAPI has *Q*=3, then the *Q*-grams are as follows:

(1,##A),(2,#AR),(3,RI),(4,IF_),(5,N_),(6,_NA),(7,AP),(8,PI#).

Here the initial '##' determines that the name is started, and the later determines the end. This method also considers the space between the words as well.

3.1.4. Constraint of the Alliance Algorithm

a. This approach specifically deals with the study of error types and their detection related to string data types.
b. It mostly concentrate on the "name" field and does not focus on any other field
c. Another weakness was that it could not detect the duplicity error competently in some circumstances and it needed the date of birth of that person as a reference.
d. In cases the DOB field is incorrect or blank field then the cleaning process could give the inaccurate output.
e. Loads of manual work during pre-processing phase could lead to the error prone and time consuming.

3.2. HADCLEAN Algorithm

The spelling errors and ambiguity in terms is a common data entry error found in the database system, to serve this type of dirty data the concept of dictionary is used where the spellings errors are detected using the standard dictionary. Many organizations have different terms assigned to the posts for their employees which may not match with other organizations and serve as jargons. To address this issue many organizations make use organization specific dictionary. Also some records have blank fields that can be filled using transitive closure algorithm [14].

3.2.1. HADCLEAN Steps

PNRS

The Personal Name Recognition Strategy corrects the phonetic and typo errors using standard dictionaries.it employs two strategies:

a. Near error approach: It states the faults in the words which are closely misused and displays faults.it is completed by put in a blank space e.g. preprocessing , by interchanging two letters e.g.:'rgreen' with 'green', by altering/adding/removing a letter. The activities are occupied with the support of reference to standard dictionaries.
b. Phonetic algorithm: It practices the idea of phonetic codes which is computed for every word and then it is matched with the phonetic code of the standard dictionary.it aids to notice faults for words like 'seen' and 'scene' which sound same (phonetic) but have dissimilar meaning and different phonetic codes correspondingly.
c. The modified PNRS-some administrations include the convention of nonsenses and often have their administrations description in regional languages, the situations of society specific dictionary can be used as a reference.

Transitive Closure

The records are matched using the attribute keys. Using key the records are clustered and coordinated as a set of connected archives and then faults are identified and modified after thorough analysis. This assist to fill in the blank cells (fields). Thus, remove the duplicity faults and repetition entries.The altered transitive closure creates routine of extra than one key to match and cluster the connected records. Primary secondary and tertiary key concept is used to group the related records, and when the records are matched blanks are filled and redundancies are deleted.

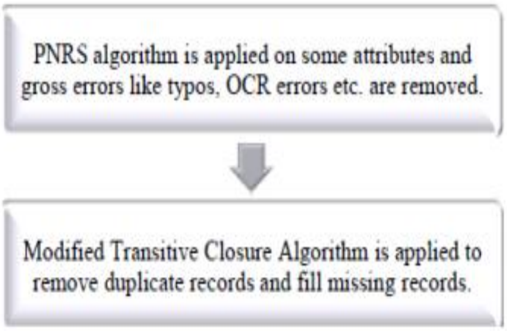

PNRS algorithm is applied on some attributes and gross errors like typos, OCR errors etc. are removed.
Modified Transitive Closure Algorithm is applied to remove duplicate records and fill missing records.

Figure 2:HADCLEAN Steps

The Figure 2above indicates, the flowchart above summarizes the working of the HADCLEAN approach. The PNRS approach is executed proceed by modified transitive closure algorithm [14]

3.2.2. Constraint of HADCLEAN Algorithm

a. The prioritization to the attribute keys [7] in modified transitive closure creates the algorithm data to be more detailed, therefore required manual involvement.

b. The modified transitive closure algorithm has a few conditions outlined in order to merge the records as connected. The guidelines being firm, occasionally the algorithm not able to merge the entries even if they are connected since it has only one secondary key and two tertiary key matches. The situation can be detected from the work completed in the [14] in the form of table which the algorithm execute before and after records

3.3. Comparison Features between Alliance Rules and HADCLEAN Algorithms

The alliance rules algorithm focuses on the error detection and only emphasizes on string data types. The HADCLEAN also highlight on the detection of outliers in the string data types but with a slight different method. Though alliance rules algorithm creates the step by step procedures of the mathematical strategy to detect errors, meanwhile HADCLEAN algorithm implements dictionary based strategy to identify spelling mistakes.The alliance rules algorithm idea generally on the 'name' field which consists of string data type. While HADCLEAN method covers the spelling mistakes, typographical errors (string data format) etc. beside with blank fields mistakes i.e. Missing data [13], too using (transitive closure) [14].Besides, alliance rules algorithm includes a tons of mathematical computations/calculations that leads to time consuming and error prone, HADCLEAN algorithm implements the dictionary based method and keys of the database that ease the executions and reduced amount of error prone and less complex.The crucial disadvantage of the HADCLEAN approach is only beneficial to English language. Besides, one great benefit of using the alliance rules is it transforms the string data types into integer numbers (Scores), thus concentrate on the memory concerns. The redundancy entry in Alliance rule algorithm is state with the assist of score matching and the anomaly is identified using Q-Grams [13]. Transitive closure convert N no of records into connected cluster and provide identification method also remove of redundancies entries. Modified transitive closure increase the quality of the result with primary, secondary and tertiary key strategy. The modified PNRS nonetheless fixes greater quantity of anomaly than PNRS (See Table 1).

Table 1: Comparison & Analysis of Algorithms

Features	Alliances Rules	HADCLEAN
Activities involved	-Pre-processing -Alliances Rules Detection and Q-Gram	-PNRS -Transitive Closure
Strategy	Scores Calculation and Comparison	Using Dictionary
Complexity	The calculation in the pre-processing phase increases the complexity of the algorithm.	Less complex because the algorithm implement dictionary approach
Accuracy	High	Low
Weakness	-lots of calculation involve. -Manual calculation leads to error prone - Time consuming	English language only Less time consuming
Types of dirty data	Misspell Entry Duplicate Entry	Nearly misspell, typo errors

4. Proposed Methodology

The theoretical study has been completed from various sources like journals, research papers, books and internet. Data cleaning techniques has been used for the cleaning of a diversity of data. On the beginning of results gained comparison of techniques has been executed to find the best techniques for data cleaning.In order to develop, enhance and evaluate the effectiveness of detecting the duplication of email entry in dataset using Alliance rules, four phases of research activities are involved namely the Planning Phase, Requirement Elicitation, Implementation and Verification of Correctness Phase and Documentation Phase in order to obtain the results (See Figure 3).

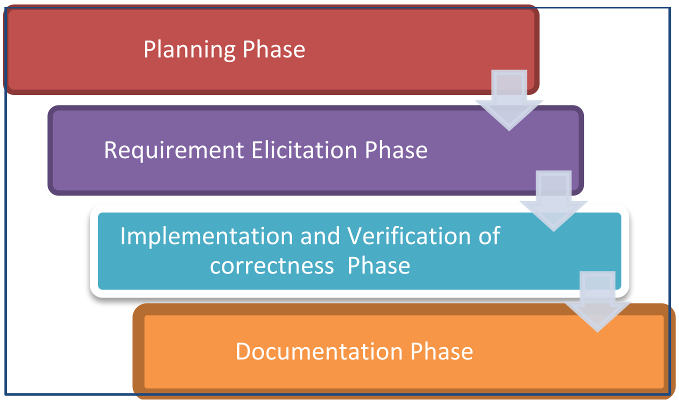

Planning Phase
Requirement Elicitation Phase
Implementation and Verification of correctness Phase
Documentation Phase

Figure 3: Proposed method

The Figure 3 above is well explained in the sub-topics below.

4.1. PHASE I: Planning Phase

In this first phase, the survey of the literature review has been conducted to establish the state-of-art on software testing. Once the background of the research has been familiarized, the aim, objectives and research scope are set to fulfil the requirements of the research.

4.2. PHASE II: Requirement Elicitation Phase

This phase is named the requirement elicitation phase. In this particular phase, the understanding gained from the literature review is put into practice in order to describe the research requirements.

In particular, the existing algorithms are studied in order to identify the improvements that have been observed.

4.3. PHASE III: Implementation and Verification of Correctness Phase

Here, the existing algorithms will be improved and at the same time, another algorithm will be developed to verify the effectiveness and correctness.

4.4. PHASE IV: Documentation Phase

Lastly, this phase involved the experiences and lessons learned as well as the elaboration concerning the relevant documentation.

5. Implementation and Comparison Results

The enhancement algorithm developed in the course of this study would be implemented using a scripting language, PHP and MySQL is to act as the database (See Table 2).

Table 2: Hardware specification for implementation

Hardware	Specification
Processor	Intel(i5)) CPU 1005M @ 1.90GHz 1.90 GHz
Memory (RAM)	4.GB the bigger the better for performance of the execution
System type	WINDOWS 8 (64-Bit operating system)

For the software specification, Brackets an open source editor is used (optional – See Table 3). This software provides a multiple platform to be used in programming environment in this case is PHP platform.

Table 3: Software specification for implementation

Software	Specification
Brackets	WINDOW 8, Internet Explorer or another web browser

The de-duplicity algorithm for strings or specifically email field consists of following three main stages:

- Pre-processing
- Condition rules application
- Detection of errors

In pre-processing the strings in the email field are converted into a numerical value which is stored in a file for ready reference. Then using the alliance rules and minimum confidence the duplicity is detected and reported in the detection of error phase.

5.1. Pre Processing

Data sets that contain email are measured to checked and matched for duplicity with the emails insert in the searching field. The strings in the email field are involved during alliance rules implementation. So, data sets of an organization and apply alliance rules to detect the duplicity in the entries.The data sets taken consist of email fields which are converted to numerical integer values. The converted integer values are stored in a file called. The conversion is done by first evaluating the total number of words in an email say, Najmi3$@gmail.com. The string is converted into numbers using relation [(radix) place value face value] mod m where radix is defined as a set of 48 characters (48 alphabets + special symbol.). The letters are taken as case-insensitive. The face value of each character is marked by the sequence number with which they arrive in alphabetic order starting with 0 -a — 25-z and 26 - (.) and m is any large prime number.

Letter: 0-26 (a-z)
Number: 27-35 (0-9)
Special Character = 36-48 (# - _ ~! $ ‘( ) * + , ; :)

The place value of letters is marked from right to left starting with O.All these evaluated scores for each email corresponding for each word in the email is stored in tabular form in the *Score* file.

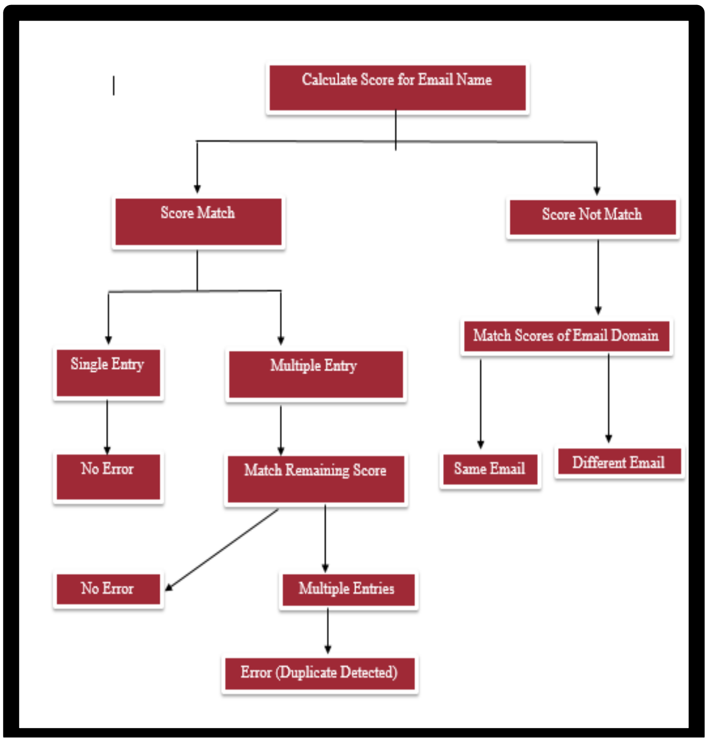
5.2. Proposed Algorithm Rules Step

The algorithm for detecting duplicity in email filled of datasets is as follows:

a. Take an email from D1.
b. Determine the no. of words in the email. Let it be denoted by *N*.
c. Email namenajmi3$@gmail.com.
d. Calculate the scores for the email each corresponding to a word present in email. The email name score is score of email name not include the domain name.
e. Calculate email name scores for all names in the datasets insert in searching field.
f. Match the domain scores *Sn* of name in D1.
g. Now match email name and domain name insert by the user in interface with the email name and domain name in the email datasets D1.

Najmi3$
Score = [13(49) ^6 + 0(49) ^5 + 9(49) ^4 + 12(49) ^3 + 8(49) ^2 + 42(49) ^1 + 29(49) ^0]
(179936733613 + 0 + 51883209 + 1411788 + 19208 + 2058 +29) mod 733
179990449905 mod 733 = 702

5.3. Condition Rules



5.4. Result and Discussion

The result is stated that the email name for *Arif_Napie#*** produce a score 686. Accordingly, the result for matching is produce to detect any duplication if there have similar email name in the database. The original Alliance Rules algorithm was only focus on name field. It was unable to identify the special character redundancies. It is required to manually calculation in every data sets record to find out the duplication. But in the modified version, the special characters and numbers can also generate score for matching process (See Table 4).

Table 4: Comparison with Alliances Rules HADCLEAN algorithms

Features	**Alliances Rules**	**HADCLEAN**	**Proposed Algorithm**
Step Involve	-Pre-processing -Alliance Rules	-PNRS -Transitive Closure	-Pre-Processing -Condition Rules
Strategy	Scores Calculation and Comparison	Using dictionary	Score Calculation
Complexity	High	Low	Moderate
Accuracy	High	Low	High
Weakness	Manual Calculation Time Consuming	-English Language only	Focus on Email field

6. Conclusion

Poor quality data costs businesses vast amounts of money every year. Defective data leads poor business decisions, and inferior customer relationship management. Data are the core business asset that needs to be managed if an organization is to generate a return from it. Based on the study in data cleaning, it is essential to have a clean data in an organization in order to generate and process the right output in whether in business profits or non-profits purpose. Hence, the output generate by the organization data storage determine the reliability of the method data cleaning itself, if the data cleaning method using by the company/organization performed well in removing or detecting the outliers therefore the output will be accurate and consistent.The study of proposed algorithm is basically based on Alliance Rules algorithm where it delivered a remarkable solution for data duplicate detection in datasets. The data duplicate detection using proposed algorithm in email field has been amazingly operational and able to function and execute smoothly. Besides, the developed system also reduce the time consume in calculation of data scores it is show that the swift progress of the propose algorithm compare to the manual calculation, that can lead to calculations error.The proposed algorithm system limitation in this study is where itonly focusses on email field data type. Another constraint is the system concentrate only on duplicate detection thus only the partial process in data cleaning is done where it should involves data correctness as well. The system data duplicate detection only calculates scores from small datasets that consists of 10-100 entries. The future of the works is to continue the implementation of propose algorithm in another data type or field column. Concentrate more on misspelled detection and correction also the q-gram matching implementation. Besides, the system also able to process the large datasets 1000-5000 entries and adding more features to the system and improve the GUI for user purpose. The future work comprised a study on taking a decision of substituting lots of calculations with minor calculation.

References

1. Whittaker, S. and Sidner, C., 1996, April. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 276-283). ACM.

2. Rose, A.N.M., Awang, M.I., Hassan, H., Zakaria, A.H., Herawan, T. and Deris, M.M., 2011, August. Hybrid reduction in soft set decision making. In *International Conference on Intelligent Computing* (pp. 108-115). Springer Berlin Heidelberg.

3. Herawan, T., Rose, A.N.M. and Deris, M.M., 2009. Soft set theoretic approach for dimensionality reduction. In *Database Theory and Application* (pp. 171-178). Springer Berlin Heidelberg.
4. Herawan, T., Ghazali, R. and Deris, M.M., 2010. Soft set theoretic approach for dimensionality reduction. *International Journal of Database Theory and Application*, *3*(2), pp.4-60.
5. Ma, X., Qin, H., Sulaiman, N., Herawan, T. and Abawajy, J.H., 2014. The parameter reduction of the interval-valued fuzzy soft sets and its related algorithms. *IEEE Transactions on Fuzzy Systems*, *22*(1), pp.57-71.
6. Ma, X., Sulaiman, N., Qin, H., Herawan, T. and Zain, J.M., 2011. A new efficient normal parameter reduction algorithm of soft sets. *Computers & Mathematics with Applications*, *62*(2), pp.588-598.
7. Zhang,S.,C.Zhang,andQ.Yang,Datapreparationfordata mining.*AppliedArtificialIntelligence*, 2003. 17(5-6):p. 375-381.
8. Brandt, R., & Chong, G. (2010). *Design informed: Driving innovation with evidenced-based design*. Hoboken, N.J.: John Wiley & Sons.
9. Yang, Q., Yuan, S., & Rajasekera, J. (2008). An Important Issue in Data Mining-Data Cleaning. (2002): 455-464.
10. Y.Patil, R., & Kulkarni, D. (2012). A Review of Data Cleaning Algorithms for Data Warehouse Systems. *International Journal of Computer Science and Information Technologies*, 3(5212 - 5214), 5-5.
11. Hellerstein, J. (2008). Quantitative Data Cleaning for Large Database.*United Nations Economic Commission for Europe (UNECE).*
12. Choudary, N. (2014). A Study over Problems and Approaches of Data Cleansing/Cleaning. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(2).
13. R. Arora,P. Pahwa and S. Bansal,"*Alliance Rules for Data Warehouse Cleansing*", 2009.IEEE Press, Pages 743-747
14. Arindam, P., and Varuni, Ganesan,"*HADCLEAN:A Hybrid Approach to Data Cleaning in Data Warehouses*",2012.IEEE Press,Pages 136-142.
15. Adu-Manu Sarpong, K., Davis, J., & Panford, J. (2013). A Conceptual Framework for Data Cleansing – A Novel Approach to Support the Cleansing Process International Journal of Computer Applications (0975 – 8887)., 77(12).
16. Amini, A., Saboohi, H., Herawan, T. and Wah, T.Y., 2016. MuDi-Stream: A multi density clustering algorithm for evolving data stream. *Journal of Network and Computer Applications*, *59*, pp.370-385.
17. Mohebi, A., Aghabozorgi, S., Ying Wah, T., Herawan, T. and Yahyapour, R., 2016. Iterative big data clustering algorithms: a review. *Software: Practice and Experience*, *46*(1), pp.107-129.
18. Qin, H., Ma, X., Herawan, T. and Zain, J.M., 2014. MGR: An information theory based hierarchical divisive clustering algorithm for categorical data. *Knowledge-Based Systems*, *67*, pp.401-411.
19. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y. and Herawan, T., 2014, June. Big data clustering: a review. In *International Conference on Computational Science and Its Applications* (pp. 707-720). Springer International Publishing.
20. Abdullah, Z., Herawan, T., Ahmad, N., Ghazali, R. and Deris, M.M., 2014, June. Mining Indirect Least Association Rule from Students' Examination Datasets. In *International*

Conference on Computational Science and Its Applications (pp. 783-797). Springer International Publishing.

21. Abdullah, Z., Mohd, F., Saman, M.Y.M., Deris, M.M., Herawan, T. and Hamdan, A.R., 2014. Mining critical least association rule from oral cancer dataset. In *Recent Advances on Soft Computing and Data Mining* (pp. 529-538). Springer International Publishing.
22. Abdullah, Z., Herawan, T. and Deris, M.M., 2014. Detecting Definite Least Association Rule in Medical Database. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp. 127-134). Springer Singapore.
23. Abdullah, Z., Herawan, T. and Deris, M.M., 2014. Mining Indirect Least Association Rule. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp. 159-166). Springer Singapore.
24. Abdullah, Z., Herawan, T., Chiroma, H. and Deris, M.M., 2014, June. A sequential data preprocessing tool for data mining. In *International Conference on Computational Science and Its Applications* (pp. 734-746). Springer International Publishing.