I.J. Information Technology and Computer Science, **2019, 7, 17-25**
Published Online July 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2019.07.03

A hybrid Technique for Cleaning Missing and Misspelling Arabic Data in Data Warehouse

Mohammed Abdullah Al-Hagery
Department of Computer Science, College of Computer, Qassim University, KSA
E-mail: hajry@qu.edu.sa, drmalhagery@gmail.com

Latifah Abdullah Alreshoodi, **Maram Abdullah Almutairi**, **Suha Ibrahim Alsharekh**, **Emtenan Saad Alkhowaiter**
Master Students, Department of Computer Science, College of Computer, Qassim University, KSA
E-mail: La.Alreshoodi@qu.edu.sa, maram.a.almutairi@gmail.com, soha.e.sh@gmail.com, e.alkhowaiter@qu.edu.sa

Received: 26 June 2018; Accepted: 19 May 2019; Published: 08 July 2019

***Abstract*—**Real-World datasets accumulated over a number of years tend to be incomplete, inconsistent and contain noisy data, this, in turn, will cause an inconsistency of data warehouses. Data owners are having hundred-millions to billions of records written in different languages, hence continuously increases the need for comprehensive, efficient techniques to maintain data consistency and increase its quality. It is known that the data cleaning is a very complex and difficult task, especially for the data written in Arabic as a complex language, where various types of unclean data can occur to the contents. For example, missing values, dummy values, redundant, inconsistent values, misspelling, and noisy data. The ultimate goal of this paper is to improve the data quality by cleaning the contents of Arabic datasets from various types of errors, to produce data for better analysis and highly accurate results. This, in turn, leads to discover correct patterns of knowledge and get an accurate Decision-Making. This approach established based on the merging of different algorithms. It ensures that reliable methods are used for data cleansing. This approach cleans the Arabic datasets based on the multi-level cleaning using Arabic Misspelling Detection, Correction Model (AMDCM), and Decision Tree Induction (DTI). This approach can solve the problems of Arabic language misspelling, cryptic values, dummy values, and unification of naming styles. A sample of data before and after cleaning errors presented.

***Index Terms*—**Data Cleaning, Missing Data, Arabic Misspelling, Data Quality, Data Consistency.

I. Introduction

Quality improvement in Data has become a critical issue for several companies and organizations because low data quality degrades organizational performance whereas enhanced data quality results in cost saving, customer satisfaction, and better analysis that will lead to a good Decision-Making. There are a lot of researchers developed various data cleaning methods of variant types of errors to improve data quality.

Data cleaning starts with identifying the expected problem when integrating various sources of data. A number of algorithms and techniques have been proposed and examined to improve the quality and consistency of the data warehouse, which is a central repository for integrated data. A real data tends to be incomplete, inconsistency, and noisy. These become problems of the data warehouse. Therefore, a comprehensive method is required to solve such problems.

There are many solutions to keep the data warehouse consistency, but all that solutions are limited to specific problems. In data warehouses, data cleaning is the main part of Extraction, Transformation, and Loading (ETL). Data cleaning helps data specialists to determine and detect all inconsistent and incomplete data items [1]. The automate data cleaning provides a solution to find and solve some of these problems and make datasets partially on a consistency level. Hence, there are many techniques such as Decision Tree algorithm that is handling missing values, it can be merged with other techniques to solve the most of data errors.

The proposed approach is an extension to improve the smart technique presented in [2], Which focused on the cleaning of some errors of datasets, such as redundant values, error formats, numeric errors of values, and Arabic mistakes in sentences. Many techniques that have applied to solve common problems and specific kinds of errors still limited. Consequently, we proposed a hybrid method that can deal with many issues, such as Arabic language misspelling, missing data, cryptic values, dummy values, and unification of naming style problem, where one attribute like department name can hold different naming styles.

The paper is organized as follows: Section 2 describes the various methods and techniques used for data cleaning and corrections. Section 3 describes the research methodology, it explains the different steps of

the proposed approach and ended with a proposed algorithm. In section 4, the results are discussed, while section 5 presents the conclusions. Finally, section 6 describes the future work.

II. Related Works

Data into the data warehouse usually collected from different sources that contain errors, missing values, contradicting data, cryptic data, noisy data, misspelling errors, etc. Currently, many types of research works focus on data cleaning from many types of errors to maintain data consistency and enhance its quality. There are many techniques to deal with missing values. Some techniques deal with missing values by deleting the records that had these values. However, deleting these records of a small dataset generally reduces the sample size and its usability for several statistical analyses. Moreover, the statistical analysis results getting from a data set having an insufficient number of records are not strong, because of incomplete and noisy data deletion. In the following, we will discuss common methods of this regard. These methods are employed to fill in missing data using the median, mode or mean.

In terms of knowledge discovery from data, the application of the mean replacement method can often lead to more mislead results than the method of simple record deletion [3]. The transitive closer algorithm employed in data cleaning to find the duplicates in datasets explained in detail [4]. A transitive closure infilling method of missing values, removing redundancies of data items and the combination of similar data items together discussed [5]. Wang et al. analyzed a set of problems relevant to data cleaning illustrated by the results [6]. As well, a hybrid method developed to clean data using enhanced versions of two basic techniques namely PNRS and Transitive Closure explained in [7]. On the other hand, an educational data mining field a special system had developed and examined by the Decision Tree to solve the educational problems of datasets [8]. The missing value is one of the most problems found, so researchers use some common algorithms to solve this problem such that ID3, CART, and C4.5.

Likewise, a simple technique was proposed to handle and copy of missing values by using the Decision Tree technique. This technique called messiness in the MIA and the EMMI algorithm used for the same purpose for range data type then copying in missing values [9]. Yang et al. discuss the test costs method to query missing values, by applying the Test Sensitive Classification Learning (TSCL) algorithm to minimize the test and classification cost [10]. Also, some imputation techniques such as the K-Nearest Neighbor, Mean-Mode, Hot_Deck, Expectation Maximization, and C5.0 used to impute artificially created missing data from various datasets. The techniques performance was compared based on the accuracy of the classification of the original data and the imputed data.

Identifying useful imputation technique provided accurate results of classification as presented [11]. A number of solutions were provided to clean cryptic, dummy values, and contradicting data, which represent the most values, in this regard. For instance, Swapna et al. assume some dummy values before comparing with the database values. Whenever Data mining query finds this assumed value, it directly knows that it is a dummy value [12]. In the same direction, there has been a remarkable series of tasks for capturing the general errors of data as violations of data integrity restrictions as accomplished [13-17]. Many data repairing methodologies have been proposed, developed, and applied [18-25]. A set of heuristic algorithms were developed for data preprocessing [18,20], based on functional dependencies [17,26], involved in inclusion dependencies [18], Conditional Functional Dependencies (CFDs) [27], the CFDs, Matching dependencies [22], and denial constraints [19].

Some research works employ confidence values placed by users to guide a repairing process [18,20,22] or use master data [28]. The statistical inference process is proposed to derive missing values [24]. In order to ensure the quality and the accuracy of data repairs required consulting professionals or users [24,25,28].

In more specific, the spelling errors detection and correction in the English Language mostly investigated, many types of research are dealing with the English language errors. On the other hand, there is a limited set of approaches were applied to solve this problem the Arabic language because it is very complicated, where [29] established an error detection and candidate generation model to detect the misspellings, and generate their possible candidates then design a misspelling correction system. Furthermore, a hybrid system based on the confusion matrix and Noisy Channel Spelling Correction Model developed to detect and correct Arabic spelling errors [30], A spelling error detection and correction method targeted at non-native Arabic language learners evaluated by a team of specialists. They applied edit distance algorithm and rule-based transformation approach. Rules are formulated according to their study on common spelling errors made by Arabic learners to handle the noisy data [31,32]. However, these approaches are partially different from our proposed approach, where, the most of them were concentrated on limited problems and used different mechanisms.

The Arabic edited distance algorithm proposed to measure the similarity between two Arabic strings. This type of algorithm is working based on a Levenshtein algorithm [32], likewise, a system developed for 'Arabic character' recognition using guessing the Fast Fourier Transform descriptors for counting the basic part of characters presented by [33]. The system used ten features to classify the character models. Also, Higazy et al. had developed an English/Arabic enabled web-based framework, considering the wide range of variations in the Arabic language to allow the fast processing, improved indexing/blocking techniques are

used [34]. There is a set of cleaning tools available for data cleaning. It concentrates on specific fields, such as cleaning address and name or a specific attribute to remove the duplication. Also, it is restricted to a specific domain, these tools usually execute very well but must be integrated with other tools to fix most of the cleaning problems that affecting the data integration.

In addition, Hamad and Jihad developed an enhanced method works on the identification of error corrections in data. Such as domain format error, lexical mistakes, constraint violation, irregularities, integrity, and redundant values [36].

Other tools developed for Extract, Transform and Load (ETL), which have comprehensive transformation and workflow capabilities help to transform and lad then clean raw data. These tools developed to cover the most of the data transformation and cleaning tasks. There is a common problem with these tools that is the difficulty to merge the functionality of various tools together [3,37]. A set of Key elements discussed in a survey for cyberspace [6]. An automated technique provided for proof negative and positive during data cleaning, based on reference tables and Sherlock Rules to update them when enough information is available. The Sherlock Rules methods are able to repair and annotate un-cleaned instances in a deterministic fashion with high precision [38]. Data cleaning also improved and it puts forward a method named an association mining data cleaning [39].

In addition, Al-Hagery et al. had developed a software tool to clean, integrate, and transform the content of a big dataset of the Hepatitis Disease written in the English language [40]. Furthermore, a smart algorithm proposed to identify the most data errors including errors in data formats, redundant data, data duplicity, limited Arabic language mistakes, numeric errors, and symbol errors. However, it did not consider some type of errors in the data, that is why we propose our method [41].

III. Research Methodology

The proposed technique is similar to the smart technique presented by [41], It is focusing on the Arabic language misspelling because there are no more perfect techniques to deal with this subject. This Technique planned to work on the data gathered from multiple sources before the loading to the data warehouse. This technique tested and validated using a set of real data instances. Fig.1 illustrates the headlines of the research methodology. The methodology includes; Arabic misspelling detection, misspelling correction, fill missing data, cryptic values, dummy values, unification of naming styles, and development of a hybrid algorithm to clean missing and misspelling arabic data.

The data set used in this paper consists of seven attributes as shown in Table 1. The data sample used before processing is presented in Table 2 while Table 3 will contain the data after processing. The Students' Names placed in Arabic and in English (STD_Name_AR and STD_Name_EN), were included in these tables.

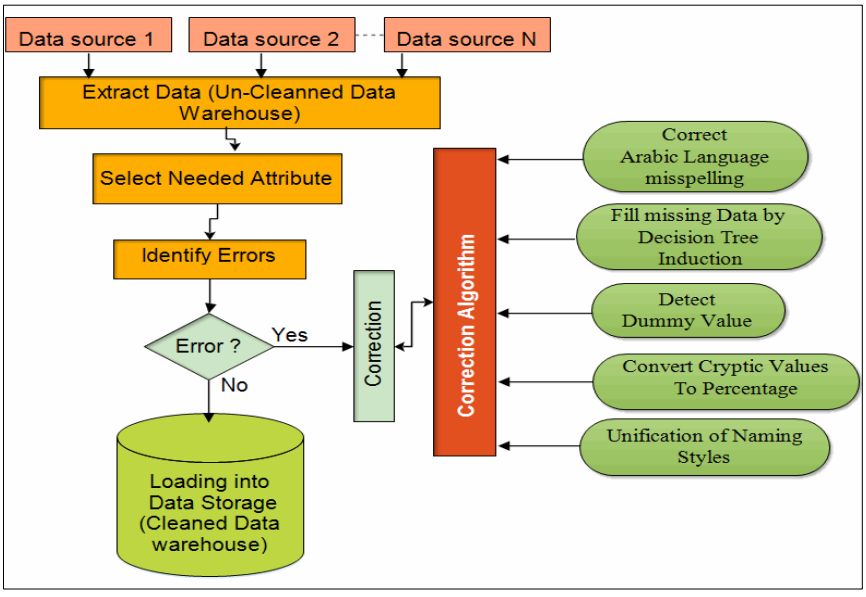

Fig.1. Flowchart of the Proposed Approach for Data Cleaning

A. Arabic Misspelling Detection and Correction

The structure of AMDCM used to solve the Arabic Misspelling problem is presented in Fig.2, the steps of this approach will be discussed in the following sections:

1. Misspelling Detection: The spelling error detection is a process of detecting incorrect word/words (non-word) using different detection methods. The most common technique is the N-grams algorithm. This algorithm is an N letter subsequence of a string, where N is usually mono, bi or tries to determine if each gram in an

input word is expected to be valid or invalid in the language. Some techniques use two-level morphological analysis to check whether an input value follows those language morphological rules. In addition, some machine-learning algorithms as hybrid techniques are using for Misspelling Detection. However, the most widely used technique in many applications for Misspelling Detection is the dictionary lookup technique, where the input word is compared with the words in the dictionary, if the input word is not found in the dictionary, the word is considered as an incorrect word. This technique is unlike the morphological analysis, it guarantees that the most common words used are covered, but it will not provide a complete coverage of the language. In our approach we used the dictionary lookup techniques to detect Arabic Misspelling errors since this technique can handle the most common words used, faster than others, it is a language independent. In this model, the size of our proposed dictionary is about 427 K unique words, where these words organized in the dictionary in an ascending order to speed-up the comparison using a search algorithm. All words that begin with the same letter will be placed in a separate table, to improve the comparison.

2. Misspelling Correction: Techniques used for spelling errors correction, it can be classified into context-dependent and context-independent error corrections. In this paper, our focus is on context independent error correction techniques, which perform in an isolated word or words. In order to propose an efficient misspelling correction model for the Arabic language, we significantly need to study and categorize common error patterns. Fig.2 shows the steps and the chart of AMDCM.

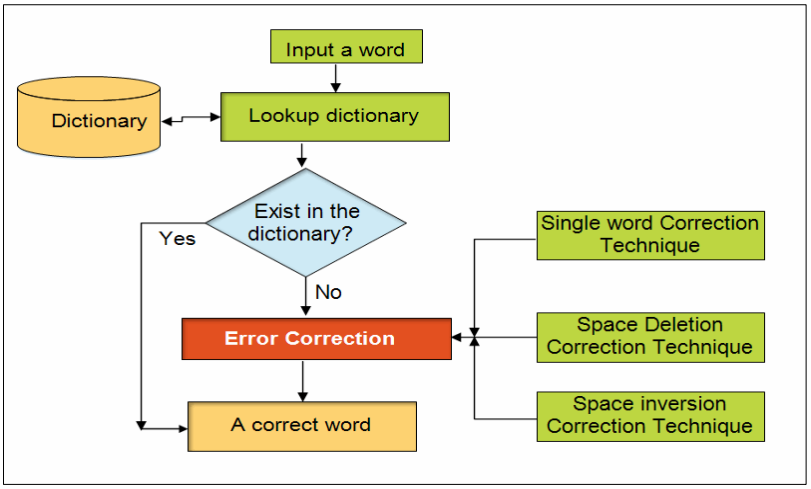

Fig.2. Arabic Misspelling Detection and Correction Model

The Misspelling Correction includes the following tasks:

- Single-word Errors: The misspellings appear because of omitting, adding, replacing or duplicating a letter within a given word by mistake. For example, spelling the word 'سلييم' instead of 'سليم - Saleem', in this case, misspelling happened by duplicating the letter 'يـ'. Also, may omit a letter identical to 'ب', cause the misspelling such as: 'مرت' instead of 'مرتب – muratab' and adding an extra letter 'ن' in a word as 'خريفن' instead of 'خريف - khareef', or by replacing a letter in a word as 'قروب' instead of 'قريب- qareeb'.

For these types of errors, several techniques and algorithms can be used together to perform a context-independent error correction such as rule-based, N-gram, probabilistic, and neural network techniques. According to [31] a rule-based technique represents common spelling error patterns in the form of rules, which are used to transform misspelt words with the correct word. A neural network technique performs associative recall based on the incomplete input. A probabilistic technique works based on N-gram technique, which is used either as a standalone or with other techniques to perform error correction. Also, the similarity key technique can employ to transform words with keys as applied [29]. Therefore, computing a key to misspelling word will point to similarly spell words of the dictionary.

The most common technique used for misspelling correction is the edit distance, which use the dynamic programming to find out the lowest cost of editing operations (insert, delete, and replace) in order to transform string Str_1 into new string Str_2 by giving a cost value of edit operations [32]. As mentioned in the example given by [41], to correct the word 'كبيير' to 'كبير kabeer' based on the Arabic dictionaries and unifying the text into standard word. The edit distance is a more efficient technique for calculating the distance between

the incorrect, the correct word in the dictionary and select the minimum cost word. In that example, we measure the distance between the previous two Arabic words, as shown in Fig.3, the edit distance technique deletes the second letter in that word, which is 'ي-yaa'. The deletion operation of one letter in this technique is costing one.

		ك	ب	ي	ر
	0	1	2	3	4
ك	1	0	1	2	3
ب	2	1	0	1	2
ي	3	2	1	0	1
ي	4	3	2	1	1
ر	5	4	3	2	1

Fig.3. Edit Distance

- Space Deletion Errors: When the user/editor forgets to put spaces between words that caused the space deletion errors. A sequence of Arabic letters was selected with the maximum marginal probability of A* lattice search, N-gram probability estimation, and using 15-grams language model of Arabic letters. This technique is more suitable for long Arabic texts because it will reduce the complexity of processing [29]. However, our exhaustive search model is more appropriate because the field of the search will not contain more than 3-4 words so, merging these words will not be much complicated.

Correction of the deleted spaces or adding spaces in the word and try to divide it into S_i correctly spelt words. For example, the STD_Department column contains the value 'علومالحاسب' as a single entry, the dictionary will classify it as an incorrect word. In this case, an exhaustive search process will find all possible correct words *Si*, where *Si*<1 from the dictionary, then the sequence of letters of 'علومالحاسب', It will be spelt into two words 'علوم' and 'الحاسب', by insertion of the deleted space.

- Space Insertion Errors**:** When the user adds one or more spaces in a single word that causes the space insertion errors. The proposed algorithm [29] can be used to find the possible merging suggestions for an incorrect phrase. In our model, we will use the exhaustive search as in the correction of space deletion errors (discussed above) by trying to concatenate *Si* incorrect words sequence to find all possible correct words of the dictionary. For example, if the user enters in the STD_Department name column, in Table 2, this string consists of two different names of a student; the first name and the second name: ' سلي مان مح مد' after the detection processes many spaces in these two words of both names, it will process these words as errors. The exhaustive search will be applied to concatenate these incorrect words by removing the space everywhere from followed incorrect words to find all possible correct words, it will be corrected for the previous example as ' سليمان' and ' محمد'.

B. DTI Algorithm to Fill Missing Data

The DTI algorithm applied [12] to fill the missing data for each data attribute, there is one binary tree that will be generated. Each Decision Tree involves a set of nodes and each node is a test on that attribute. Two paths will come out from each value/node except for the end leaf nodes. The first path is "No" and the second is "Yes". The missed values are placed in the Leaf Nodes.

The missing data is replaced by the data placed in the leaf nodes of the tree. Initially, data in the form of rows are collected from different educational data sources. Each row is a combination of several attributes.

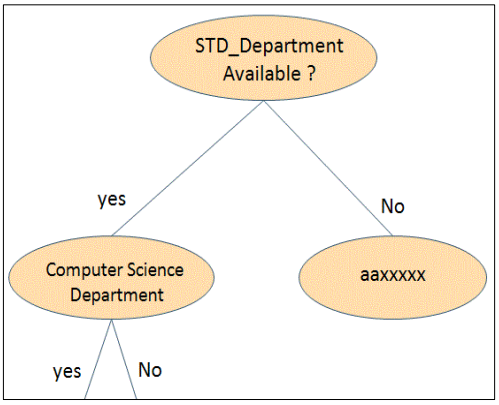

Fig.4. Decision Tree for STD_Department

Fig.4 illustrates a Decision Tree for a set of students, it presents how Decision Trees is created for STD_Department attribute values, it checks the value of each row until it reaches to the missing value then replaces it with the leaf Node value. likewise, a Decision Tree is used to fill in the missing data by the value found in the leaf node. The Decision Tree was chosen because it is simple, fast, and it does not require any domain knowledge.

Table 1. Field Description of Student Data Set

Field Name	Description	Field type
STD_Name_AR	Student Name in the Arabic language	String
STD_Name_EN	Student Name in the English language	String
STD_Age	Student Age	Number
STD_Department	Student Department	String
STD_GPA	Student Grade Point Average	Number
Nationality	The nationality of Student	String
Gender	Male or Female	String

C. Cryptic Values

A simple example of Cryptic Values is the Cumulative Grade Points Average (CGPA). If the student fills his/her CGPA as a 4.6, the cleaning process should convert this value to the percentage, because most of the universities prefer percentages than CGPA [12]. In our example, if the user enters the CGPA such as 4.50 out of five, it will be converted to percentages by the division of five then multiplied the result by 100.

D. Dummy Values

Usually, the dummy values are those values generated in a huge range on the Internet. For example, Email account registrations, other accounts in Tweeters, Facebook, WhatsApp or other applications. Currently, there is no existing algorithm to find the dummy values. Each field follows a special method to identify and solve the dummy values. The following are a set of examples given as dummy values.

عععع نننننن قققفف هههها أأأأأ ككككك تتت ييييي ببب
جججج سسسس.

The next step of identifying these types of values is replacing each one with a global value "Dummy". Later after a query is applied in data mining, whenever mining process finds value "Dummy" in Arabic 'قيمة وهمية', it directly knows that this value is a dummy value. The mining process will not use this kind of values. Therefore, the final results will be more accurate after the cleaning task.

E. Unification of Naming Styles

The use of various naming styles makes problem in data analysis and data mining because of the inconsistencies of data. The unification of naming styles is important for data cleaning to improve its quality. This helps to get a better analysis of results. These are two examples of different naming. The first is the weight can be represented in tons 'طن' or in kilogram 'كيلوجرام' and the distance can be represented in kilometres 'كيلومتر' or in mils 'ميل'. In our sample of data, the CGPA represented in two different measures as in Table 2, in the column of STD_CGPA.

Table 2. A sample of Dataset for Students in Computer College Before Cleaning

ID	STD_Name_AR	STD_Name_EN (name translation)	STD_Age	STD_Department	STD_CGPA	Gender
1	أح مد خا لد	Ahmed Khaled	'ببب'	علومالحاسب	83.0	ذكر
2	بدر احمد	Bader Ahmed	24	علوم حاسوب	4.45	ذكر
3	يوسف سيد	Yousef Sayed	22	تقنية معلومات	60.5	انثى
4	دلال ابرا هيم	Dalal Ibrahim	20	تقنية المعلومات	3.15	انثى
5	سعد خخليل	Saad Khalil	21	تكنولوجيا المعلومات	5.00	null
6	عمر محمو	Omar Qamar	23	هندسة الحاسب	82.0	ذكر
7	فاتنسامي	Faten Sami	22	هندسة البرمجيات	72.5	انثى
8	سلي مان مح مد	Suliman Mohammed	24	علوم الحاسوب	84.8	ذكر
:	:	:	:	:	:	:
98	رهفف ناصر	Rahf Nasser	24	هندسة البرامجيات	4.66	انثى

F. The proposed Algorithm

The following algorithm is summarizing the mechanism of the hybrid method for data cleaning depend on various techniques.

1) *Start*
2) *Use the dataset (table/tables) collected from the Original Sources So_1, So_2, …, So_n.*
3) *Chick the dataset field by field, identify the error types*
4) *Process the **Arabic Language Misspelling**, check all words based on lookup dictionary technique, in case of misspelling errors, and the AMDCM model will correct the misspelling through several techniques based on the following error (Single Word Errors, Space Deletion Errors, Space Insertion Errors).*
5) *For each **Missing Data**, use the DTI to fill the required value.*
6) *For **Cryptic Data** values, check the field of this attribute then convert the values to the preferable*

range

7) *For **Dummy Values**, check the data in this field if it's dummy value by comparing it with the assumed dummy values such as 'ببب, تتت, !!!!, ...' then replace it with the "Dummy" word.*
8) *For all incompatible naming styles, use the most common name in that field to unify all names.*
9) *Save the cleaned data and exit*
10) *End.*

IV. Results Discussion

When integrating data gathered from multiple sources, the noisy data appears and the needs for data cleaning are increased. In this paper, the proposed approach applied to a sample of data contains 98 data records (presented in Table 2) to test the validation of the proposed approach, this sample contains data before the cleaning process. It contains various types of errors occurred during data entry, such as missing letters, missing some fields, extra space 'أح مد خا لد' & 'دلال ابرا هيم', (column 2, row 1 and column 2, row 4), these items converted into Table 3 to 'أحمد خالد' & 'دلال ابراهيم', respectively. Moreover, there is another type of errors was discovered and corrected, it is the doubling of one letter, as in the content of row 5 and row 98 (field of STD_Name_AR).

In addition, some words there is Arabic Misspelling related to adding or deleting spaces in a single words or may omit a letter in a word or words, these problems were solved by adding the omitted letter or by inserting of a space in the required position, as shown in the final result (Table 3, STD_Name_AR row 6 and 7).

Additionally, various naming styles found in the dataset sample, as in column "STD_Department", which contain a set of variable names with different naming styles in the following list:

{ علوم حاسب، علوم حاسوب، تقنية معلومات، تقنية المعلومات هندسة البرمجيات، هندسة البرامجيات}.

These names had unified based on the application of the proposed model. The results of the unification process appear in Table 3 contain only the elements of this list:

{علوم الحاسوب، تقنية المعلومات، هندسة البرمجيات}

In STD_CGPA attribute, there are different measures to calculate this value. In some sources, CGPA calculated as scores out of five and others calculated as a percentage that may cause inconsistency problem with the whole data (see the contents of column sixth). Besides, some missing data found in that sample, precisely in the gender column, which contains a null value. According to the proposed approach, the "null value" found at the last column Table 2, row 5 replaced by 'ذكر', which means "male", based on the type of the Arabic name using the dictionary, the final result shown in Table 3.

The dummy value denoted by 'ببب' found in the first row, column STD_Age was replaced with a global value 'Dummy', this value will be replaced by the average value of all values of this column during the data analysis process. The application of the proposed approach is a transformation process of the dirty dataset as in Table 2 (low-quality data) to a new content placed in Table 3, as a (high-quality dataset). This approach is increasing the accuracy of the final results produced from data analysis using the traditional methods or using the advanced methods such as mining in data or knowledge discovery from data.

The AMDCM model has applied in this paper to detect and correct Arabic Misspelling and unification of naming styles and other types of errors that appear as a result of merging different sources of a dataset.

Table 3. A sample of Data for Student's Information in Computer College, After Cleaning

ID	STD_ Name_AR	STD_Name_EN	STD_Age	STD_ Department	STD_CGPA	Gender
1	أحمد خالد	Ahmed Khaled	'Dummy'	علوم الحاسوب	83.0%	ذكر
2	بدر احمد	Bader Ahmed	24	علوم الحاسوب	89.0%	ذكر
3	يوسف سيد	Yousef Sayed	22	تقنية المعلومات	60.5%	انثى
4	دلال ابراهيم	Dalal Ibrahim	20	تقنية المعلومات	63.0%	انثى
5	سعد خليل	Saad Khalil	21	تكنولوجيا المعلومات	100%	ذكر
6	عمر محمود	Omar Qamar	23	هندسة الحاسب	82.0%	ذكر
7	فاتن سامي	Faten Sami	22	هندسة البرمجيات	72.5%	انثى
8	سليمان محمد	Suliman Mohammed	24	علوم الحاسوب	84.8	ذكر
:	:	:	:	:	:	:
98	رهف ناصر	Rahf Nasser	24	هندسة البرمجيات	93.2%	انثى

V. Conclusion

One of the most important steps of data processing is the verification of data values are correct. Data cleaning is an important task for data warehousing development. In this paper, we proposed a comprehensive approach includes most of the data cleaning methods, this approach is able to check and repair the Arabic words that have misspelling, through the DIT model to fill in missing data and AMDCM model to correct the misspelling based on error type: Single Word Errors, Space Deletion Errors and Space Insertion Errors, etc. Moreover, this approach can solve the problems of cryptic values, dummy values, and unification of various naming styles. A sample of data has presented before and after errors cleaning to show the validation

of this approach. The implementation of this model as a real application will increase the data consistency and quality, this, in turn, will help to get accurate results for data analysis, especially in mining purposes in data, knowledge discovery, and improve the Decision-Making.

VI. Future work

The proposed approach can be modified according to the following perspectives:

1. Increase the size of the Lookup Techniques Dictionary to cover all possible cases.
2. Develop this approach as a software application, as an implementation of the algorithm of the proposed approach to be applied on big datasets.
3. Improve this approach to cover multilingual, exceeded the Arabic and English languages.

References

[1] N. Debbarma, "Analysis of Data Quality and Performance Issues in Data Warehousing and Business Intelligence," vol. 79, no. 15, pp. 20–26, 2013.

[2] S. B. Alotaibi, "ETDC," in *Proceedings of the International Conference on Advances in Image Processing - ICAIP 2017*, 2017, pp. 135–138.

[3] G. Rahman and Z. Islam, "iDM I : A Novel Technique for Missing Value Imputation using a Decision Tree and Expectation-Maximization Algorithm," no. March, pp. 8–10, 2014.

[4] M. Hernández and J. Stolfo, "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem," *Data Min. Knowl. Discov.*, vol. 2, pp. 9–37, 1998.

[5] W. N. Li, R. Bheemavaram, and X. Zhang, "Transitive closure of data records: Application and computation," *Int. Ser. Oper. Res. Manag. Sci.*, vol. 132, pp. 39–75, 2010.

[6] J. Wang, H. Zhang, B. Fang, X. Wang, and L. Ye, "A Survey on Data Cleaning Methods in Cyberspace," *2017 IEEE Second Int. Conf. Data Sci. Cybersp.*, pp. 74–81, 2017.

[7] A. Paul, V. Ganesan, J. S. Challa, and Y. Sharma, "HADCLEAN: A hybrid approach to data cleaning in data warehouses," *2012 Int. Conf. Inf. Retr. Knowl. Manag.*, pp. 136–142, 2012.

[8] P. Patidar and A. Tiwari, "Handling Missing Value in Decision Tree Algorithm," *Int. J. Comput. Appl.*, vol. 70, no. 13, pp. 975–8887, 2013.

[9] B. E. T. H. Twala, M. C. Jones, and D. J. Hand, "Good methods for coping with missing data in Decision Trees," *Pattern Recognit. Lett.*, vol. 29, no. 7, pp. 950–956, 2008.

[10] Q. Yang, S. Member, C. Ling, X. Chai, and R. Pan, "Test-Cost Sensitive Classification on Data with Missing Values," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 5, pp. 626–638, 2006.

[11] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," *Proc. 2016 Int. Conf. Data Sci. Eng. ICDSE 2016*, 2017.

[12] S. Swapna, P. Niranjan, B. Srinivas, and R. Swapna, "Data Cleaning for Data Quality," *2016 3rd {{International Conf. {{Computing}} {{Sustainable Glob. Dev.*, pp. 344–348, 2016.

[13] X. Lian, L. Chen, and S. Song, "Consistent query answers in inconsistent probabilistic databases," *Proc. 2010 Int. Conf. Manag. data - SIGMOD '10*, p. 303, 2010.

[14] L. Bravo, W. Fan, and S. Ma, "Extending Dependencies with Conditions," *Constraints*, pp. 243–254,

[15] J. Chomicki and J. Marcinkowski, "Minimal-change integrity maintenance using tuple deletions," *Inf. Comput.*, vol. 197, no. 1–2, pp. 90–121, 2005.

[16] W. Fan, F. Geerts, N. Tang, and W. Yu, "Inferring data currency and consistency for conflict resolution," *Proc. - Int. Conf. Data Eng.*, pp. 470–481, 2013.

[17] S. Kolahi and L. V. S. Lakshmanan, "Inconsistency in Databases," *Icdt*, no. March 2009.

[18] P. Bohannon, W. Fan, M. Flaster, and R. Rastogi, "A cost-based model and effective heuristic for repairing constraints by value modification," *Proc. 2005 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '05*, p. 143, 2005.

[19] I. F. I. P. P. Xu Chu, "Holistic data cleaning: Put violations into context," no. L, pp. 458–469, 2013.

[20] G. Cong *et al.*, "Improving data quality: consistency and accuracy," *Proc. 33rd Int. Conf. Very large data bases*, vol. Vienna, Au, pp. 315–326, 2007.

[21] A. C. Gohel, A. V Patil, P. P. Vadhwana, and H. S. Patel, "A commodity data cleaning system," *Int. Res. J. Eng. Technol.*, vol. 4, no. 5, 2017.

[22] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu, "Interaction between record matching and data repairing," *Proc. 2011 Int. Conf. Manag. data - SIGMOD '11*, vol. 1, no. 1, p. 469, 2011.

[23] F. Geerts, G. Mecca, P. Papotti, and D. Santoro, "The L LUNATIC Data-Cleaning Framework," *Proc. VLDB Endow.*, vol. 6, no. 9, pp. 625–636, 2013.

[24] C. Mayfield, J. Neville, and S. Prabhakar, "ERACER: A Database Approach for Statistical Inference and Data Cleaning," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 75–86, 2010.

[25] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas, "Guided data repair," *Proc. VLDB Endow.*, vol. 4, no. 5, pp. 279–289, 2011.

[26] M. Volkovs, J. Szlichta, and R. J. Miller, "Continuous data cleaning," *2014 IEEE 30th Int. Conf. Data Eng.*, pp. 244–255, 2014.

[27] P. Bohannon, W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, "Conditional functional dependencies for data cleaning," *Proc. - Int. Conf. Data Eng.*, pp. 746–755, 2007.

[28] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu, "Towards certain fixes with editing rules and master data," *VLDB J.*, vol. 21, no. 2, pp. 213–238, 2012.

[29] M. I. Alkanhal, M. A. Al-badrashiny, M. M. Alghamdi, and A. O. Al-qabbany, "Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions," vol. 20, no. 7, pp. 2111–2122, 2012.

[30] H. M. Noaman, S. S. Sarhan, and M. A. A. Rashwan, "Automatic Arabic Spelling Errors Detection and Correction Based on Confusion Matrix- Noisy Channel Hybrid System," *J. Theor. Appl. Inf. Technol.*, vol. 40, no. 2, pp. 54–64, 2016.

[31] K. Shaalan, R. Aref, and A. Fahmy, "An Approach for Analyzing and Correcting Spelling Errors for Non-native Arabic learners," 2017.

[32] H. H. A. Ghafour, A. Ei-bastawissy, and A. F. A. Heggazy, "AEDA : Arabic Edit Distance Algorithm Towards A New Approach for Arabic Name Matching," pp. 307–311, 2011.

[33] S. A. Mahmoud and A. S. Mahmoud, "Arabic character recognition using Modified Fourier Spectrum (MFS) Vs. fourier descriptors," *Cybern. Syst.*, vol. 40, no. 3, pp. 189–

210, 2009.
[34] A. Higazy, T. El Tobely, A. H. Yousef, and A. Sarhan, "Web-based Arabic/English duplicate record detection with nested blocking technique," *Proc. - 2013 8th Int. Conf. Comput. Eng. Syst. ICCES 2013*, pp. 313–318, 2013.
[35] S. Kamble and V. Kohle, "A novel approach of data cleaning/cleansing detecting, editing," *Int. J. Acad. Res. Dev.*, vol. 2, no. 3, pp. 84–88, 2017.
[36] M. M. Hamad and A. A. Jihad, "An enhanced technique to clean data in the data warehouse," *Proc. - 4th Int. Conf. Dev. eSystems Eng. DeSE 2011*, pp. 306–311, 2011.
[37] E. Rahm and H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
[38] M. Interlandi and N. Tang, "Proof positive and negative in data cleaning," *Proc. - Int. Conf. Data Eng.*, vol. 2015–May, pp. 18–29, 2015.
[39] L. Zhai, M. Wu, S. Zhang, Q. Zhao, and T. Li, "Research on association mining data cleaning for the professional field," in *Proceedings of 2013 2nd International Conference on Measurement, Information and Control, ICMIC 2013*, 2013, vol. 1, pp. 563–566.
[40] M. A. Al-hagery, "Knowledge Discovery in the Data Sets of Hepatitis Disease for Diagnosis and Prediction to Support and Serve Community," *Int. J. Comput. Electron. Res.*, vol. 4, no. 6, pp. 118–125, 2015.
[41] S. B. Alotaibi, "ETDC : An Efficient Technique to Cleanse Data in the Data Warehouse," pp. 135–138, 2017.

Authors' Profiles

Mohammed Abdullah Al-Hagery: received his B.Sc in Computer Science from the University of Technology in Baghdad Iraq-1994. He got his MSc. in Computer Science from the University of Science and Technology Yemen-1998. Al-Hagery finished his Ph.D. in Computer Science and Information Technology, (Software Engineering) from the Faculty of Computer Science and IT, University of Putra Malaysia (UPM), November 2004. He was a head of the Computer Science Department at the College of Science and Engineering, USTY, Sana'a from 2004 to 2007. From 2007 to this date, he is a staff member at the College of Computer, Department of Computer Science, Qassim University, KSA. He published more than 20 papers in various international journals. Dr. Al-Hagery was appointed a head of the Research Centre at the Computer College, and a council member of the Scientific Research Deanship Qassim University, KSA from September 2012 to October 2018. Currently, he is teaching the master degree students and a supervisor of four master thesis. He is a jury member of a number of PhD and master thesis, as an internal and external examiner in his field of his specialist.

Latifah Abdullah Saleh Al-Reshoodi: received her B.Sc. in Information Technology from the College of Computer, Qassim University, Saudi Arabia in 2014. Currently, she is studying MSc in Computer Science, at Qassim University. In addition, from 2016 to this date, she is working as a Teaching Assistant in the Computer College, Qassim University, Saudi Arabia.

Maram Abdullah Al-Mutairi: received her B.Sc in Software Engineering from Ha'il University-KSA in 2014. Currently, she is studying MSc in Computer Science, Qassim University.

Suha Ibrahim Abdullah Alsharekh: received her B.Sc in Computer Science, from the College of Computer, Qassim University, Saudi Arabia in 2015. Currently, she is studying MSc in Computer Science, Qassim University.

Emtenan Saad Al-Khowaiter: received her B.Sc. in Information Technology from the College of Computer at Qassim University, Saudi Arabia -2014. the Currently, she is studying MSc in Computer Science, at Qassim University. In addition, from 2016 to this date, she is working as a Teaching Assistant in the Computer College, Qassim University, Saudi Arabia.

How to cite this paper: Mohammed Abdullah Al-Hagery, Latifah Abdullah Alreshoodi, Maram Abdullah Almutairi, Suha Ibrahim Alsharekh, Emtenan Saad Alkhowaiter, "A hybrid Technique for Cleaning Missing and Misspelling Arabic Data in Data Warehouse", International Journal of Information Technology and Computer Science(IJITCS), Vol.11, No.7, pp.17-25, 2019. DOI: 10.5815/ijitcs.2019.07.03