A Token-Based Data Cleaning Technique for Land Application System

Tin Nilar Aye, May Aye Khine
University of Computer Studies, Yangon, Myanmar
tinilaraye18@gmail.com, tinnilaraye18@gmail.com

Abstract

The process of detecting and removing database defects and duplicates is referred to as data cleaning. The fundamental issue of duplicate detection is that inexact duplicates in a database may refer to the same real world object due to errors and missing data. Duplicate elimination is hard because it is caused by different types of errors like typographical errors, missing values, abbreviations and different representations of the same logical value. If the database contain duplicate records, it is difficult to analyze the database as well as difficult to extract the required data. To get qualified data, data cleaning must be performed. This paper proposes a system to resolve dirty data in the database and ensuring to get clean data. This paper concentrates on the duplicate data problem and this study can be smoothed by using token-based data cleaning technique.

1. Introduction

Cleansing data from impurities is an integral part of data processing and maintenance. Data cleansing is the process of eliminating the errors and the inconsistencies in data, and solving the object identity problem. The major areas of data cleansing are: data warehousing, knowledge discovery in databases (also termed data mining), and data/information quality management. Within the data warehousing field, data cleansing is applied especially when several databases are merged. Records referring to the same entity are represented in different formats in the different data sets or are represented erroneously. Thus, duplicate records will appear in the merged database. The issue is to identify and eliminate these duplicates. The problem is known as the merge/purge problem.

Instances of this problem appearing in literature are called record linkage, semantic integration, instance identification, or object identity problem. Data cleansing is applied with varying comprehension and demands in the different areas of data processing and maintenance. The original aim of data cleansing was to eliminate duplicates in a data collection, a problem occurring already in single database applications and gets worse when integrating data from different sources. Data cleansing is therefore often regarded as integral part of the data integration process. Besides elimination of duplicates, the integration process contains the transformation of data into a form desired by the intended application and the enforcement of domain dependent constraints on the data. Data cleansing is the process of eliminating the errors and the inconsistencies in data, and solving the object identity problem.

2. Motivation

Data cleaning 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. Quality of data mining results crucially depends on quality of input data. The meaning of data quality means completeness, uniqueness, consistency, accuracy, etc. Converting standard data form is obviously an important step to identify the tokens. To eliminate duplicated records from a dataset, the main fact is how to decide that two records are duplicated? We need to compare records to determine their degree of similarity by using defined tokens. The comparison of fields to determine whether or not two values are alternative representations of the same semantic entity. Duplicated information will produce incorrect or misleading statistics. All data cleaning is a preprocessing stage before loading the transformed data into the warehouse. So, the need for data cleaning increases significantly in order to provide access to accurate and consistent data.

3. Data Quality and Data cleaning Process

To be processable and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. Data Quality criteria are:

- **Accuracy**: An aggregated value over the criteria of integrity, consistency and density
- **Integrity**: An aggregated value over the criteria of completeness and validity
- **Completeness**: Achieved by correcting data containing anomalies
- **Validity**: Approximated by the amount of data satisfying integrity constraints
- **Consistency**: Concerns contradictions and syntactical anomalies
- **Uniformity**: Directly related to irregularities
- **Density**: The quotient of missing values in the data and the number of total values ought to be known
- **Uniqueness**: Related to the number of duplicates in the data

Data cleaning first detects dirty records by determining whether two or more records represented differently refer to the same real world entity, and then, it cleans the dirty records by either (i) collapsing them to get a consolidated whole devoid of missing parts, (ii) unifying them with a single entity identity and (iii) retaining only one copy of records that are exact duplicates. Two main causes of "Dirt" or conflicts in data are synonyms and homonyms, though there are many others such as: "incomplete, missing or null values", "spelling, phonetic or typing errors", "Miss-fielding", "noise or contradicting entry", values outside the accepted range, "scanning errors", "type mismatch". Homonymous dirt arises when the same "term" or "expressions" refer to two or more entities, e.g., many occurrences of "Daw Aye" in a data source may refer to different persons. Common causes of dirt in data include:

- Synonyms (i.e. Different names for same real life object)
- Homonyms (i.e. The same name for different objects):
- Abbreviations:
- Non-standard Naming of fields:
- Format Differences:

The cleaning tasks consist of:

- Record Duplicate detection (starting with dimension tables)
- Record Duplicate Elimination (only one copy of duplicates in dimension tables)
- Record unification (assigning same warehouse id to duplicates in the fact table).

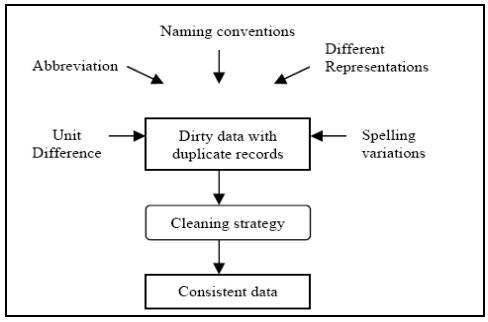

Figure1.Cleaning Model

3.1. Data Cleaning Approaches

Data cleaning is an automated method for examining the data, detecting missing and incorrect values and correcting them. It focuses on eliminating variations in data contents and reducing data redundancy aimed at improving the overall data consistency. In general, data cleaning involves several phases

- *Data analysis*: In order to detect which kinds of errors and inconsistencies are to be removed, a detailed data analysis is required.
- *Definition of transformation workflow and mapping rules*: Depending on the number of data sources, their degree of heterogeneity and the "dirtiness" of the data, a large number of data transformation and cleaning steps may have to be executed. Sometime, a schema translation is used to map sources to a common data model; for data warehouse, typically a relational representation is used. Early data cleaning steps can correct single-source instance problems and prepare the data for integration. Later steps deal with data integration and cleaning multi-source instance problems.
- *Verification*: The correctness and effectiveness of a transformation workflow and transformation definitions should be tested and evaluate. Multiple iterations of the analysis, design and verification steps may be needed.
- *Transformation*: Execution of the transformation steps either by running the extraction, transformation and loading workflow for loading and refreshing data.
- *Backflow of cleaned data:* After errors and duplications are removed, the cleaned data should also replace the dirty data in the original sources.

4. Overview of the Proposed System

4.1. Land Application System

Land application system is the system for applying plot of land. Each interested person allowed buying only

one plot of land. The plot of land owner will be given a 30-year grant. The plot owner is not allowed to build more than one house on a single plot. Interested persons are to apply to the Yangon and Mandalay Town planning and Land Management Department.

4.2. The Proposed System

Land application forms can be applied to Yangon or Mandalay. The applicants are sent to apply many application forms many times from Yangon or Mandalay. Yangon application forms can be stored in Yangon database and Mandalay application forms can be stored in Mandalay database. Then the system merges these two databases into single database. The proposed token-based data cleaning technique first define token and differentiate by cluster formation. Then the system searches duplicate records and eliminate duplicates records. The system will process the whole process according to token-based data cleaning algorithm.

In proposed system Table I (Yangon Application Table) and Table II (Mandalay Application Table) are used to record the application forms. The (Standard Dimensional Table) given as Table III are yet to be cleaned.

5. Preprocessing Stage

5.1 Data Parsing. The process of analyzing a continuous stream of input and information within the document is filtered into the context of the elements in which the information is structured.

5.2 Data Standardization. The process of standardizing the information represented in certain fields with some special content. Different data representation in different data sources are converted to a uniform representation before the record matching. Sometimes, we may use some tables to resolve data conflicts to standardize records in databases.

5.3 Data Transformation. This process refers to simple conversions that can be applied to the data in order to confirm to the data types of their corresponding domains.

In proposed system, there are two data sources: Yangon database and Mandalay database. In Yangon database, there are twelve fields such as ID, NRC, Sex, ContactAddress, Occupation, Name, DateOfBirth, PhoneNumber, Citizen, AppliedDate, Email and Comment. In Mandalay database, there are eleven fields such as ApplicantID, NRC, Gender, PermanentAddress, Rank, ApplicantName, BirthDate, PhoneNumber, Citizen and AppliedDate. These databases differ by the number of columns and field names. Therefore, we first convert field name and number of columns in these databases according to the Standard database and standard table before merging these databases.

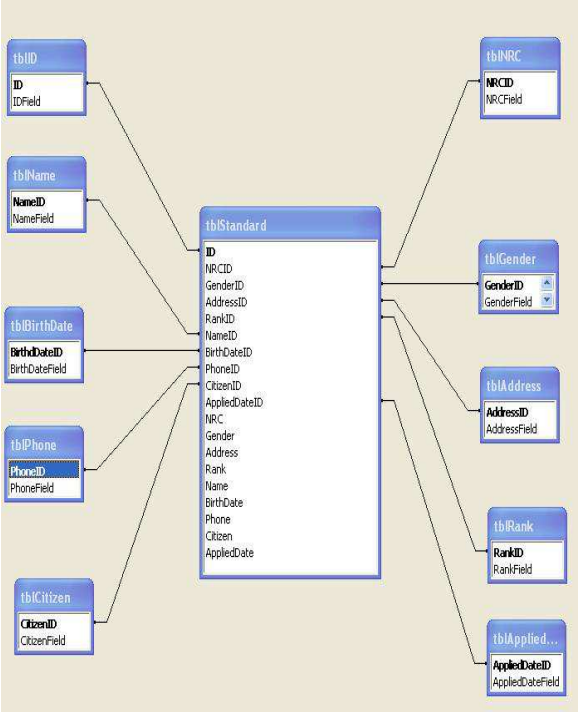

Figure2.Database Design for Standard Database

5.4 Dirt in the Standard Dimensional Table

Two levels of dirt exist in Table III, namely field level dirt and record level dirt. The field level dirt is the dirt that occurs when each field in a record is isolation. For example "PhoneNumber" and "NRC" fields have format differences. The National Registration Card (NRC) of the applicant in row (7) written as"6/LLN (N) 123557" has a different format (6-La La Na (n) 123557) in row (8). Other Field level dirt in Table III include (i) typographical error. Record level dirt is the combination of all the fields' dirt in a given row. Row (3) of Table III is a record with following content (excluding ID) "4/ma ta na (n) 134759, Male, Yangon, Staff, Zaw Win Htut, 12/13/1975". This appears to be same person as row (6) with following content (excluding ID) ("4/MTN (N) 134759, M, Yangon, Staff, Ko Zaw Win Htut, 12/13/75"). An obvious implication of record level dirt is "that duplicates are not easily determined".

6. The Proposed Token-based Data Cleaning Algorithm

The proposed token-based data cleaning algorithm, Token Based cleaner, accepts "dirty" source tables, such as Tables I (Yangon Application Table), II (Mandalay Application Table), III (Standard Dimensional Table) and IV (Tokenized Dimension Table) and returns "clean"

table, such as Table V (cleaned Dimension Table)

Table I. (Yangon Application Table)

ID	NRC	Contact Address	Occupation	Name	Phone Number
Y00001	12/magada(n)126446	No(6),Thiri St, Mingalardon,Yangon	Teacher	Aye	635470
Y00002	1/mkn(N) 178141	No(3), Sinmin St, Sanchaung, Yangon	Engineer	U Tin Tun	01-637412
Y00003	4/ma ta na (n) 134759	No(11), Baho Street, Thamine,Yangon	Staff	Zaw Win Htut	635481
Y00004	12/pabata (Naing) 111345	No.21,Shwe St, Sagaing	Officer	Daw Aye	095197887

Table II. (Mandalay Application Table)

Applicant ID	NRC	Gender	Permanent Address	Rank	Applicant Name	Phone Number
M00001	5/Ya Ba Na (Naing) 123456	F	No(54),12 St, Amarapura	Teacher	Hla Win	665766
M00002	4/MTN(N) 134759	M	No(11), Baho St, Thamine,Yangon	Staff	Ko Zaw Win Htut	01 -635481
M00003	6/LLN(N) 123557	F	No.2,Sagaing st, Hlaing,Yangon	Tutor	Thandar Win	8617880
M00004	6-La La Na (n) 123557	F	No.2,Sagaing st, Hlaing,Yangon	Tutor	Daw Thandar Win	098617880

Basically, the most important field is selected and ranked based on the power to uniquely identify records. The elements in the selected field are tokenized resulting in a table of tokens. Tokens are formed using numeric values, alphanumeric values and alphabetic values. Tokens are used to determine whether the records are match or not. If there are some matching records, eliminate the duplicate record. Warehouse id (WID) is generated for cleaned records. The steps in the algorithm are described below.

Step 1: Select field based on record identifying abilities: The user familiar with application domain is expected to select and rank the most important field that could be tokenized to perfectly discriminate one record from another. In this paper, there are many fields according to land application form such as: Name, Birth Date, NRC, Gender and Address, Phone Number. Applicants can have same birth date. Same name can be possible. Moreover, applicants are family members and they have same address. However, NRC field cannot be duplicated. Therefore, National Registration Card (NRC) field is selected as the unique field.

Step 2: Token Composition: This step Token – Based cleaner starts token formation with a (Standard Dimensional Table) Table III. The first step is to compose tokens from the most uniquely identifying fields for a given record. In this paper, NRC is used as tokens. The separator in NRC is not important for token-based data cleaning technique. Whether the NRC number is written like “12/ykn(n) 741852”, “12\ykn (n) 741852” or “12-ykn(n) 741852”. The consideration for that NRC number is 12/ykn (n) 741852. Therefore delimiters for that NRC number are “-”,”\” .When the system sees the delimiters defined by the developers in the NRC number, it will replace standard format. NRC tokens are as shown in Table IV.

Tokens can be defined three types of tokens, as follows.

(i) Numeric Tokens: These tokens comprise only digits (0,1,2,3,4,5,6,7,8,9) and are obtained from numeric-dominated fields such as "Date of Birth", "Telephone numbers" ,etc.

(ii) Alphabetic Tokens: Tokens in this category consists only of alphabets (a-z A-Z) and are obtained from fields consisting mainly of alphabets, e.g., "Name", "Father Name", etc. To form the alphabetic token, the first character of each word in the field is obtained, and the token is made up of these characters.

(iii) Alphanumeric Tokens: These tokens comprise both numeric and alphabetic tokens and could be obtained from fields containing both numbers and strings, e.g., "NRC", "address", etc. A function is defined, which (1) decomposes a given alphanumeric element into its constituent members, (2) scans through the set of members and selects only tokens that are either numeric or alphanumeric part to its numeric and alphabetic parts ,and (3) define the set of tokens to get an alphanumeric token key.

Table III. (Standard Dimensional Table)

ID	NRC	Address	Rank	Name	Phone Number
Y00001	12/magada (n)126446	No(6),Thiri St, Mingalardon, Yangon	Teacher	Aye	635470
Y00002	1/mkn(N) 178141	No(3), Sinmin St, Sanchaung, Yangon	Engineer	U Tin Tun	01-637412
Y00003	4/ma ta na (n) 134759	No(11), Baho Street, Thamine,Yangon	Staff	Zaw Win Htut	635481
Y00004	12/pabata (Naing) 111345	No.21,Shwe St, Sagaing	Officer	Daw Aye	095197887
M00001	5/Ya Ba Na (Naing) 23456	No(54),12 St, Amarapura	Teacher	Hla Win	665766
M00002	4/MTN(N) 134759	No(11), Baho St, Thamine,Yangon	Staff	Ko Zaw Win Htut	01 -635481
M00003	6/LLN(N) 123557	No.2,Sagaing st, Hlaing,Yangon	Tutor	Thandar Win	8617880
M00004	6-LaLaNa (n) 123557	No.2,Sagaing st, Hlaing,Yangon	Tutor	Daw Thandar Win	10/10/1980

Table IV. (Tokenized Dimension Table)

ID	NRC	Name	Address	Token
Y00001	12/magada(n)126446	Aye	No(6),Thiri St, Mingalardon, Yangon	12/MGD(N)126446
Y00002	1/mkn(N)178141	U Tin Tun	No(3), Sinmin St, Sanchaung, Yangon	1/MKN(N)178141
Y00003	4/ma ta na (n) 134759	Zaw Win Htut	No(11), Baho Street, Thamine,Yangon	4/MTN (N) 134759
Y00004	12/pabata(Naing) 111345	Daw Aye	No.21,Shwe St, Sagaing	12/PBT(N) 111345
M00001	5/Ya Ba Na (Naing) 123456	Hla Win	No(54),12 St, Amarapura	5/YBN(N) 123456
M00002	4/MTN(N) 134759	Ko Zaw Win Htut	No(11), Baho St, Thamine,Yangon	4/MTN(N) 134759
M00003	6/LLN(N) 123557	Thandar Win	No.2,Sagaing st, Hlaing,Yangon	6/LLN(N) 123557
M00004	6-LaLaNa (n) 123557	Daw Thandar Win	No.2,Sagaing st, Hlaing,Yangon	6/LLN (n) 123557

Step 3: Cluster Formation, Define WID: Potentially each record in a data set has to be compared with all the records in the data set. The clustering techniques are used to cluster or group the dataset into small group based on the similar values to reduce the time for the elimination process. The clustering technique will be useful in the elimination process.

Similarity functions can be categorized into two groups: sequence-based similarity functions and token-based similarity functions. Sequence-based similarity functions allow contiguous sequences of mismatched characters. It is defined as the minimum number of insertions, deletions or substitutions necessary to transform one string into another. Sequence-based similarity functions become complicated for larger strings.

Token-base similarity functions can be used as the simplest method than the sequence-based similarity functions. Tokenization is typically performed by treating each individual word of certain minimum length as a separate token or by taking first character from each word. Token has been crated for the selected attributes. Each function measures the similarity of selected attributes with other record fields and assigns a similarity value for each field. The clustering techniques have been selected to group the fields based on the similarity values.

Accurate similarity functions are important for clustering and duplicate detection problem. The matching and non-matching pairs are used to group as cluster and eliminate the duplicates. Generating WID operation performed in this step is expressed as follow:

- If cluster is empty; then generate WID from NRC token t1, for a new applicant
- Else if cluster has only one element, then use WID of this existing applicant
- Else if cluster contains more than one element, then perform related task for eliminating duplication.

Step4: Duplicate Detection and Elimination Duplicate Records: During the elimination process, only one copy of exact duplicated records should be retained and eliminate other duplicated records. The elimination process is very important to produce a cleaned data. This step is used to detect or remove the duplicate records from one cluster or many clusters. In the elimination process, select all possible pairs from each cluster and compare records within the cluster using the selected attributes. The duplicate record detection and elimination are crucial for improving the quality of the extracted data with imprecise and noisy.

In this paper, the system detect the duplication records of Yangon land application forms and Mandalay land application forms by using token shown in table IV. The system compares the tokens of records within the clusters and eliminates the duplicated records. As a result following table V cleaned data can be produced.

Table V. (Cleaned Dimension Table)

ID	NRC	Name	Address	Token
Y00001	12/magada(n)126446	Aye	No(6),Thiri St, Mingalardon, Yangon	12/MGD(N)126446
Y00002	1/mkn(N) 178141	U Tin Tun	No(3), Sinmin St, Sanchaung, Yangon	1/MKN(N)178141
Y00003	4/ma ta na (n) 134759	Zaw Win Htut	No(11), Baho Street, Thamine,Yangon	4/MTN (N) 134759
Y00004	12/pabata (Naing) 111345	Daw Aye	No.21,Shwe St, Sagaing	12/PBT(N) 111345
M00001	5/Ya Ba Na (Naing) 123456	Hla Win	No(54),12 St, Amarapura	5/YBN(N) 123456
M00003	6/LLN(N) 123557	Thandar Win	No.2,Sagaing st, Hlaing,Yangon	6/LLN(N) 123557

7. Performance Evaluation

This section presents some results of the experiments conducted to measure the performance of the proposed token-based algorithm. The performance of proposed system is evaluated according to the following percentages:

- Recall
- False positive error (FP)
- False negative error (FN) and
- Precision

7.1. Recall

This is also known as percentage hits. It is identified as the percentage of duplicate records being correctly identified by the system.

Recall= Number of identified duplicates x100%
Number of actual duplicates

7.2. False Positive Error

This is the percentage of records wrongly identified as duplicates.

FP= Number of wrongly identified duplicates x100%
Total number of identified duplicates

7.3. False Negative Error

False Negative is the percentage of duplicate records that are not detected by the system.

FN=100%-Recall

7.4. Precision

Precision is the percentage of the information reported as relevant by the system that is correct. It can be calculated using the following formula.

Precision =100% - False Positive Error

The performance evaluation of proposed system is indicated in Table VI.

Table VI. (Recall, False Positive Error (FP), False Negative Error (FN) and Precision)

For Total Record (1000)	Recall (%)	FP (%)	FN (%)	Precision (%)
Duplicate Record 200	95	5	5	95
Duplicate Record 420	90	10	10	90
Duplicate Record 736	88	12	12	88

8. Conclusion

The integration of information is an important area in databases. The duplicate elimination problem of detecting database records is an important data cleaning problem. To ensure, high data quality, data warehouse must cleans data by detecting and eliminating the redundant data. This system use token-based data cleaning method to detects and eliminates duplication by using well-defined tokens for record matching. Land application system needs high data quality for administrator. This reduced time consuming using Token-Based cleaner. By cleaning duplicate records in application dataset, it can be used effectively in decision making and can achieve high quality dataset. And the space required for storing cleaned database is saved.

9. Future Work

Other duplicate detection system can reference this proposed system in the future. This paper provides the basic method of data preprocessing. Although several methods of data preprocessing have been developed, data preparation remains an active area of research.

10. References

[1] A.E.Monge, C.P.Elkan, "An Efficient Domain-independent Algorithm for Detecting Approximately Duplicate Database Records," *In Proceedings of the ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining*. Tucson, AZ, 1997.

[2] D.Bitton, D.J DeWitt, "Duplicate Record Elimination in large data files," *ACM Trans Databse System* 8 (2) 1995.

[3] Erhard Rahm and H.Hai Do. Data Cleaning: "Problems and Current Approaches." *IEEE Data Engineering Bulletin, 23(4):3-13*, December 2000.

[4] Heiko Muller, Johann-Christoph Freytag "Problems, Methods, and Challenges in Comprehensive Data Cleansing" Hum boldt University, 10099, Berlin, Germany. **(4)**

[5] Mong Li Lee, Hongjun Lu, Tok Wang Ling, Yee Teng Ko, "Cleansing data for mining and warehousing,"*In Proceedings of the 10th International Conference on Database and Expert System Applications* (DEXA99), 1999.

[6] M Hernandex and S.Stolfo, "The merge, purge problem for large databases." *In proceedings of the ACM SIGMOD*, pages 127-138, San Jose, CA, 1995.

[7] Matthew A Jaro "Advances Record Linkage Methodology as Applied to Matching the Census of Tampa Florida" *Journal of the American Statistical Association*.

[8] M Chochinwala, S Dalal, A K Elmagarmid V S Verykios, "Record Matching Past, Present and Future", Applied Research Division Telcordia Technologies, Computer Sciences Department, Purdue University, College of Information Science and Technology, Drexel University.

[9] M Hernandez and S.Stolfo, J.S.: "Real World Data is Dirty: Data Cleansing and The Merge/Purge Problem", *Journal of Data Mining and Knowledge Discovery*, No.2, 1998,pp. 9-37.

[10]Peng Wang (a1111318), "Data Cleaning Services for Distributed Biological Data Sets, "School of Computer Science, University of Adelaide.

[11] Htike.Thin Thin, "An Association Rule Based Approach to Duplicate Detection in Bibliographical Records." *In Proceedings of the 3rd International Conference on Computer Applications* (ICCNA), 2005.

[12] Wai Lup Low, Mong Li Lee, Tok Wang Ling, "A knowledge based approach for duplicate elimination in data cleaning." School of Computing, National University of Singapore, 3 Science Drive 2, 117543.