ATTRIBUTE CORRECTION-DATA CLEANING USING ASSOCIATION RULE AND CLUSTERING METHODS

R.KAVITHA KUMAR[1] *and* DR. RM.CHADRASEKARAN[2]

[1]Department of Computer science and Engineering, Pondicherry Engineering college, Pondicherry
rkavithakumar@pec.edu
[2] Annamalai University, Chidrabram, India
aurmc@sify.com

ABSTRACT
Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases,. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. Data cleaning is the necessary condition of knowledge discovery and data warehouse building. In this paper two algorithms are designed using data mining technique to correct the attribute without external reference. One is Context-dependent attribute correction and another is Context-independent attribute correction.

KEYWORDS
Data Cleaning, Pre-processing, Attribute correction, Missing data, Clustering.

I. INTRODUCTION

Data cleaning is deeply domain-specific. Data quality problems are quite trivial, complex and inconsistent. There is no international common standard for reference. So the process of data cleaning is vary from domain to domain but basically a process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g. taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking and correction of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules, and conventions [1]. The principle of data cleaning is to find and rectify the errors and inconsistencies.

II. PROBLEMS WITH DATA

Missing data occur two ways

- data are expected but are absent
- data are appropriately not available or inapplicable in the real world. Detection of the missing data is often relatively straightforward.

DOI : 10.5121/ijdkp.2011.1202

Erroneous data occur

- when an incorrect value is recorded for a real world value. Detection can be quite difficult. (E.g. the incorrect spelling of a name)

Duplicated data occur in two ways

- repeat records, perhaps with some values different
- different identifications of the same real world entity. Repeat records are common and usually easy to detect. The different identification of the same real world entities can be a very hard problem to detect. (E.g In medical domain 'heart attack" and "cardic arrest" having same meaning)

Heterogeneities arise

when data from different systems are brought together in one analysis.

Two of the problems are

- structural heterogeneity which arises when the data structures reflect different business usage
- semantic heterogeneity which arises when the meaning of data is different in each system that is being combined

These heterogeneities can be very difficult to resolve because they usually involve quite a lot of contextual knowledge that is not defined as metadata.

III. DATA QUALITY

High quality data needs to pass a set of quality criteria. Those include:

- ***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

The quality of data is often evaluated to determine usability and to establish the processes necessary for improving data quality. Data quality may be measured objectively or subjectively. Data quality is a state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use. The hierarchy of data quality

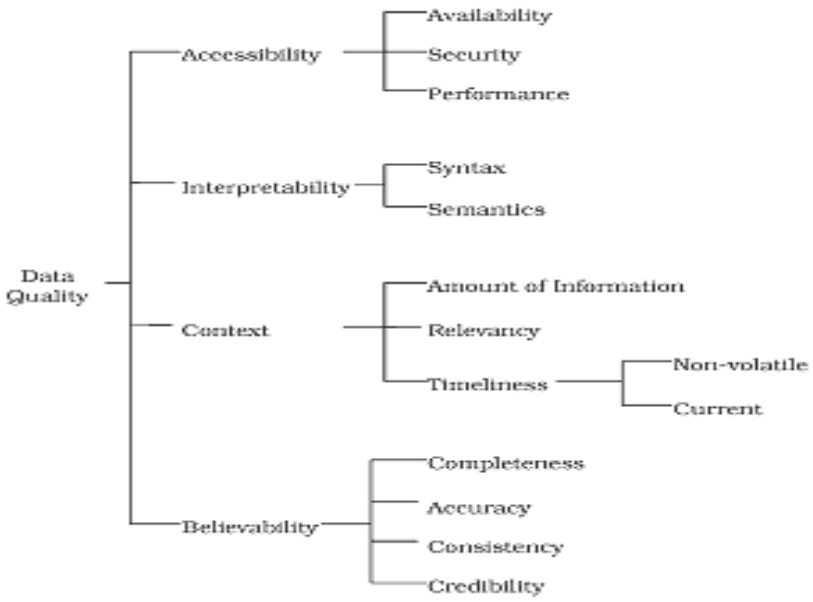

Figure 1. Data quality hierarchy

Data quality has two distinct aspects: one is the "correctness" of data (such as accuracy and consistency), and the other involves the appropriateness of data for some intended purposes. Data producers and users generally assume that the purpose of data quality assurance is to provide the best data possible. However, this obscures the need to evaluate data. The implication is that if a data set is the best available and is as good as it can be made, then there are no other options than to use it. In this case, there is no point in worrying about just how good it can be made. The flaw in this is that merely saying that a data set is as good as it can be made does not tell us how good it is or whether it is any good at all. What may be considered good data in one case may not be sufficient in another case. For example, an analysis of the financial position of a firm may require data in units of thousands of dollars while an audit requires precision to the cent. Therefore, the term "data quality" may best be defined as "fit to use," which implies the quality of the data in the warehouses is accurate enough, timely enough, and consistent enough for the organization to make reasonable decisions

IV. RELATED LITERATURE

Related researches are as follows [5, 6]:
• To design efficient algorithms for abnormal detection to avoid the traversal of huge data set
• To insert human judgment between automated abnormal detection and cleaning process to prevent error in data processing
• To parallelize data cleaning and data set file Processing
• To eliminate the duplicate data brought in by data combination
• To build a universal domain-independent data cleaning framework

• To realize schema integration especially, methods to solve the data abnormal detection problems are as follows [7]: statistical method based on Chebyshev's theorem which selects samples randomly to analyze and gets a greater speed; pattern recognition method which is based on data mining and algorithm of machine learning to search abnormal data with association rule algorithm; aggregation algorithm based on distance which takes Euclid Edit distance as the criteria to judge class so as to find the duplicate records in data set; increment method which employ random methods to get tuples and input a random tuple stream if data source supports.
Duplicate detection in data cleansing frameworks for fuzzy duplicates *is* pertaining two or more tuples that describe the same real-world entity using different syntaxes. Eliminating fuzzy duplicates is applicable in any database but is critical in data-integration and analytical processing domains, which involve data warehouses, data mining applications, and decision support systems [9]

Several recent approaches incorporate machine learning into the duplicate detection process. Tejada et al. use a decision tree forest to learn both duplicate detection rules and weights for string transformations, which are used for comparing fields. The string edit distance is a metric commonly used in duplicate detection procedures. Bilenko and Mooney have shown that machine-learning techniques increase the accuracy of the field-matching task when string edit distance is used and in some cases even when token-based measures are used [10]
Record linkage follows a probabilistic approach [11, 12]. For each record pair, a comparison vector is produced by comparing corresponding attribute values. The record pairs are classified as matched, possibly matched, and unmatched using a linkage rule that assigns each observed comparison vector with a probability for each class. To reduce the number of comparisons, application specific blocking criteria can be used.

V. DATA QUALITY PROBLEMS

According to [2,7] data quality issues may ne divided into two main categories

i. Issues regarding data coming from one source
ii. Issues regarding data coming from multiple source

Both main categories may be further divided into subcategories

i. Data quality issues on instance
ii. Data quality issues on the record level

In Fig -2 show the categories of the data quality

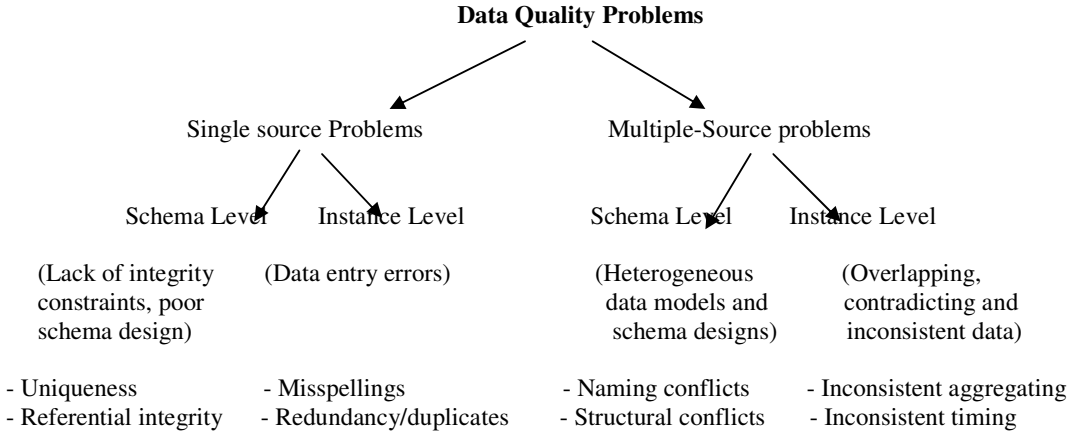

Figure – 2 Categories of data quality

VI .FRAMEWORK FOR DATA CLEANING

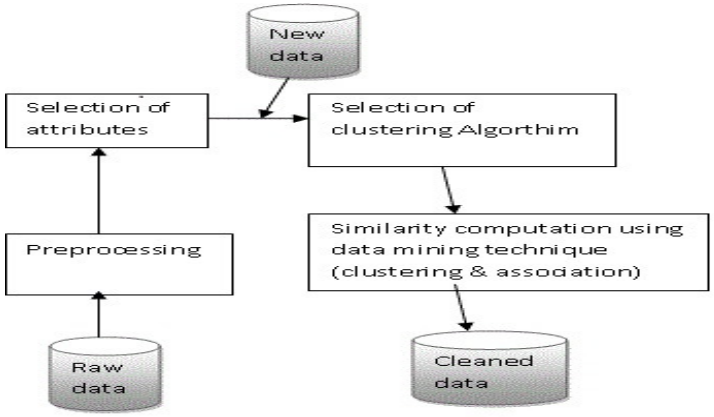

Figure 3. Framework for data cleaning

Each steps in the framework is well suited for the different purposes . Framework in Figure 3 is suitable for all purpose . Some of the data cleaning techniques will be suited for the particular work of the data cleaning process. This framework offers the user interaction by selecting the suitable algorithm. The user has to know each step clearly. This framework will be effective in handling noisy data.

The principles of the framework are

1. From the raw data process the data using Pre-processing technique to select the attribute
2. Clustering algorithm or blocking method is selected to group the records based on the clustering key.
3. Similarity computation using data mining technique i.e. Clustering is used correct the Context independent Incorrect the Context –Dependent attribute. Detail discussion about this is in next chapters.

VII. ATTRIBUTE CORRECTION USING DATA MINING TECHNIQUES

In this paper we focused on the attribute correction. In attribute correction we used two methods of data mining i. Clustering techniques for Context-independent correction and ii. Associations rule for Context-dependent correction.

a. Association Rules Methodology -Context-Dependent Correction

Context-dependent means that attribute values are corrected with regard not only to the reference data values it is most similar to but also takes into consideration values of other attributes within in a given record.

Algorithm

In this algorithm we use association rules methodology to discover validation rules for data sets. To generate frequent item sets Apriori[14] algorithm is utilized.

Two parameters is used

i. Minsup is defined analogically to the parameter of the same name for the Apriori algorithm used here.
ii. Distthresh is the minimum distance between the value of the "suspicious" attribute and the proposed value being a successor rule it violates in order to make correction.

Levenshtein distance (LD) is a measure of the similarity between two strings, which we will refer to as the source string (s) and the target string (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t. For example,

- If s is "test" and t is "test", then LD(s,t) = 0, because no transformations are needed. The strings are already identical.
- If s is "test" and t is "tent", then LD(s,t) = 1 , because one substitution (change "s" to "n") is sufficient to transform s into t.

The following is the modified Levenshtein distance

$$Lev(s_1, s_2) = 1/2.(Lev(s_1, s_2)/|s_1| + Lev(s_1, s_2)/|s_2|)$$

Where Lev(s1,s2) denotes Levenshtein distance between strings s1 1ns s2. The modified Distance for strings may be interpreted as an average fraction of one string that has to be modified to be transformed into the others. For instance, The LD between "Articulation" and "Articaulation" is 2. In modified Levenshtein distance is 0.25. The modification was introduced to be independent of the string length during the comparison.

The algorithm is follows

- Generate all the association rules from the sets generated in the previous step. The rules generated may have 1, 2 or 3 predecessors and only one successor. The association rules generated form the set of validation rules.
- The algorithm discovers records whose attribute values are the predecessors of the rules generated with an attributes whose value is different from the successor of a given rule.
- The value of the attribute for a "suspicious" in a row is compare with all the successors.
- If the relative Levenshtein distance is lower the threshold distance the value may be corrected. If there are more values within the accepted range of the parameter, a value most similar to the value of the record is chosen.

b. Clustering Technique - Context-independent Correction

Context-independent correction means that all the record attributes are examined and cleaned in isolation without regard to values of others attributes of a given record. The main idea behind this algorithm is based on an observation that in most data sets there is a certain number of values having large number of occurrences within the data sets and a very large number of attributes with a very low number of occurrences. Therefore, the most -representative values may be the source of reference data. The values with low number of occurrences are noise or misspelled instance of the reference data.

The same Levenshtein is used in these methods which is discussed in the previous algorithm . Two parameter is used here

i. Distthresh – being the minimum distance between two values allowing them to be marked as similar and relation

ii. Occrel – used to determine whether both compared values belong to the reference data set.

The algorithms is follows

- First cleaning process, first convert all attributes convert from lower case to upper case. all the values is cleaned and set is calculated
- Each value is assigned to separate cluster. The cluster element with highest occurrence is treated as cluster representative. Depends upon the clustering representative sort in descending order.
- Starting from first cluster compare all the cluster and also calculate the distance between the cluster using the modified Levenshtein distance.
- If the distance is lower that the distthresh parameter and the ration of occurrences of cluster representative is greater or equal the occrel parameter the cluster
- After all the clusters are compared , the clusters are examined whether they contain values having distance between them and the cluster representative above the threshold value .if so, they are removed from the cluster and added to the cluster list as separate clusters.
- Repeat the same step until there are no changes in the cluster list i.e.no clusters are merged and no cluster are created. The cluster representative is from reference data set and the cluster define transformation rule for a given cluster values should be replace with the value of the cluster representative.

As far as the reference dictionary is concerned, it may happen that it will contain values where the number occurrences are very small. These values may be marked as noise and trimmed in order to preserve the compactness of the dictionary.

VIII. RESULTS

a. *Context Dependent Attribute correction*

The Algorithm was tested using the sample Cardiology dataset which is from Hungarian data.The rule-generation part of the algorithm is performed on the whole data set. The Attribute correction part was performed on random sample.

Following measures are used for the correctness algorithm.

P_c – Percentage of correctly altered values
P_i – Percentage of incorrectly altered values
P_o- Percentage of values marked during the review as incorrect, but not altered during cleaning

The measure as defined

$P_c = n_c / n_a * 100$
$P_i = n_i / n_a * 100$
$P_0 = n_{00} / n_0 * 100$

n_c - correctly altered values
n_i - number of incorrectly altered values
n_0 - total number of altered values
n_{00}- the number of elements initially marked as incorrect .

From Table – I observed that the relationship between the measures and the distthresh parameter. The result show that the number of values marked as incorrect and altered is growing with the increase of the distthresh parameter. This also proves that the context-dependent algorithm perform better to identifying incorrect entries. The number of incorrectly altered values is growing with increase of the parameter. However , a value of the distthresh parameter can be identified that gives optimal results. i.e. the number of correctly altered values is high and the number of incorrectly altered values is low.

Table – I Dependency between the measures and the parameter for Context-dependent algorithm

Distthresh	P_c	P_i	P_o
0	0.0	0.0	100.0
0.1	90	10	73.68
0.2	68.24	31.76	46.62
0.3	31.7	68.3	36.09
0.4	17.26	82.74	33.83
0.5	11.84	88.16	31.33
0.6	10.2	89.8	31.08
0.7	9.38	90.62	30.33
0.8	8.6	91.4	28.82
0.9	8.18	91.82	27.32
1.0	7.77	92.23	17.79

b. Context-Independent Attribute correction

The Algorithm was tested using the sample Cardiology dataset which is from Hungarian data There are about 44000 records divided into 11 batches of 4 thousand records. The attribute CP (Chest pain type) in that Angial is one of the types which is occurs when an area of your heart muscle doesn't get enough oxygen-rich blood. During process 4.22% i.e. 1856 element of whole set were identified as incorrect and hence subject to alteration. Table –II contains the example transformation rules discovered during the execution.

Table I Example Transformation Rules

Original value	Correct value
Angail	Angial
Anchail	Angial
Angal	Angial
Ancail	Angial

The measure is used
P_c – Percentage of correctly altered values
P_i – Percentage of incorrectly altered values
P_o- Percentage of values marked during the review as incorrect, but not Altered during cleaning

Table III – Dependency between the measures and the parameter for Context -Independent algorithm

Distthresh	P_c	P_i	P_o
0	0.0	0.0	100.0
0.1	92.63	7.37	92.45
0.2	79.52	20.48	36.96
0.3	67.56	32.44	29.25
0.4	47.23	52.77	26.93
0.5	29.34	70.66	23.41
0.6	17.36	82.64	19.04
0.7	7.96	92.04	8.92
0.8	4.17	95.83	1.11
0.9	1.17	98.83	0.94
1.0	0.78	99.22	0

The algorithm display better performance for long strings as short string would require higher value of the parameter to discover a correct reference value. High values of the distthresh parameter results in larger number of incorrectly altered elements. This method produces as 92% of correctly altered elements which is an acceptable value. The range of the application of this method is limited to elements that can be standardized for which reference data may exist. Conversely, using this method for cleaning last names could end with a failure.

IX. CONCLUSION AND FUTURE WORK

From the above work that shows the attribute correction is possible without external reference data and can give good results using data mining technique. From the experimental result it is observed that the result of context-dependent attribute correction is better than the context-independent attribute correction in terms of distthresh-parameter. In context-dependent method it is observed that the relationship between the measures and the distthresh parameter and the number of values marked as incorrect and altered is growing with the increase of the distthresh parameter. It is better to identify incorrect values. In this work only one string matching distance was used. It is possible that other functions could result in better output and this should be explored in future work and also applies other data mining technique for data cleaning.

REFERENCES

[1] Chapman, A. D. “Principles and Methods of Data Cleaning – Primary Species and Species- Occurrence Data “, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. 2005

[2] Lukasz Ciszak, “ Application of clustering and Association Methods in data cleaning”, 978-83-60810-14-9/08, 2008 IEEE

[3] Paul Jermyn, Maurice Dixon and Brian J Read, “Preparing clean views of data for data minig”,

[4] Huang Yu, Zhang Xiao-yi, Yuan Zhen , Jiang Guo-quan , “A Universal Data Cleaning Framework Based on User Model “, 2009 ISECS

[5] Martin J. Eppler, René Algesheimer, Marcus D31.33impfel. "Quality Criteria of Content-Driven Websites and their Influence on Customer Satisfaction and Loyalty: an Empirical Test of an Information Quality Framework". the 8th International Conference on Information Quality (IQ 2003), November 7-9, 2003: 108-120

[6] KDnuggets Polls. "Data Preparation Part in Data Mining Projects", Sep 30-Oct 12, 2003. http://www.kdnuggets.com/polls/2003/data_preparation.htm.

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

[8] Mauricio Hernandez, Salvatore Stolfo, "Real World Data Is Dirty: Data Cleansing and The Merge/Purge Problem", Journal of Data Mining and Knowledge Discovery, 1(2), 1998.

[9] H.H. Shahri; S.H. Shahri, Eliminating Duplicates in
Information Integration: An Adaptive, Extensible Framework, Intelligent Systems, IEEE, Volume 21, Issue 5, Sept.-Oct. 2006 Page(s):63 – 71

[10] M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable string similarity measures". ACM SIGKDD, 39-48, 2003.

[11] W. E. Winkler. Matching and record linkage. In Business Survey Methods. Wiley-Interscience, 1995.

[12] M. G. Elfeky, V. S. Verykios, and A. K. Elmagarmid. TAILOR: A record linkage toolbox. In Proceedings of the International Conference on Data Engineering (ICDE), pages 17–28, 2002.

[13] YAN HAN, DIAO Xing-chun. "The design and Implementation of Data cleaning Knowledge Modeling", International symposium on knowledge Acquisition and modeling , 2008

[14] J.WEBB, "Association Rules", in the Handbook of Data Mining , ISBN-139780805855630, 724pp, pp25-39

AUTHORS PROFILE

R. Kavitha Kumar Working as a programmer in Department of Computer Science and Engineering , Pondicherry Engineering College., Puducherry, India. Has 11 years teaching experience and 3 years Inducstry Experience as a Programmer. Did M.Phil (Computer Science) Degree in Mother Thersa Women's University ., Kodaikanal, India. M.Sc (Computer Science) in Bharathidasan University, Trichy, India. B.Sc (Computer Science) in Madras University, India. At present pursing Ph.D (Compter Science) in Mother Tersa Women's University ., Kodaikanal, India.

Dr.RM.Chandrasekaran Working as Professor in Annamalai University , Chidabaram, India., has 18 years teaching experience, 3 years worked as a Registrar ,in Trichy Anna University, Trichy . 2 years Worked as software consultant in USA. Area of interest and guiding in Data Mining, Image Mining, Software Slicing, Document Image Segmentation