Volume 4, No. 1, January 2013

Journal of Global Research in Computer Science

ISSN-2229-371X

RESEARCH PAPER

Available Online at www.jgrcs.info

INFORMATION CLEANUP FORMULATION: PRAGMATIC SOLUTION

Qamar Rayees Khan[1]
Department of Computer Science, BGSB University, Rajouri (J&K)
E-mail: qamarrayees@rediffmail.com

Muheet Ahmed Butt[2]
Department of Computer Science, University of Kashmir, Hazratbal (J&K)
E-mail:ermuheet@gmail.com

Majid Zaman[3]
Directorate of Information Technology & Support Systems, University of Kashmir, (Hazratbal) (J&K)
Email: zamanmajid@rediffmail.com

M. Asger[4]
Department of Computer Science, BGSB University, Rajouri (J&K)
E-mail:m_asger@rediffmail.com

Abstract: The magnitude of information accessible that we all transact with is growing rapidly. With this growth information and the current component of small and big data sets, conformity of information essential. Information cleanup also know as information cleanup or scrubbing trades with detection and removal of errors and inconsistencies from information in state to change the quality of data, improve. The effective methodology for error detection goes beyond wholeness reasoning are analyze in the proposed research. The practice includes: applied mathematics, pattern matching, clustering, and data mining techniques.

Keywords: data cleansing, data cleaning, data quality, error detection.

INTRODUCTION

To truly quantify attributes of information is still challenge at large, it relies upon numerous issues which ascertain quality of information [1][32][33]. The information acquiring or information entry operation play very crucial part in guarantee choice of information since it comprises source of data which is being saved into system. Good deal of work is being done to decrease faults due to the information entry procedure exploitation definite establishment regulation which wholly depends on kind of information which is being pushed into system. Unless an organization ensures measures in perseverance to avoid data errors and inconsistencies tract mistake rates are still typically around 5% [2][32][33]. For existing information sets, the coherent solution to this difficulty is to undertake cleanse of information in some way or the other. Course of study, for whatever real world data set, doing this job automatic is reasonably impossible. Enterprise which transact with tremendous information spends millions of dollars every year to detect information errors [3][32][33]. Data cleansing process which is done non-automatic is hard, time consuming, and is inclined to errors. The obligation for priceless and standardized tools that alter or to a great degree aid in information cleansing process is must. The tools proposed should be very much price effective and should also guarantee that quality is delivered. This may seem to be an evident answer but precise small elementary investigation has been immediately consciously at methods to activity such tools. In this paper, study and reassessment of the disagreeing method aspect of information cleaning are carried out. A broad model of the information improvement operation and a set of broad mechanisms are immediate that can be used to address the problem. Eventually, future directions with regards to investigation to reference the information purifying job in the real world are deliberated.

AN OVERVIEW OF DATA CLEANSING

Relatively new research area, procedure is computationally costly on precisely large collection sets hence it is nearly unfeasible to do with old engineering. With the state of art infrastructure available allows performing information cleanup process having a minimal time complexity and guarantee choice results, still there are many difficulties in information cleanup that engineers are working on. There is usually no agreed explanation of information cleanup,

various explanations look on peculiar topic in which procedure is applied, thus no universal explanation is agree upon. The leading topic that includes information cleanup as part of their shaping activity are: data warehousing, knowledge discovery in databases (also termed data mining), and information quality management. Within the data warehousing field, information cleanup is practical particularly with heterogeneous information sources, records associated with same entity are depicted in various formats in different collection sets or are depicted erroneously-resulting in duplicate records in merged(central) database, issue is to eliminate these duplicate records, this problem is known as the *merge/purge problem* [8][32][33].

Information cleanup is also defined as procedure to get rid of errors and inconsistencies in collection, and solving object identity problem [9]. A researcher primarily dealt upon information cleanup as merge/purge problem and proposes basic grouped-vicinity method to solve it. Total Information Quality Management (TDQM) is an area of interest both within the research and business communities. The information cognitive content and its integration in the entire information business process are handled from different points of view in literature (e.g., [11, 12, and 13]). The broadest study of research in this area is accessible in [14]. Unfortunately, none of mentioned papers mention explicitly or implicitly to the information cleanup problem as such. Some of the researchers are interested only with process management issues from data quality perspective, others with explanation of information quality only. The latter category is of interest to this research.

In proposed theoretical account of information life cycles with practical application to attribute [15][32][33] the information acquiring and collection usage cycles incorporate a series of activities: assessment, analysis, adjustment, and discarding of information, although it is not generally addressed in the paper, if one merged the collection cleansing procedure with information life cycles, this ordering of track would specify it in the proposed model from the information choice perspective. In the same model of collection choice, Fox [16] proposes four choice magnitude of the information: accuracy, correctness, completeness, and consistency. The rightness of information is outlined in terms of these dimensions. Again, a simple effort to specify the information cleansing procedure within the model would be the procedure that measure the rightness of information and better its quality. More lately, aggregation cleansing is regarded as a first step, or a reprocessing measure, in the KDD process [17, 18],[32][33]. Respective KDD and Data Mining scheme execute information cleansing activities in a very sphere specific fashion. In [19] discovering of informative form is utilized to execute one type of collection cleansing by covert *garbage patterns* – insignificant or illegal patterns. Machine learning methods are used to use the information cleansing procedure in the backhand quality categorization problem. In [20] information cleansing is definite as the procedures that instrumentality computerizes methods of analyze databases, detection missing and incorrect data, and correcting errors. The Recon Data Mining scheme is used to assist the quality expert to place a series of error types in financial data systems [32] [33].

DATA CLEANSING METHODS

Keeping all the above facts in view, information cleanup must be viewed as a procedure. This procedure may be tied directly to information acquiring and explanation or it may be practical after the information, to better collection choice in an present system. The following three phases define a information cleanup procedure:

- Specify and ascertain the different error types
- Lookup and determine the various error instances
- Correct the exposed different errors

All of preceding mentioned phases constitute a complex problem in itself. A wide variety of specialized methods and technologies can be applied to each phase. The focus of the research here is on the first two aspects of this generic framework. The later aspect is very difficult to automate outside of a strict and well defined domain. Many of the aforementioned data cleansing tools utilize integrity analysis to locate data errors [32] [33].

Many database systems (e.g., Oracle, SQL Server) support this type of data cleansing to some extent. It is the duty of database administrator to resolve which data cleansing technique is used but does not possess the knowledge of the domain and/or clearly defined approach to correctly carry out this task. There are tools, usable by non-database experts, which support this type of cleansing. Errors that involve relationships between one or more fields are often very difficult to uncover. These types of errors require deeper inspection and analysis in a much detailed manner. One can view this as a problem in outlier detection way as for example if a large data elements say 100 where in 99 conform to a general form then; the remaining 1 data element is likely to have error. These data elements are considered outliers. Two things are done here, identifying outliers or strange variations in a data set and identifying trends in data. Knowing what data is supposed to look like allows errors to be uncovered and resolved. But, the fact of the matter is that real world data and information often are very diverse and rarely conforms to any standard statistical distribution [32][33]. This is especially acute when data is viewed in several dimensions. Therefore, more than one method for outlier detection is often necessary to capture most of the outliers. Below is a set of general methods that can be utilized for error detection:

Statistical: Identifying the outlier fields and records using the values of mean, standard deviation, range or any other statistical method, based on Chebyshev's theorem [21], considering the confidence intervals for each field [22][32][33].

- **Clustering**: Identifying outlier records using clustering based on Euclidian (or other) distance. Existing clustering algorithms provide little support for

identifying outliers [23].However, in some cases clustering the entire record space can reveal outliers that are not identified at the field level inspection [22][32][33]. The main drawback of this method is computational time. The clustering algorithms have high computational complexity. For large record spaces and large number of records, the run time of the clustering algorithms is prohibitive.

- **Pattern-based**: Identify outlier fields and records that do not conform to existing patterns in the data. Combined techniques (partitioning, classification, and clustering) are used to identify patterns that apply to most records [32] [33]. A pattern is defined by a group of records that have similar characteristics ("behavior") for p% of the fields in the data set, where p is a user-defined value (usually above 90).

APPLIED METHODS FOR CLEANSING

Each of the above mentioned methods was implemented at University of Kashmir, Srinagar J&K, and India. Each method was tried using a data set consist of real world data supplied by the Examination Wing, University of Kashmir. It is mentioned here that the University declares around 900 results for around 3.5 lakh students appearing in around 450 courses at under and post graduate level. The data set represents part of the student information system which includes its registration and examination records. A subset of 40000 records with 130 fields of the same type from this data set is used to demonstrate the methods. The size of the data and the type of the data elements allowed fast and multiple runs without reducing much the generality of the proposed methods. The goal of this experimentation is to prove that these methods can be successfully implemented so as to identify outliers that represent potential errors. The executions were intentional to work on larger collection sets and without large amounts of domain cognition. The only information needed from the user is the size of the data set and the values of some threshold parameters [32][33].

CLUSTER METHOD

A combined cluster method was enforced based on the group-average clustering algorithm [25] reckon the geometrical distance between records. The clustering algorithmic rule was run n times correcting the maximum size of the clusters. The main aim was to determine as outliers at most those records that were identified before as containing outlier values. However, procedure time proscribe multiple runs in an every-day concern usage, on larger information sets as the data sets could be already utilized for assorted read or write function. After several executions on the same data set, it inverted out that the bigger the commencement value for the maximal distance allowed between clusters that are to be merged together the better the outlier sensing[32][33].

A faster clustering algorithmic rule could be utilized that may allow automatic tuning of the maximal cluster size, as well as scalability to bigger data sets. Also, using some domain cognition, an "important" mathematical space can be elect to guide the clustering, to trim the size of the data. The method acting can be used to cut down the search space for different method. The known clusters combine together line that bear certain sameness. Therefore, it is highly likely that the records in a clustering would follow a certain form. The test data set has a particular distinctive: many of the data conditions are empty. This specialness of the data set does not make the know-how less unspecific, but allowed the definition of a new similarity step that relies on this feature. Here, strings of cardinal and ones, referred to as *Hamming value* [26], are subordinate with each record. Each twine has as many conditions as the number of fields. A "1" in the string correspond a non-empty field on the same place in the record as the 1 in the string. A "0" in the string stand for an bare field on the aforesaid place in the record as the 0 in the string. The Hamming distance [26] is utilized to clustering the records into groups of similar records. Initially, clusters having zero Hamming distance between records were identified [32][33].

PATTERN-BASED DETECTION METHOD

Patterns are known in the information accordant to the arrangement of the line per each field. For each tract the line are gregarious using the geometrical spacing and the k-mean algorithmic rule [27]. The six protrusive conditions are not indiscriminately chosen, but at equal spacing from the average, one of them being an blank field. A pattern is definite by a large grouping of line that clustering the aforesaid way for fewest of the fields. Each clustering is categorized accordant to the amount of capacity unit it comprise (i.e., cluster number 1 has the largest size and so on)[32][33].

CONCLUSIONS

A well definite conceptualization were offering and practical which reference the difficulty of involuntary identification of mistakes in statistics sets and the conforming results showed that some of the approaches could be effectively pragmatic to real-world data, while others need fine-tuned and improvement. Each of the preset conceptualization has property and failing. Unsuitably, little basic enquiry within the material systems and computer science groups has been showed which is straight related to error find and data cleansing. Few in-depth appraisals of data cleansing acting have been published. Some intense effort by the database and information systems groups is needed to reference this problem [32] [33].

REFERENCES

[1] English, J., "Plain English on Data Quality," DM Review, Webzine, Date Accessed: 02/10/99, http://www.dmreview.com, 1999.

[2] Orr, K., "Data Quality and Systems Theory," *CACM*, vol. 41, no. 2, February 1998, pp. 66-71.

[3] Redman, T., *Data Quality for the Information Age*, Artech House, 1996.

[4] Ballou, D. and Tayi, K., "Methodology for Allocating Resources for Data Quality Enhancement," *CACM*, vol. 32, no. 3, 1989, pp. 320-329.

[5] Centrus, "Qualitative Marketing Software Centrus Merge/Purge Module," Centrus, Webpage, Date Accessed: 01/15/2000, http://www.qmsoft.com/solutions/Merge.htm, 2000.

[6] EDD, "Home page of DataCleanser tool," EDD, Webpage, Date Accessed: 01/15/2000, http://www.npsa.com/edd/, 2000.

[7] Flanagan, T. and Safdie, E., "A Practical Guide to Achieving Enterprise Data Quality," Webpage, Date Accessed: 12/01/99, http://www.techguide.com/, 1999.

[8] Moss, L., "Data Cleansing: A Dichotomy of Data Warehousing," *DM Review* February 1998.\

[9] Galhardas, H., Florescu, D., Shasha, D., and Simon, E., "An Extensible Framework for Data Cleaning," Institute National de Recherche en Informatique ét en Automatique, Technical Report 1999.

[10] Hernandez, M. A. and Stolfo, J. S., "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem," *Journal of Data Mining and Knowledge Discovery*, vol. 2, 1998, pp. 9-37.

[11] Strong, D., Yang, L., and Wang, R., "Data Quality in Context," *CACM*, vol. 40, no. 5, May 1997, pp. 103-110.

[12] Svanks, M., "Integrity Analysis: Methods for Automating Data Quality Assurance," *EDP Auditors Foundation*, vol. 30, no. 10, 1984, pp. 595-605.

[13] Wang, R., Strong, D., and Guarascio, L., "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, Spring 1996, pp. 5-34.

[14] Wang, R., Storey, V., and Firth, C., "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4, August 1995, pp. 623-639.

[15] Levitin, A. and Redman, T., "A Model of the data (life) cycles with application to quality," *Information and Software Technology*, vol. 35, no. 4, 1995, pp. 217-223.

[16] Fox, C., Levitin, A., and Redman, T., "The notion of Data and Its Quality Dimensions," *Information Processing and Management*, vol. 30, no. 1, 1994, pp. 9-19.

[17] Brachman, R. J. and Anand, T., "The Process of Knowledge Discovery in Databases: A Human-Centered Approach," in *Advances in Knowledge Discovery and Data Mining*, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurasamy, R., Eds., MIT Press/AAAI Press, 1996, pp. 97-58.

[18] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., "From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining*, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurasamy, R., Eds., MIT Press/AAAI Press, 1996, pp. 1-36.

[19] Guyon, I., Matic, N., and Vapnik, V., "Discovering Information Patterns and Data Cleaning," in *Advances in Knowledge Discovery and Data Mining*, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurasamy, R., Eds., MIT Press/AAAI Press,1996, pp. 181-203.

[20] Simoudis, E., Livezey, B., and Kerber, R., "Using Recon for Data Cleaning," in Proceedings of KDD, 1995, pp. 282-287.

[21] Bock, R. K. and Krischer, W., *the Data Analysis Briefbook*, Springer, 1998.

[22] Johnson, R. A. and Wichern, D. W., *Applied Multivariate Statistical Analysis*, 4th ed., Prentice Hall, 1998.

[23] Knorr, E. M. and Ng, R. T., "A Unified Notion of Outliers: Properties and Computation," in Proceedings of KDD 97, 1997, pp. 219-222.

[24] Barnett, V. and Lewis, T., *Outliers in Statistical Data*, John Wiley and Sons, 1994.

[25] Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B. T., and Liu, X., "Learning approaches for Detecting and Tracking News Events," *IEEE Intelligent Systems*, vol. 14, no. 4, July/August 1999.

[26] Hamming, R. W., *Coding and Information Theory*, New Jersey, Prentice-Hall, 1980.

[27] Kaufman, L. and Rousseauw, P. J., *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[28] Agrawal, R., Imielinski, T., and Swami, A., "Mining Association rules between Sets of Items in Large Databases," in Proceedings of ACM SIGMOD International Conference on Management of Data, Washington D.C., May 1993, pp. 207-216.

[29] Srikant, R., Vu, Q., and Agrawal, R., "Mining Association Rules with Item Constraints," in Proceedings of SIGMOD International Conference on Management of Data, Montreal, Canada, June 1996, pp. 1-12.

[30] Korn, F., Labrinidis, A., Yannis, K., and Faloustsos, C., "Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining," in Proceedings of 24th VLDB Conference, New York, 1998, pp. 582--593.

[31] Marcus, A. and Maletic, J. I., "Utilizing Association Rules for the Identification of Errors in Data," The University of Memphis, Division of Computer Science, Memphis, Technical Report TR-14-2000, May 2000.

[32] Jonathan I. Maletic1 and Andrian Marcus, "Data Cleansing: A Prelude to Knowledge Discovery".

[33] Jonathan I. Maletic, Andrian Marcus, "Data Cleansing: Beyond Integrity Analysis"