A Framework for Automatic Data Maintenance In Business Applications

Mr. Badugu Ranjith Kumar[1], Mr. Vishnu Prasad Goranthala[2]
Asst.Prof Dept of CSE
Balaji Institute of Technology & Science Narsampet, Telangana, India.

Abstract:
The crisis of detecting errors, different patterns, duplication, and misbehavior are known as Anomaly detection. The problem based on applications is anomaly detection, and it is considered as one of the leading research. Different non-conforming patterns, aberrations, peculiarities or exceptions in a range of application domains are often referred to as anomalies. The other aspects from these are the duplicate records, error, misbehavior based data which have been treated as anomalies. Anomaly detection was normally used in online applications such as banking, credit card fraud, insurance, and healthcare. In recent database management system meets issues every day due to increased volume of data and it is a difficult problem for database administrators. Digital library, e-commerce records have a different data structure and different schemes. It makes a little response in terms of time, relevancy, security and poor quality in probing and managing the huge amount of data.

Keywords: Anomalies, e-commerce, Digital library novel framework

1 Introduction:

For keeping repositories with quality data, this situation needs a solution after reducing "dirty" data like replicas, identification errors, and different patterns for same data. It also affects the speed or presentation of DBMS and this kind of issue can rectify by providing a quality data improved with managed database. Here in this research, the main goal is to supply a novel framework to handling duplicates, removing the error, correcting the error, aligning data in application specific domains. The contribution of the novel framework is:

- Analyzing the data, before and after deployment.
- Verify the attributes of each data entity before deployment and persist a log for future verification.
- Check the input data with link and without a link.
- Make and compare data index each time of data manipulation for finding duplicates.

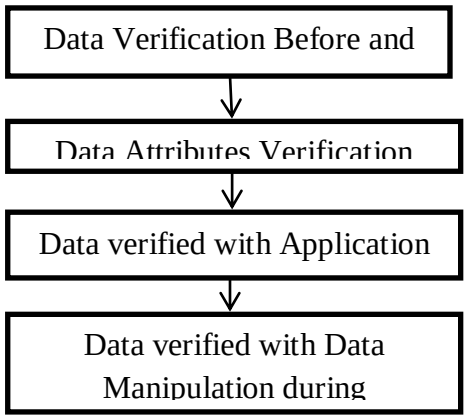
2 Creation of Novel Framework

In the earlier, any problem in data requires more manual work, and it is a tedious task. Testing is a development that can be applied before and after deploy the software in the real time industry. Testing of Regression is one of the ways which detects the problems early in the application's process. One of the stable solutions is clearing the data before used in the application. Data cleaning concept is used to reducing the duplicate, removing error and aligning the data in a correct manner such as how the application is going to access the data.

It is provoked to afford a most general, application specific approach by comparing with the existing approaches discussed in the literature review, for de-duplication, error-free data by verifying the data attributes and log file of the data shaped during data creation. For providing a better solution in this research, a novel framework is presented here, where it cares about all the necessary functions to provide a quality data for the application. The entire architecture of the proposed framework is shown in figure 4.2. The main objective of this research is to provide an automatic data cleaning method using a novel framework presented here. The automatic tasks assigned in the framework can sense anomalies and improve the quality of the data which is going to use the application software. The simulation based experiment has been obtained from DOTNET software, and the routine metrics have been compared with the existing systems result.

Proposed System Model

The entire work is divided into four phases such as (i). Data verification is before and after the creation of the data on the contribution of the novel framework (ii). Attributes of the data have been confirmed according to the application where the data is going to be used (iii). Data have been verified while getting manipulated during application execution (iv). Data verified while getting manipulated during application execution and these phases is shown in figure 4.1 clearly



Process Flow Diagram

During the application execution, data is connected either in connection based or not connection based. This connection based and connection-less based are available features integrated with the RDBMS "SQL-Server." The input data perfectness has been verified by alignment of the data, size, attributes, and duplicate. Once the investigation report says that the data is faultless, and then the data will be served by the specific application and persisted, else the data will be eliminated from the data storage. The perfectness has been carried in four stages. The data verification is before and after deploying the application.

Experimental Datasets

In this Paper, a set of real-world datasets is experimented and verified the performance of the proposed framework. The datasets used for an experiment during the entire framework is given in the following table 1. It gives the information about dataset name, number data in the dataset and the main attributes used for the test. For attribute verification, the main attributes used in this experiment. With the input datasets given the application, the attributes of the synthetic datasets have been verified.

The data manipulation has been handled based on the data, in the application such as data insertion, editing data structure, editing data, updating data, deleting data and searching data. The DBMS itself verifies the data size, data type, alignment, ID of the data and data format during these data manipulations.

It is well known that if any contradiction occurs in the data, there will be DBMS error created automatically by the DBMS software. However, the chief aim is to provide quality application software which can do data cleaning by itself in each stage of the process.

Number	Dataset Name	Number of data	Number of Duplicate Records	Number of Error Records	Attributes
1	Cora Bibliography	1295	95	10	Author-id, author name, year, title, venue, the number of pages, volume, and DOI
2	Restaurant	864	112	13	Name, address, city, state, specialty
3	Synthetic dataset generator	32	5	5	Name, surname, address, locality, street, city, pin code, phone number, personal ID
4.	User Defined	100	10	10	Name, Acc No, address, city, state, balance, pin code, personal ID

3. Results and Discussion

The synthetic dataset is the banking dataset which is user defined and also created by the user. For evaluating the performance of the framework by experimenting the algorithm and entire process of the other datasets used are real-time benchmarking datasets. There is a possibility of changing manually as error data among some of the data in the banking. Remaining data are duplicating by the user itself to confirm the duplication and error correction. Error created by changing the data size and data design dynamically.

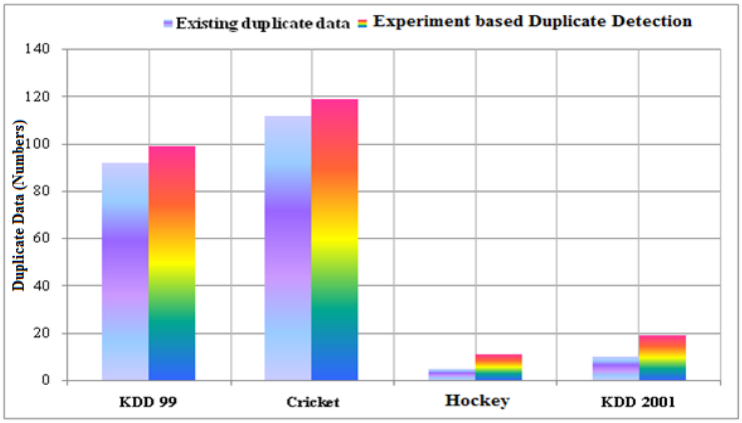


Figure 1.Comparison of Duplicate Data Detection

\For dynamic comparison, some of the data was perfectly checked and considered as the good data and these complete datasets are used as a benchmark data by the user. It happens in parallel. During the application based data manipulations, the result of the framework is examined by checking the duplication verification, error identification, and error correction functions. All the findings are applied using the user proof pair based comparison in this experiment.

The research-based data cleaning has been given in the following figure 1 and figure 2.. From the experimental results, it has been gathered that total number present data versus many duplicate data; error data have been sensed without human intervention by the software which has been compared with the manual results. There are four datasets are used for the experiment, 90, 115, 5 and 10 are the number of additional data exist in the KDD 99, Cricket, Hockey and KDD 2001 correspondingly. Out of this, 100, 120, 10 and 19 are all identified as replica data automatically by the framework.

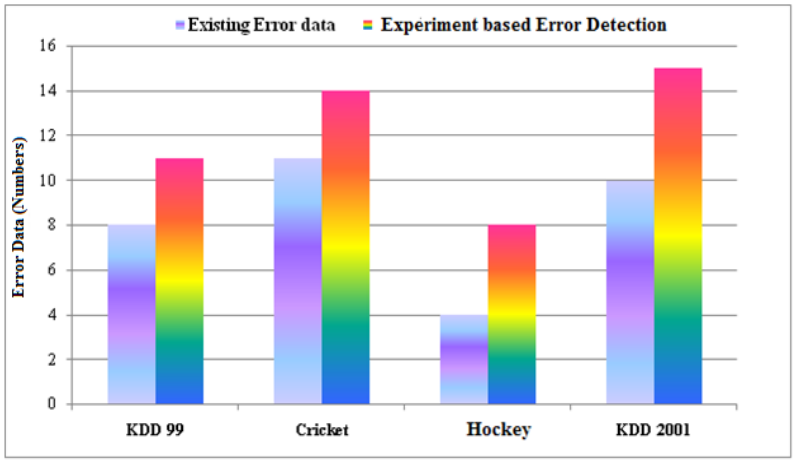


Figure 2.Comparison of Error Data Detection

This manual duplication versus automatic duplicate detection has been shown in1. In the four datasets, the number of existing error data is 8, 11, 4 and 10. Out of this 11, 14, 8 and 15 number of data is detected as error data constantly in the experiment by the framework and it is shown in 2.. One of the major objectives of this research is to detect the error data and correct the errors automatically by the framework to improve the quality of the application software.

It is also confirmed in this experiment using the framework, and the result has been shown in figure 3. From figure 3, it is clear that the number detected error data are 10, 13, 5 and 10 for KDD 99, Cricket, Hockey and KDD 2001 respectively. Out of this, 3, 4, 3 and 5 are the number of error data are rectified automatically by the framework in KDD 99, Cricket, Hockey and KDD 2001 correspondingly. From this, it has been completed that nearly 95% to 98.5% of the error data has been corrected by comparing with the benchmark datasets.

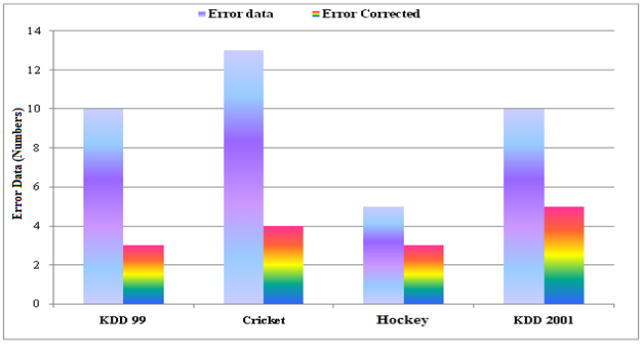


Figure 3. Error Data vs. Corrected Data

Also, another purpose of this research is to de-duplicate the data without disturbing a new data having less percentage of variation than the present data. In this experiment, it has been also confirmed that some data are de-duplicated with and without removing the data eternally. In the four datasets KDD 99, Cricket, Hockey and KDD 2001, 90, 10, 2 and 9 number of data is duplicate data. Out of this 5, 100, 3 and 5 number of data is de-duplicated by comparing with the existing benchmark data entirely in block level.

At the same time, 90, 10, 2 and 9 number of data is removed permanently from the dataset and those data are not able to include after successful changes on the data. This de-duplication of the dataset is shown in figure 4. From given figure 1to figure 4, it is evident, and it is completed that the novel framework can do data cleaning based anomaly detection automatically then the present approaches discussed in the literature survey.

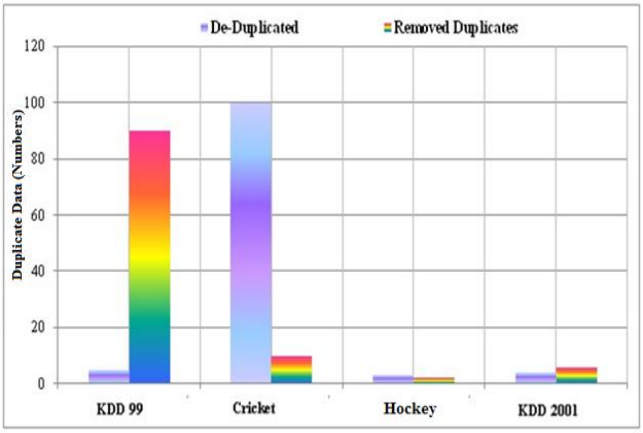


Figure 4. Data De-Duplicated vs. Duplicate Data Deleted

4. CONCLUSION:

Here in this paper, the main purpose is to provide an error free and non-duplicated data to get better in the application software and its performance to the present data computing industry. For this, a novel framework is offered here with automatic detection and rectification of duplicate, error, and alignment based processes. The experiment is handled on four different datasets with the most number of similar attributes to assess the performance. The experiment is handled in .DOTNET software and the results are established. From the obtained results, it is clear that the proposed framework is better for improving the quality of data in the application software. In future, the performance of the framework is confirmed with more number of systems connected to cloud surroundings and client-server technique.

5. REFERENCE

[1]. Ahamed, B. B., & Hariharan, S. (2011). A survey on distributed data mining process via grid. International Journal of Database Theory and Application, 4(3), 77-90.

[2]. Nodine, Marian "Active Information Gathering in Infosleuth TM", International Journal of Cooperative Information Systems, pp.3-27, 2000.

[3]. Ahamed, B. B., & Yuvaraj, D. (2018, October). Framework for Faction of Data in Social Network Using Link Based Mining Process. In International Conference on Intelligent Computing & Optimization (pp. 300-309). Springer, Cham

[4]. Patcha, A. and Park, J.M. "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends", The International Journal of Computer and Telecommunications Networking, Computer Networks, Vol.51, pp.3448-3470, 2007.

[5]. Phua, C., Alahakoon, D. and Lee, V. "Minority Report in Fraud Detection: Classification of Skewed Data", ACM SIGKDD Explorations Newsletter Special Issue on Learning from Imbalanced Datasets, Vol.6, pp.50-59, 2004.

[6]. Poesia, Gabriel and Cerf, Loic "A Lossless Data Reduction for Mining Constrained Patterns in n-ary Relations", Machine Learning and Knowledge Discovery in Databases, Springer- Verlag, Vol.8725, pp.581-596, 2014.

[7]. Qingwei, Ye., Dongxing, Wu., Yu, Zhou and Xiaodong, Wang "The Duplicated of Partial Content Detection based on PSO", IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications, pp.350-353, 2010.

[8]. Quinlan, S. and Dorward, S. "Venti: A New Approach to Archival Storage", Conference on File and Storage Technologies (FAST '02), Bell Labs, Lucent Technologies, pp.89-101, 2002.

[9]. S bnis, Vikrant and Khare, Neelu "An Adaptive Iterative PCA-SVM Based Technique for Dimensionality Reduction to Support Fast Mining of Leukemia Data", Soringer Professional., 2012.

[10]. Sebyala, A.A., Olukemi, T. and Sacks, L. "Active Platform Security through Intrusion Detection using Naive Bayesian Network for Anomaly Detection", In London Communications Symposium, 2002.

[11]. Ahamed, B. B., & Ramkumar, T. (2015). Deduce User Search Progression with Feedback Session. Advances in Systems Science and Applications, 15(4), 366-383.

[12]. Siqueira, ThiagoLuís Lopes., Ciferri, Cristina Dutra de Aguiar., Times, Valéria Cesário., Oliveira, Anjolina Grisi de and Ciferri, Ricardo Rodrigues "The Impact of Spatial Data eRdundancy on SOLAP Query Performance", Journal of the Brazilian Computer Society, Vol.15, pp.19-34, June 2009.

[13]. Snyder, D. "Online Intrusion Detection using Sequences of System Calls", Master of Science, Department of Computer Science, Florida State University, Spring, 2001.

[14]. Song, X., Wu, M., Jermaine, C. and Ranka, S. "Conditional Anomaly Detection", IEEE Transactions on Knowledge and Data Engineering, Vol.19, pp.631-645, 2007.

[15]. Spence, C., Parra, L. and Sajda, P. "Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model", In IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, IEEE Computer Society, 2001.