Methods and techniques to evaluate the performance of Data Cleansing Algorithms for very Large Database Systems

R. Deepa. Dr. R Manicka Chezian

Ph.D. Scholar, Department of computer Science (Aided), NGM College Pollachi. Associate professor, Department of Computer Science(Aided), Pollachi

Abstract: -The data cleansing algorithm has a key role in this competitive environment as for decision making considered the system requires more precise information. . Yet the inconsistency in the data submitted makes it difficult to aggregate data and analyze results which may lead to delay or data compromises in the reporting of results. This paper gives a detailed view on different algorithms which is used for cleansing very large dataset to get for the need for more consistent data. The need of cleansing algorithms is to increase the quality of dataset as well as it reduces the computational cost after filtering and ignoring the outliers. This paper also presents some methods and techniques to evaluate and master the performance of the data cleansing algorithms used for the very large database systems.

Keywords: Algorithms,Dataset, cleanse, Computational Cost, outliers.

Introduction

Data mining is the process of database analysis that attempts to discover useful information's from a large dataset. The analysis uses several advanced statistical and other methods, like cluster analysis, and sometimes it uses artificial intelligence or neural network techniques. A major objective of data mining is to find previously hidden relationships between the data, especially when the data is collected from different databases. Data mining is used in several areas like insurance, banking, retail, astronomy, medicine detection of criminals and terrorists. The process of converting data to knowledge has several phases that is shown in the figure -1

DATA CLEANSING

Data cleansing or data scrubbing is the process of finding and rectifying improper or wrong information or records from a large record set, collection of records (table), or from collection of tables (database). This concept is mainly is used in databases, the term data cleansing refers to finding incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and subject to replacing, modifying, or deleting this outlier data or coarse data. After data cleansing, the data set will be consistent for analyzing or any other operations with other similar data sets in the system. The inconsistency data should be analyzed and detected or removed from the original database. The process diagram is shown for the data cleansing is shown in the figure-2.

ISSN: 2278 – 1323

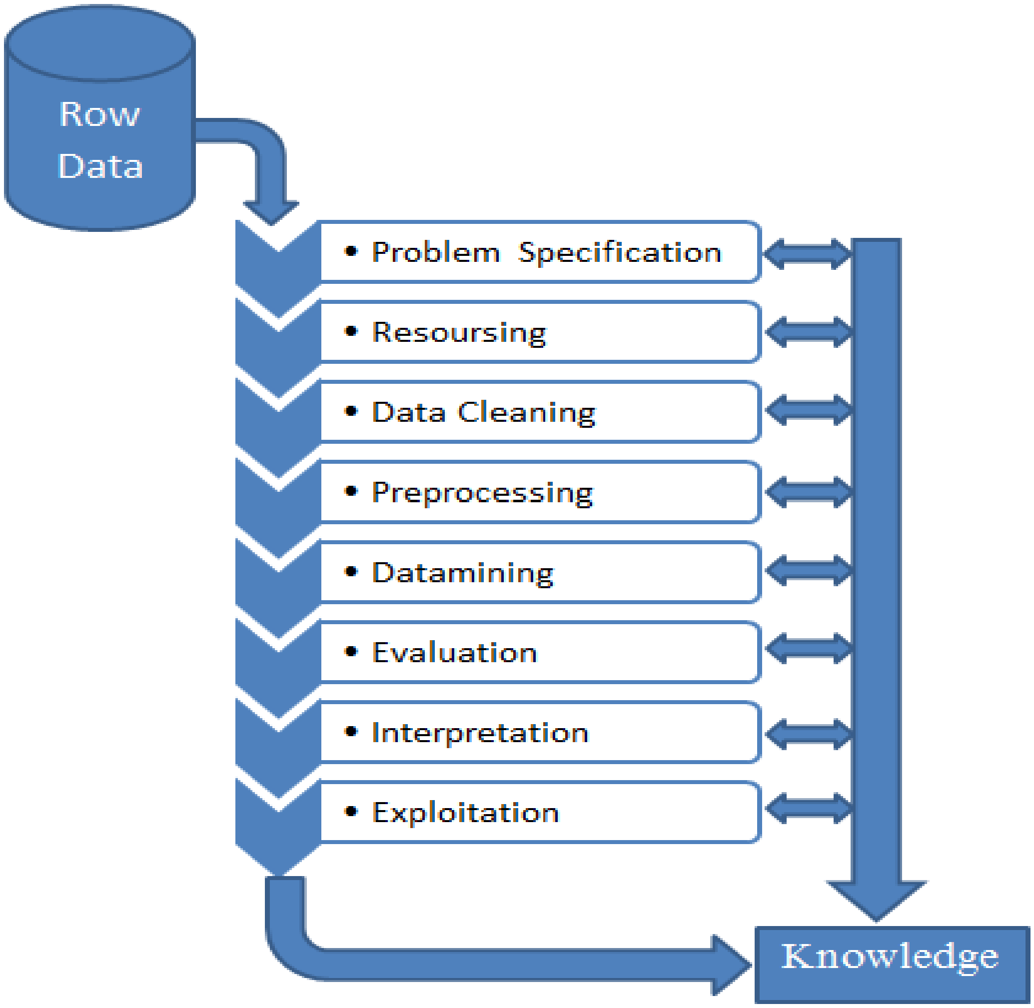

Figure -1 Process Diagram of Data mining

DATA CLEANING APPROACHES

In general, data cleaning has several phases

Data analysis: The first phase of data cleansing is data analysis as the data is collected from the heterogeneous background or from different data sets or databases. So the possibility of errors or bugs is high in order to detect and remove such errors and inconsistencies in the data set a detailed data analysis is required. In addition the analyzed data must be subjected to manual verification or inspection with the data or data samples and some sort of analysis programs should be used to gain metadata about the data properties and find the data quality problems.

ISSN: 2278 – 1323

Figure -2Process diagram of Data Cleansing

Definition of transformation workflow and mappingRules: This phase is purely depends upon the number of data sources. To fulfill this phase large number of transformation or cleaning steps may have to be involved. This is purely based on their degree of heterogeneity and the "dirtyness" of the data. Sometime, a schema translation is used to chart sources to a common data model. Typically a relational representation is used for data ware houses. Early data cleaning steps can correct single-source instance problems and prepare the data for integration. Later steps deal with schema/data integration and cleaning multi-source instance problems, The schema related data transformations as well as the cleaning steps should be specified by a declarative high query and mapping language as much as possible, to enable automatic generation of the transformation code. In addition, it should be possible to embed the user written cleaning code and special purpose tools during a data transformation workflow. The transformation steps may request user feedback on data instances for which they have no built-in cleaning logic.

Verification: The correctness and effectiveness of a transformation workflow and the transformation definitions should be tested and evaluated both manually and algorithmically, e.g., on a sample or copy of the source data, to improve the definitions if necessary. It requires Multiple iterations of the analysis in order to improve quality of the iterated or analyzed dataset, in addition design and verification steps may be needed, e.g., since some errors found only after applying some kind of transformations techniques.

Transformation: Execution of the transformation steps is maintained by either

ISSN: 2278 – 1323

running the ETL workflow for loading and refreshing a data warehouse or when answering queries on datasets or multiple tables.

Backflow of cleaned data: After each error are removed from the dataset, the cleaned data shouldalso replace the outlier or dirty or unwanted data in the original sources in order to give legacyapplications the improved data and to avoid repetition of cleaning work for future data extractions. For datawarehousing, the cleaned data is available from the data staging area.

EMPIRICAL REVIEW

An effective cleansing method with low computational cost using association rule mining is achieved by Wejje Wei at. Al[1]. Another novel method using applied brain and vision is proposed [2], which is used to cleanse the ECG data. Chaudhuri at. Al [4] introduced a textual cleansing algorithm; the experimentation is done on the dummy a set of bibliographic references. A novel learning-based algorithm is designed to reduce the web pages by cleansing the information by Yiqun Liu at. Al[3]. For cleansing long string dataset a novel approach is proposed by C.I.Ezeife [5]. Another method is designed and proposed by exploiting statistical relationship of records in a database[7]. Another optimization method to solve the issue in the use of picturing for data mining is proposed by Yu Qian, Kang and Zhang [6]. A mathematical morphology based cleansing algorithm is designed by utilizing possibility of frequent noise that occurs and deteriorates [8]. An alternative method by identifying the identical records in a dataset is proposed by Kazi Shah Nawaz Ripon et. Al [9][10].Context-dependent attribute based detection and correction and Context-independent attribute based detection and correction is proposed by R.Kavitakumar at.al[11]. Forest based technique [12] and sampling methods to identify the potential buyers. This method has two phases: data cleaning and classification, both methods is purely based on random forest. Another novel methodology to cleanse World Wide Web is a monolithic repository. They emphasize on the Web Usage and content Mining process and exploits in the area of data cleaning [14]. Li Zhao, Sung Sam Yuan, Sun Peng and Ling Tok Wang They propose [15] a method based on based on the longest common subsequence. Aye T.T[16] has explained the data cleaning algorithm eliminates inconsistent or unwanted or dirty items in the preprocessed data. an extended tree-like knowledge base and proposed a novel knowledge base data cleaning algorithm is proposed by Yan Cai-rong, Sun Gui-ning, GaoNiangao[17]. A new method is proposed using statistical method for detecting missing element and bugs automatically [18].parsing based cleansing method is proposed by Mohammad, H.H. Shawn R. Jeffery, Minos Garofalakis, Michel J. Franklin [19] has proposed SMURF method, this is the first declarative, adaptive smoothing filter algorithm for effective RFID data cleaning. A novel cleansing method is implemented for dirty data identification and data correction for both normal and no normal multivariate dataset[20]. A new outlier detection engine by combining an FD discovery technique with an existing outlier detection technique and this optimization called "Selective Value". This leads to decrease the number of identified FDs[21]. A study is done on cleansing algorithms for very large datasets.

PARAMETERS AND METHODS TO MEASURE THE PERFORMANCE OF DATA CLEANSING ALGORITHM.

Depending on the nature of the application there are various criteria to measure the performance of a data cleansing algorithm. When measuring the performance, the main concern would be the accuracy in data cleansing. The time efficiency is another factor and amount of data loss is also considered.

ISSN: 2278 – 1323

Definition-1

Assumes that U ={x/x ε N ,x can be any data} where U is not ϕ. C={x/xε N ,x is non-redundant based on some criteria's} in other words x_i is not same as x_0 to x_{i-1} and x_{i+1} to $x_{n(c)-1}$. Then the C is a finite set with no redundant data with some criteria. The criteria purely based on the requirements. Where$C \subset U$. so C □(shown in figure -9) represents the dirty data in U. so data cleaning is the process of transforming U to C. but in real time when transforming there are four possibilities. Figure-3 and figure-4 represents the data set U and C Respectively.

- ✓ Data loss but no outlier
- ✓ Outlier but no data loss
- ✓ Data loss and outlier
- ✓ No outlier and no data loss
- ✓ Full data lsoss

So B is the data set after transformation. Then B={x/xε N ,x is may be one of the above possibilities}

Data loss but no outlier *– if* $B \subset C \subset \mathrm{U}$, *where*$B \leftrightarrow C$ *and* $n(C) > n(B)$then *DL>0* where DL represents the data loss that is shown in figure -5

Outlier but no data loss *-if* $C \subset B$, *where*$B \leftrightarrow C$ *and* $n(C) < n(B)$then *DL=0*.This is shown in figure -6

Data loss and outlier *-if* $C \subset U$ *and* $B \subset U$*where*$B \leftrightarrow C$ *and* $n(c \cap \mathrm{B}) > 0$then *DL>0* this is shown in figure-7

No outlier and no data loss*if* $B \subset C$ *and* $C \subset B$ *where*$B = C$ *and* $(c \cap \mathrm{B}) = (c \cup \mathrm{B})$ *so* $n(B) = n(C)$then *DL>0* where DL represents the data loss that is shown in figure -8

Full data loss *–if* ***B and C are distingt set or*** $(c \cap \mathrm{B}) = \{\}$ *then accuracy is Zero. This criteria is shown in figure-10*

Figure-3 data set U(which contain dirty data)

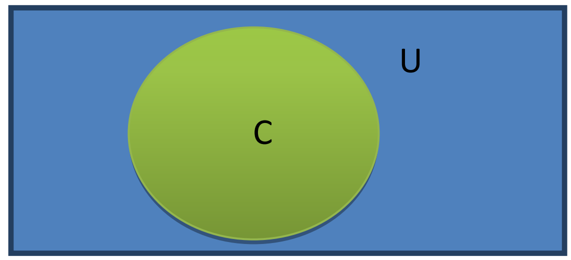

Figure-4 data set C (which contain Actual data)

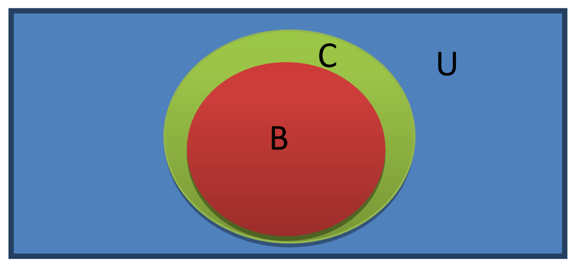

Figure-5data set B (Data loss but no outlier)

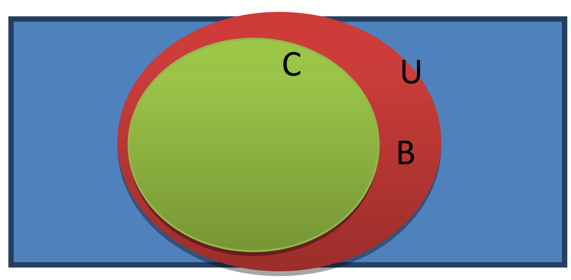

Figure-6data set B (Outlier but no data loss)

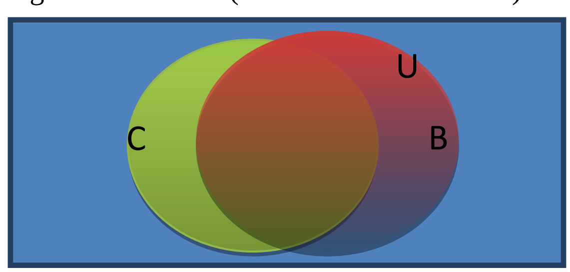

Figure-7data set B (Data loss and outlier)

ISSN: 2278 – 1323

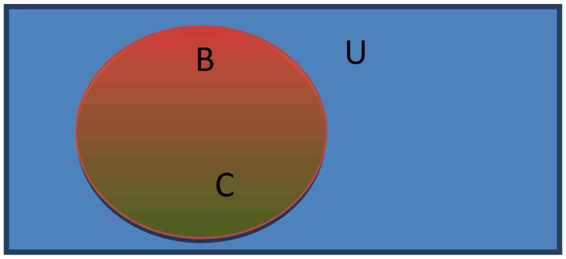

Figure-8 data set B (Data loss and outlier)

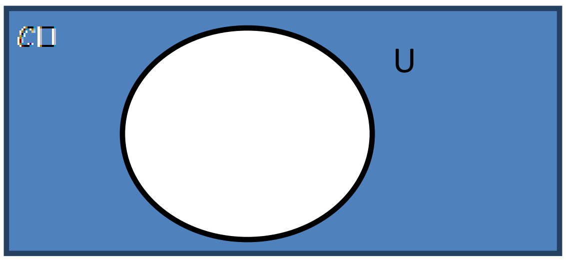

Figure-9Data set C□ (represents the dirty or outlier data)

ACCURACY OF DATA CLEANSING

When the U is transformed to B then the accuracy is calculated based on data lose, identical elements in both C and B so the accuracy is calculated as follows.

$$AC = \frac{n(C \cap \mathrm{B})}{n(C)}$$

Lemma -1 if C and B are disjoint sets then the performance of data cleansing are poor.

Proof: if $B \subset C \; and \; C \subset \mathrm{U}$ but $(c \cap \mathrm{B}) = \{ \;\; \}$ then $B \subset C$ □ so B has only outlier value then the DL =100% and AC=0%. Shown in figure-10

Lemma-2 if $B \subset C$ then this indicates the data lose

Lemma-3 B=C then AC=100%

Proof:if$(c \cap \mathrm{B}) = (c \cup \mathrm{B})$ where $n(B) = n(C)$ then C □ and B □ are identical so C and B also identical then the DL=0% so the efficiency is 100%.

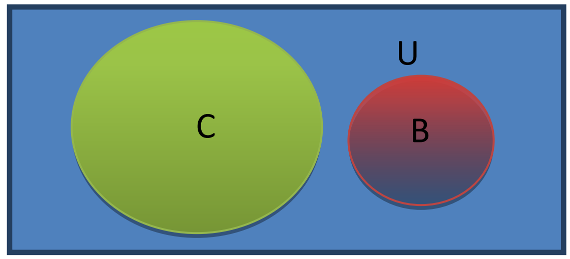

Figure-10 B and C are disjoint

DATA LOSE

Data lose is calculated as follows

$$DL = \frac{n(C - B)}{n(C)}$$

OUTLIER PERCENTAGE

Outlier percentage is calculated as follows

$$\frac{n(C \;\square)}{n(U)}$$

COMPUTATIONAL COST

All the above methods evaluate the effectiveness of Cleansing algorithm. There are some other methods to evaluate the performance of cleansing algorithms are Compression time or computational complexity. Time taken for the transforming U to B should be considered to check the efficiency based on the time. If the time required for cleansing time of an algorithm is less or acceptable level, it implies that the algorithm is acceptable with respective to the time factor. With the development of high speed computer accessories this factor may give very small values and those may depend on the performance of computers, where $t(\mathrm{U})$ represent time taken to transform U to B. In real, the time taken for a process is not constant during in all execution, and the average is not a correct term to represent the time taken. It always lies between the minimum time required $Min(t)$ for a process and maximum time $Max(t)$ requires for a process.

ISSN: 2278 – 1323

Table -1 Data set U (to be cleansed)

Name	place	Pin
CHELLI	Kuruvankandi	643233
THRUPAATHI	Kookkampalayam	645633
RANKAMA	Cheramankandi	702344
ABHINAYA	MelePrappanthara	643721
CHELLII	Kuruvankandi	643233
ABHINAYA	MelePrappanthara	643721
TRUPATHI	Kookkampalayam	645633
TRUPATHIE	Kookkampalayam	645633
RANKAMA	Cheramankandi	702344
MANUPRIYA	ThazhePrappanthara	643271
RANKAN	Plamaram	643833
RANKAMAA	Cheramankandi	702344
MANUPRIYAA	ThazhPrappanthara	643271
RANKANN	Plamaram	643833
GUNDY	GANAPATHY	657028
DEEPA	Kunnanchala	634444
MARUTHAN	Metticolony	566999
THALI	NakkupathiPirivu	234242
GUNDYy	GNAPATHY	657028
CHANDRAN	Karadipara	242424

Table -2 Data set C (Actual data without dirty)

Name	place	Pin
CHELLI	Kuruvankandi	643233
THRUPATHI	Kookkampalayam	645633
RANKAMMA	Cheramankandi	702344
MANUPRIYA	ThazhePrappanthara	643271
ABHINAYA	MelePrappanthara	643721
RANKAN	Plamaram	643833
GUNDY	GANAPATHY	657028
SINDHU	THAMANDAN	653010
DEEPA	Kunnanchala	634444
MARUTHAN	Metticolony	566999
THALI	NakkupathiPirivu	234242
CHANDRAN	Karadipara	242424

Table-3 dataset B (After Cleansing)

Name	Place	Pin
CHELLII	Kuruvankandi	643233
TRUPATHIE	Kookkampalayam	645633
RANKAMA	Cheramankandi	702344
MANUPRIYA	ThazhePrappanthara	643271
SINDHU	THAMANDAN	653010
DEEPA	Kunnanchala	634444
MARUTHAN	Metticolony	566999
THALI	NakkupathiPirivu	234242
CHANDRAN	Karadipara	242424

Table –1, table -2 and table-3 represents the datasets U, C, B respectively. The n(U) is 20, n(C) is 12 and n(B) is 9 so n(B) tells that the data loss as well as the presents of the dirty data in the B. so the accuracy of B when comparing with C is 58% and the data loss is 25% and the dirty data in B is .23%.

Conclusion

This paper surveys the data cleansing algorithm for large dataset and some fundamental steps in the data cleaning and data mining. Data cleaning is a very is very young field in the area of computer science research. This paper represents the current research and practices in data cleansing for large data set. This paper also presents some methods and techniques to evaluate and masher the performance of the data cleansing algorithms used for the very large database systems.Although the large number of tools indicates both the importance and difficulty of the cleaning problem.This paper discussed several implementation of various algorithm effectively used in data cleaning which deserve for further research.

ISSN: 2278 – 1323

References

[1]Weijie Wei, Mingwei Zhang, Bin Zhang “A Data Cleaning Method Based on Association Rules”, Northeastern University, Shenyang PP.1-6

[2] Applied Brain and Vision Science-Data cleaning algorithm

[3] Yiqun Liu, Min Zhang, LiyunRu, Shaoping,“Data Cleansing for Web Information Retrieval using Query Independent Features” Journal of the American society for information scienceand technology, Vol58(12) pp-1-15

[4] Chaudhuri, S., Dayal, U” An Overview of Data Warehousing and OLAP Technology”.ACM SIGMOD Record 26(1), 1997.

[5] Timothy E. Ohanekwu and C.I. Ezeife ,“A Token- Based Data Cleaning Technique for Data Warehouse systems” -University of Windsor P.P. 7-12

[6] Yu Qian, Kang Zhang ,“The role of visualization in effective data cleaning” SAC '05 Proceedings of the 2005 ACM symposium on Applied computing P.P 1239-1243

[7] Chris Mayfield, Jennifer Neville, Sunil Prabahakar, “A Statistical Method for Integrating Data Cleaning and Imputation” - Purdue University(Computer Science report-2009) Report Number -09 -008

[8] Sheng Tang “Data cleansing based on mathematical morphology” published in ICBBE 2008 the second InternationalConference-2008

[9] Kazi Shah Nawaz Ripon, AshiqurRahman and G.M. AtiqurRahaman“A Domain-Independent Data Cleaning Algorithm for Detecting Similar- Duplicates”,Journal Of Computers, VOL. 5, NO. 12, December 2010 . P.P 1800-1809

[10] SurbhiAnand and Rinkle Rani Aggarwal "An Efficient Algorithm for Data Cleaning of Log File using File Extensions" International Journal of Computer Applications Volume 48– No.8, June 2012. P.P 13-18

[11] R. Kavitha Kumar and Dr. R.M Chandrasekaran“Attribute correction-data cleaning using association rule and clustering methods” International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.2, March 2011. P.P 22 - 32

[12]JieGu - Random Forest Based Imbalanced Data Cleaning and Classification – JieGu-Software School of Tsinghua University, China .P.P.1-7

[13] Sheng TANG and Si-ping CHEN "Data Cleansing Based on Mathematic Morphology" Bioinformatics and Biomedical Engineering, 2008.ICBBE 2008. The 2nd International Conference on May(IEEE-EXplore) 2008 755 - 758

[14] SurabhiAnand ,Rinkle Rani Aggarwal. “An efficient Algorithm for Data Cleaning of Log File using FileExtension” International journal of Computer Appliactions June-2012 Vol - 48(8).PP 13-18,

[15] Li Zhao, Sung Sam Yang, Sum Peng and LingTock Wang " A New Efficent Data Clencing Method" Springer - DEXA 2002 P.P 484 -804

[16] Aye, T.T. "Web log cleaning for mining of web usage patterns" Computer Research and Development (ICCRD), 2011 3rd International Conference(IEEE Expore) Vol-2 P.P 490 - 494

[17] Yan Cai-rong,SUNGui-ning , GAO Nian-gao ,“Mass Data Cleaning Algorithm based on extended tree-likeknowledge base” –Computer Enginerring and application –PP-146-148

[18] Chris Mayfield, Jennifer Neville and Sunil Prabhakar “ERACER-A database approach for statistical inference and datacleaning” SIGMOD’10, June 6–11, 2010, Indianapolis, Indiana, USA

[19] Shawn R. Jeffery, Minos Garofalakis and Michael J. Franklin "Adaptive Cleaning for RFID Data Streams" ACM VLDB ‘06, September 1215, 2006, Seoul, Korea. P.P 163-174

[20] Manuel CastejonLimas, Joaquin B. Ordieres Mere, Francisco J. Martinezn de Pison

,AscacibarandEliseoP.VergaraGonzaalez"Outlier Detection and Data Cleaning in Multivariate Non-NormalSamples: The PAELLA Algorithm" Data Mining and Knowledge Discovery,Vol 9, P.P 171–187

[21] Mohamed H.H "E-Clean A Data Cleaning" Informatics and Computational Intelligence (ICI) 2011 (IEEE Explore)

[22] R. Deepa and Dr. R Manicka Chezian "A Study on Data Cleansing and Classification Algorithms for Large Dataset Systems" International Journal of Research in Advent Technology, Vol.2, No.9, September 2014 P.P 97-102.

[23] R. Deepa, R Manicka Chezian An Involuntary Data Extraction And Information Summarization Expending Ontology International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.44 (2015).

[24] R.Deepa, R. Manicka Chezian Keyword based web page re-ranking using relevancy and bee algorithm by incorporating user behaviour information in web mining, 103rd Indian Science Congress, 05 January 2016.

[25] R. Deepa , R. Manicka Chezian An Involuntary Data Extraction and Information Summarization Expending Ontology Artificial Intelligence and Evolutionary Computations in Engineering Systems Volume 394 of the series Advances in Intelligent Systems and Computing pp 59-68, 06 February 2016

BIOGRAPHY

R.Deepa received her Bsc.Statistics from P.S.G College of Arts and Science College, Coimbatore, India. She had her Master of Computer Applications from Bharathidasan University,Trichy, India. She holds MPhil, in Computer Sciencefrom Bharathiar University, Coimbatore, India. Shehas 9 years of experience in teaching. She is presently working as an Assistant Professor in NGM College, Pollachi. Her research interest includes Data Mining,Big Data Management, and Image Compression. Nowshe is pursuing her Ph.D Computer Science in Dr.Mahalingam Center for Research and Development atNGM College Pollachi.

Dr.R.Manickachezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published thirty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

ISSN: 2278 – 1323