Application of Framework for Data Cleaning to Handle Noisy Data in Data Warehouse

A. F. Elgamal*, N.A. Mosa, N.A. Amasha

Abstract*— Data cleaning is a complex process which makes use of several technology specializations to solve the contradictions taken from different data sources. In fact, it represents a real challenge for most organizations which need to improve the quality of their data. Data quality needs to be improved in data stores when there is an error in input data, abbreviations or differences in the archives derived from several data bases in one source. Therefore, data cleaning is one of the most challenging stages to clear repeated archives, because it deals with the detection and removal of errors, filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies to improve the quality of the data gathered from distributed sources. It is particularly crucial to extract a correct conclusion from data in decision support systems (DSS). This paper presents an application of general framework for the data cleaning process, which consists of six steps, namely selection of attributes, formation of tokens, selection of the clustering algorithm, similarity computation for the selected attributes, selection of the elimination function, and finally merge. A proposed software is developed with SQL Server 2010 and C# 2010.

Index Terms*— Data cleaning, Data quality, Data warehouse, Duplicate elimination.

I. INTRODUCTION

The explosive growth of government, business, and scientific databases has taken by storm the traditional, manual approaches to data analysis, creating a need for a new generation of techniques and tools for intelligent and automated knowledge discovery in data [1]. Data quality means using an error-free mechanism in the data warehouse. The quality of data needs to be improved by using the data cleaning techniques. Existing data cleaning techniques are used to identify record duplicates, missing values, record and field similarities, and duplicate elimination [2]. Data quality issues are often multifaceted and complex, and it is crucial for information management departments to build applications that support the goal of achieving high-quality data within an organization [3]. The main objective of data cleaning is to reduce the time and complexity of the mining process and increase the quality of datum in the data warehouse [2].

Data cleaning monitoring is an incessant activity which starts right from the data gathering stage and continues until the ultimate choice of analysis and interpretation of the results [4].

Manuscript received January 15, 2014.

A.F. Elgamal, Ass. Professor, Department of Computer Science, Mansoura University, Egypt.

N.A. Mosa, Lecturer, Department of Computer Science, Mansoura University, Egypt.

N.A. Amasha, Instructor, Department of Computer Science, Mansoura University, Egypt.

This process is essential for drawing correct conclusions from data in decision support systems [5]. Errors in data can often be found when multiple data sources are merged [3]. Moreover, the records processed within different systems may have different data formats or representations [6]. When such databases are merged, two records referring to the same entity may not match, resulting in duplicate entries or missing data.

The classical application of data cleaning occurs in data warehouses. Data warehouses are generally used to provide analytical results from multidimensional data through effective summarization and processing of segments of the source data relevant to the specific analysis. Business data warehouses are the basis of decision support systems (DSS) which provide analytical results to officials so that they can analyze a situation and make important decisions. Cleanliness and integrity of the data contribute to the accuracy and correctness of the results and hence affect the impact of any decision made or conclusion drawn[7]. The problem of detecting and eliminating duplicated data is one of the major problems in the broad area of data cleaning and data quality [8]. Duplicate elimination is a hard task, because it is caused by several types of errors. Duplicate records in databases are an essential step in the data cleaning processes. Most existing approaches rely on generic or manually tuned distance metrics for estimating the similarity of potential duplicates [9]. The goal of record matching (duplicate detection) and deduplication is to identify the matching records, defined to be records that correspond to the same real-world entity. The output of record matching (duplicate detection) is pairs of matching records while the output of deduplication is clusters of matching records. The goal of segmentation is to extract structured records from unstructured text [10]. This paper presents the application of a general framework for data cleaning. Section 2 illustrates the steps of the used framework; the application of the framework is illustrated in Section 3; and finally the conclusion is presented in Section 4.

II. FRAMEWORK DESIGN

The Data Cleaning Framework is designed with extensibility as the central idea, and it enables users to customize the data cleaning operations to meet their needs, rather than try to adapt to the rules set forth by the system [11]. Each step of the framework is well suited for the different purposes. Some of the data cleaning techniques are suited for the particular work of the data cleaning process. In addition, the framework offers the user interaction by selecting the suitable algorithm [12]. The framework steps are as follows:

A. Selection of attributes

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

B. Formation of tokens
C. Selection of clustering algorithm
D. Similarity computation for selected attributes
E. Selection of elimination function
F. Merge.
Figure 1 illustrates the framework for data cleaning

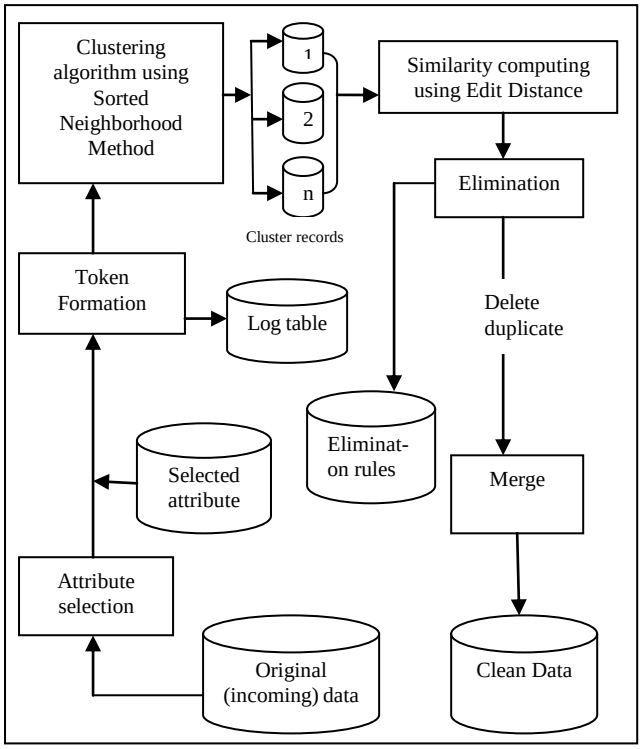


Figure 1: Framework for Data Cleaning

A. Selection of attributes

Attribute selection (feature selection) is one of the important and frequently used techniques in data preprocessing. It reduces the number of attributes by removing irrelevant and redundant attributes, which have no significance in the classification task [13].

A good attribute variable for clustering should contain a large number of attribute values that are fairly uniformly distributed; and such an attribute must have a low probability of reporting errors. For example, a field such as gender has two value states only and consequently cannot impart enough information to identify a match uniquely. Conversely, a field such as surname imparts much more information, but it may frequently be recorded incorrectly [14].

B. Formation of tokens

The smart token-based technique achieves a better result than the record-based techniques of comparable algorithms by using short tokens in record comparisons. The technique also drastically lowers the dependency of data cleaning on match score "threshold" choice [15]. This step attempts to remove typographical errors and abbreviations in data fields. This increases the probability that potentially matching records be brought closer after sorting, which uses keys extracted directly from the data fields. The following steps have to be taken for the best token key before forming the token. The steps are: Removing unimportant tokens (Appendix A contains a Reference table for the Unimportant characters used in this work), Expanding abbreviations using (Appendix B contains a sample of abbreviations), Formation of Tokens, and Maintaining a LOG Table (Figure 2 illustrates the token algorithm).

Input: A Table with dirty data (incoming table) after selected attributes.
Output: LOG token table after tokens.
Begin:
For attribute i=1 to x (last attribute)
For row j=1 to y (last record)

1. Remove unimportant characters such as (special characters, title or salutation, ordinal forms and common words).
2. Expand abbreviations using Reference table.
3. Check the type of row (j) and apply the following to numeric type
   - Convert string into number
   - Sort number
   - Put into LOG table
4. Check the type of row (j) and apply the following to isalphabet type
   - Select first character of every word
   - Sort the characters in alphabetic order
   - Combine them together to obtain the alphabetic token
   - Put into LOG table
5. Check the type of row (j) and apply the following to isalphanumeric type
   - Split alphanumeric to numeric and alphabetic
   - Combine numeric together and alphabetic together
   - Sort the components
   - Put numeric first then put alphabetic to formulate the components
   - Put into LOG table

End

Figure 2: Token-based data cleaning algorithm

C. Selection of clustering algorithm

Clustering method is also known as blocking method in data cleaning for duplicate detection. Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters) based on the block token key [16]. The blocking key is important in the blocking method to gather resemblance records. It is generated by taking the first four or three characters of the attribute, and it can be composed of more than one attribute [17].

The blocking methods use a record attribute to split the data sets into blocks. There are some blocking methods available such as Standard Blocking, Sorted Neighborhood method, Bigram Indexing, Canopy Clustering with TFIDF[18], K-way Sorting Method, Disjunctive Blocking, Fuzzy Blocking and so on[17].

D. Similarity computation for selected attributes

Data cleaning based on similarities involves identification of tuples, where closeness is evaluated using a variety of similarity functions chosen to suit the domain and application.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A variety of string similarity functions are considered, such as edit distance, jaccard similarity, cosine similarity and generalized edit distance for measuring similarities[19]. However, no single string similarity function is known to be the overall best similarity function, and the choice usually depends on the application domain [20].

E. Selection of elimination function

Duplicate elimination methods for data cleaning are based on computing the degree of similarity between nearby records in a sorted database [21]. This step is used to detect or remove the duplicate records from one cluster or many clusters. Before the elimination process, the user should know the similarity threshold values for all the records available in the data set. Several rule-based approaches are proposed for the duplicate elimination process. The distance criteria is mostly used in the rule-based approaches. The commonly available rule-based approaches are the 'Bayes decision rule' for minimum error, Decision with a Reject Region 'Equational theory' and so on [12].

F. Merge

There are different merging strategies used in collecting records as a single cluster. The user must maintain the merged record and the prime representative as a separate file in the data warehouse. This information helps the user for further changes in the duplicate elimination process. This merge step is useful for the incremental data cleaning. [22].

The following section illustrates the application of the framework. A dataset of bank customers contains ID, Name, Birth Date and Address is considered. A proposed software is developed with SQL Server 2010 and C# 2010. Appendix C represent sample of the proposed software screen.

III. APPLICATION

A. Selection of attributes

Two attributes are selected and they can be combined to uniquely identify records. This process depends on the efficiency of an expert who should be aware of the problem domain, and who can select rank attributes according to their unique identifying power. The user in the domain selects two attributes "Name" and "Address" from the table.

B. Formation of tokens

In this step, a given attribute value is transformed into a smart token. Table1 illustrates tokens for records.

Id	Name	Name Token	Address	Address Token
1001	Mohamed Ahmed Solman	AMS	11 Hassanan Dakak	11HD
1018	Mohamed Khaled Shalby	KMS	25 Abd Elsalam Aref	25AEA
1010	Tarek Saad Salah	SST	9 elterah st.	9E
1003	Khaled Mohamed Shalby	KMS	25 Abd Elsalam Aref	25AEA
1009	Mohamed Ahmed Soliman	AMS	11 Hassanan Dakak	11HD
1016	Ebrahem Mohamed Elaraby	EEM	55 Etantawy St.	55E
1017	Ebrahem Mohamed Elaraby	EEM	5 Azeza Elshenawy	5AE
1013	Eman Ahmed Elhodad	AEE	Adb Esalam Aref	AEA
1014	Tarek Saad Salah	SST	9 Elterah St.	9E
1025	Eman Ahmed Elhodad	AEE	9 Abdelsalam Aref	9AA

Table1: The table of tokens

C. Selection of clustering algorithm

Sorted Neighborhood Method (SNM) is used for duplicate detection. Its steps can be summarized as follows:

1- Creating keys: A key is computed for each record in the database by extracting relevant fields or portions of fields which form an important discriminating attribute. The choice of the key depends upon an ''error model'' that draws from domain knowledge. The key selection process is a highly knowledge-intensive and domain-specific process, which should know the characteristics of the data.

2- Sorting Data: Sort the records in the data list to find similar records using the key of step 1.

3- Merge: Move a fixed size window through the sequential list of records limiting the comparisons for matching records to those records in the window. Results after formation of token and clustering are illustrated in Table2.

Id	Cluster Key
1025	aee9aa
1013	aeeaea
1001	ams11hd
1009	ams11hd
1016	eem55e
1017	eem5ae
1003	kms25aea
1018	kms25aea
1010	sst9e
1014	sst9e

Table2: The table of clusters

D. Similarity computation for selected attributes

The edit distance algorithm is used to compute similarity. Given two strings s1[1..m] and s2[1..n] over an alphabet Σ, the edit distance between s1and s2 is the minimum number of edit operations needed to convert s1 to s2. The edit distance problem is used to find the edit distance between s1 and s2. Most common edit operations are the following:

1. Change: Replace one character of s1 by another single character of s2;
2. Deletion: Delete one character from s1;
3. Insertion: Insert one character into s2.

A well-known method for solving the edit distance problem in O *(mn)* time uses the D-table. Let D (i, j), $0 \le i \le m$ and $0 \le j \le n$, be the edit distance between s1[1..i] and s2[1..j]. Initially, D (i, 0) = i for $0 \le i \le m$ and D (0, j) = j for $0 \le j \le n$. An entry D (i, j), $1 \le i \le m$ and $1 \le j \le n$, of the D-table is determined by the three entries D(i −1, j −1), D(i −1, j), and D(i, j −1). The duplication for the D-table is as follows: for all $1 \le i \le m$ and $1 \le j \le n$.

$$D(i,j) = \min \begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+(\text{if } s1(i)=s2(j) \text{ then } 0 \text{ else } 1) \end{cases}$$

The edit similarity ES (s1, s2) is calculated as following.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$$ES(s1, s2) = 1 - \frac{ED(s1, s2)}{Max(s1, s2)}$$

As: ED (s1, s2) is last value for D(i, j).

Table3 illustrates the calculated ES and Figure3 illustrates degree of similarity between records represents chart of these results.

Id	Cluster Key	ED
1025	aee9aa	.67,28,.28,.33,.5,.25,.25,.33,.33
1013	aeeaea	.28,.28,.33,.33,.37,.37,.17,.17
1001	ams11hd	1,.14,.14,.37,.37,.28,.28
1009	ams11hd	.14,.14,.37,.37,.28,.28
1016	eem55e	.83,.37,.37,.17,.17
1017	eem5ae	.37,.37,.14,.14
1003	kms25aea	1,.25,.25
1018	kms25aea	.25,.25
1010	sst9e	1
1014	sst9e	

Table3: table of clusters

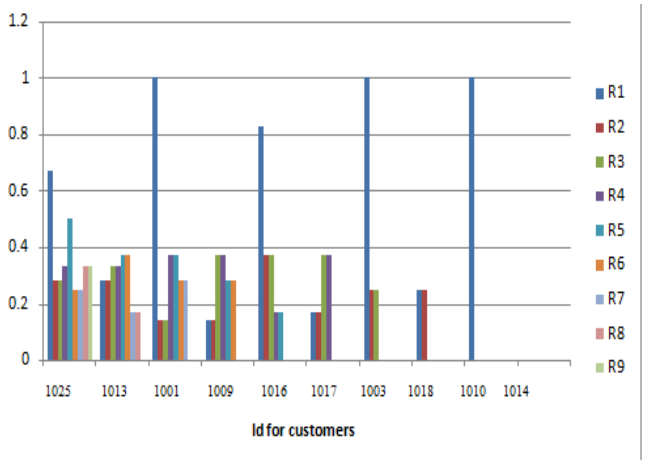


Figure3: degree of similarity

E. Selection of elimination function

In the duplicate elimination step, one copy of the duplicated records is retained and the rest of duplicate records are eliminated. The elimination process is very important to produce cleaned data. This step is used to detect and remove the duplicate records from one or more clusters. During the elimination process, select all possible pairs from each cluster and compare records within the cluster using the selected attributes. Most of the elimination processes compare records within the cluster only. Sometimes other clusters may have duplicate records, of the same value of other clusters. The comparisons of all the clusters are not always possible due to the time constraint and efficiency. Several rule-based approaches are proposed for implementing the duplicate elimination process. The distance criteria are the mostly used rule-based approaches. The commonly available rule-based approach is the Equational theory.

Given two records, r1 and r2.

The rule of duplicate elimination can be presented as:

High similarity (r1.cluster key, r2.cluster key) ∧ ((r1.id) ≠ (r2.id)) —> duplicate

F. Merge

Merged records are loaded into the data warehouse for the decision support process. These merged records are cleaned before loading data into the data warehouse. In the merge step, duplicates should be removed and records should be merged as a cluster. The records from each cluster are appended to the above cluster to form a table.

Table4 illustrates data after cleaning.

Id	Name	Birth date	Address
1001	Mohamed Ahmed Solman	12/11/1969	11 Hassanan Dakak
1003	Khaled Mohamed Shalby	1/22/1980	25 Abd Elsalam Aref
1010	Tarek Saad Salah	12/11/1970	9 Elterah st.
1016	Ebrahem Mohamed Elaraby	12/15/1981	55 Eltantawy st.
1017	Ebrahem Mohamed Elaraby	7/5/1961	5 Azeza Elshenawy
1013	Eman Ahmed Elhodad	2/22/1982	Adb Elsalam Aref
1025	Eman Ahmed Elhodad	2/22/1982	9 Abdelsalam Aref

Table4: data after cleaning

IV. CONCLUSION

Because it is the nature of data to increase and vary from time to time, it is required to clean data to guarantee their quality and facilitate using them in the decision support process. There is not a comprehensive group of techniques for clearing data in any arbitrary field. For example, there are techniques to duplicate elimination of data, while some measure the similarity degree in the fields or records and others form groups for similar records and so on.

Therefore, a framework is applied to customize the data cleaning operations to meet the needs of all users, and this framework consists of six steps working in a sequential order. The proposed software has several advantages such as easy use through interactive interface for the user, in addition to speedy development, effectiveness of the run-time when it is used in different information systems, and the flexibility of all kinds of data.

Appendix A

a. **Special characters** are
` , ‘ ” < > - % + _ ( ) . * - $ # ³ º » ¡ ¢ ¤ ¦ © ª « ®! [ ] ^ \ @ : ; ♦ ¯ ← ↑ → ↓ ™ = ? | { } **% & ‘ + - ~ = ? @ µ ¶ Æ Ç È É @** and so on.

b. **Title or Salutation** tokens are
Herr,Monsieur,Hr,Frau,Admiraal,Admiral,Baron,Brig,Brother, Canon,Capt,Captain,Cardinal,Cdr,Cik,Col,Colonel,Count,Mr,Mrs,Ms,Miss,Dr,Chief,Dean,Doctor,Dra,Drs,Father,General,Jonkheer,Judge,Justice,Kolonel,Lady,Lic,Madame,Major,Master,Miss,Mme, Prof,Prof Dr, Professor, The Hon Dr,The Hon Justice,The Hon Miss,The Hon Mr,The Hon Mrs,The Hon Ms,The Hon Sir, Sir,Sister,Sqn Ldr,Sr,Sr D and so on.

c. **Ordinal forms** are
st,nd,rd,th,ad,ado,and,an,a,din,dor,id,idol,in,ion,dial,do,lion,lir,lord,loan,no,land,nod,road,rand,radio,rin,old,ran,al,in,or and so on

d. **Common abbreviations** are ‘Pvt’, ‘Ltd’, ‘Co’, ‘Rd’, ‘St’, ‘Ave’, ‘Blk’, ‘Apt’, ‘Univ’, ‘Sch’, ‘Corp’ and etc

e. **Common words** are
by,she,or,as,what,go,their,can,who,get,if,would,her,all,my,make,about,know,will,as,up,one,time,there,the,be,and,of,a,in,to,have, to,it,that,for,you,he,with,on,do,say,this,they,at,but,we,his and etc

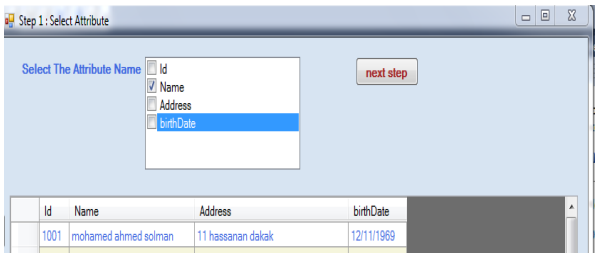
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Appendix B

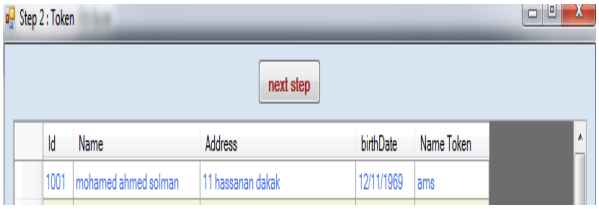
Shortcut	Full form
a	(in dates) ante
bbrev.	abbreviation (of)
Argt.	argument
Arith.	arithmetic
Arrangem.	Arrangement
art.	Article
Bk.	Book
BNC	British National Corpus
Bord.	Border
cent.	Century
Cent.	Central
Chr.	Christian
Dict.	Dictionary

And so on in : Oxford English Dictionary

Appendix C



Selection of attributes



Formation of tokens

Clustering

REFERENCES

1. Magdi Kamel, "Data Warehousing and Mining", IGI Global, 2009.(URL:http://www.igi-global.com/chapter/data-preparation-data-mining/10872)
2. Israr Ahmed and Abdul Aziz," Dynamic Approach for Data Scrubbing Process", International Journal on Computer Science and Engineering Vol. 02, No. 02, pp 416-423, 2010.
3. Jason D. Van Hulse, TaghiM. Khoshgoftaar, Haiying Huang, "The pairwise attribute noise detection algorithm", Knowl Inf Syst, Springer, 2006.
4. Enrico Fagiuoli, Sara Omerino and Fabio Stella, "Mathematical Methods for Knowledge Discovery and Data Mining ", IGI Global,2008. (URL:http://www.igi-global.com/chapter/bayesian-belief-networks-data-cle-aning/26141)
5. Hamid Haidarian Shahri and Ahmad Abdollahzadeh Barforush, "A Flexible Fuzzy Expert System for Fuzzy Duplicate Elimination in Data Cleaning", Springer, DEXA 2004, LNCS 3180, pp. 161 - 170, 2004.
6. Galhardas, H. Florescu, D. Shasha and D. Simon, "An extensible framework for data Cleaning", In Proceedings of 18th international conference on data engineering, IEEE Computer Society, San Jose, 2000.
7. Judice, Lie Yongkoh," Correlation-Based Methods for Biological Data Cleaning", DOCTOR OF PHILOSOPHY, National University of Singapore, 2007.
8. Rohit Anantha krishna, Surajit Chaudhuri and Venkatesh Ganti, "Eliminating Fuzzy Duplicates in Data Warehouses", Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002.
9. Mikhail Bilenko and Raymond J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures", Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington DC,pp.39-48,August 2003.
10. Arvind Arasu, Surajit Chaudhuri, Zhimin Chen, Kris Ganjam, Raghav Kaushik and Vivek Narasayya," Towards a Domain Independent Platform for Data Cleaning", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2011.
11. Rand Siran Gu," Data Cleaning Framework: an Extensible Approach to Data Cleaning ", degree of Master of Science in Computer Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2010.
12. J. Jebamalar Tamilselvi and V. Saravanan," A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.5, May 2008.
13. D.Lavanya and K.Usha Rani, "Analysis of Feature Selection with Classification: Breast Cancer Datasets", Indian Journal of Computer Science and Engineering, Vol. 2 No. 5, Oct-Nov 2011.
14. Lifang GU, Rohan Baxter, Deanne Vickers and Chris Rainsford," Record Linkage: Current Practice and Future Directions", "Canberra, ACT 2601, Australia.
15. (URL: http://datamining.csiro.au)
16. Christie I. Ezeife AND Timothy E. Ohanekwu, " Use of Smart Tokens in Cleaning Integrated Warehouse Data", International Journal of Data Warehousing & Mining, 1(2), 1-22, April-June 2005.
17. K. M. Bataineh, M. Naji, M. Saqer, " A Comparison Study between Various Fuzzy Clustering Algorithms", Jordan Journal of Mechanical and Industrial Engineering, Volume 5, Number 4, Aug. 2011.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

18. Jebamalar Tamilselvi J. and Saravanan V., "Token-based method of blocking records for large data warehouse", Advances in Information Mining, ISSN: 0975–3265, Volume 2, pp-05-10, 2010.
19. Rohan Baxter, Peter Christen and Tim Churches, "A Comparison of Fast Blocking Methods for Record Linkage", CMIS Technical Report, 2003.
20. S. Chaudhuri, V. Ganti, and R. Kauskik," A Primitive Operator for Similarity Joins in Data Cleaning ", (URL:http://www.yumpu.com/en/document/view/11885816/apr-imitive-operator-for-similarity-joins-microsoft-research)
21. S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, "Robust and efficient fuzzy match for online data cleaning. In Proceedings of the ACM SIGMOD, June 2003.
22. Wai Lup Low, Mong Li Lee and Tok Wang Ling, "A knowledge-based approach for duplicate elimination in data cleaning ", Information Systems 26, pages 585–606, 2001.
23. J. Jebamalar Tamilselvi and V. Saravanan, " Detection and Elimination of Duplicate Data Using Token-Based Method for a Data Warehouse: A Clustering Based Approach ", International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 5, 2009.