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
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 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.
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.

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.

D. Similarity computation for selected attributes
The edit distance computation is used to compute similarity. Given two strings s1[1..m] and s2[1..n] over an alphabet Σ, the edit distance between s1 and 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 ≤ i ≤ m and 0 ≤ j ≤ n, be the edit distance between s1[1..i] and s2[1..j]. Initially, D(i, 0) = i for 0 ≤ i ≤ m and D(0, j) = j for 0 ≤ j ≤ n. An entry D(i, j), 1 ≤ i ≤ m and 1 ≤ j ≤ 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 ≤ i ≤ m and 1 ≤ j ≤ n.

Table1: The table of tokens

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Name Token</th>
<th>Address</th>
<th>Address Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>Mohamed Ahmed Soliman</td>
<td>AMS</td>
<td>11 Hassanan Dakak</td>
<td>11HD</td>
</tr>
<tr>
<td>1018</td>
<td>Mohamed Khaled Shalby</td>
<td>KMS</td>
<td>25 Abd Elsalam</td>
<td>25AEA</td>
</tr>
<tr>
<td>1010</td>
<td>Tarek Saad Salahi</td>
<td>SST</td>
<td>9 Eltehrak st.</td>
<td>9E</td>
</tr>
<tr>
<td>1003</td>
<td>Khaled Mohamed Shalby</td>
<td>KMS</td>
<td>25 Abd Elsalam</td>
<td>25AEA</td>
</tr>
<tr>
<td>1009</td>
<td>Mohamed Ahmed Soliman</td>
<td>AMS</td>
<td>11 Hassanan Dakak</td>
<td>11HD</td>
</tr>
<tr>
<td>1016</td>
<td>Elsaham Mohamed Elaraby</td>
<td>EEM</td>
<td>35 Eltehrak St.</td>
<td>55E</td>
</tr>
<tr>
<td>1017</td>
<td>Elsaham Mohamed Elaraby</td>
<td>EEM</td>
<td>5 Azza Eshamawy</td>
<td>5AE</td>
</tr>
<tr>
<td>1013</td>
<td>Eman Ahmed Elhodad</td>
<td>AEE</td>
<td>Adh Elsalam</td>
<td>AEA</td>
</tr>
<tr>
<td>1014</td>
<td>Tarek Saad Salahi</td>
<td>SST</td>
<td>9 Eltehrak St.</td>
<td>9E</td>
</tr>
<tr>
<td>1025</td>
<td>Eman Ahmed Elhodad</td>
<td>AEE</td>
<td>9 Abdelsalam</td>
<td>9AA</td>
</tr>
</tbody>
</table>

D(i, j) = \min(D(i−1, j)+1, D(i, j−1)+1, D(i−1, j−1)+(if s1(i)=s2(j) then 0 else 1))

The edit similarity ES (s1, s2) is calculated as following.
Application of Framework for Data Cleaning to Handle Noisy Data in Data Warehouse

Given two records, r1 and r2.
The rule of duplicate elimination can be presented as:
High similarity (r1.cluster key, r2.cluster key) \land ((r1.id) \neq (r2.id)) \rightarrow\text{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.

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Birth date</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>Mohamed Ahmed Solman</td>
<td>12/11/1969</td>
<td>Hassanan Dakak</td>
</tr>
<tr>
<td>1003</td>
<td>Khaled Mohamed Shaiby</td>
<td>1/22/1980</td>
<td>Abd Elsalam Aref</td>
</tr>
<tr>
<td>1016</td>
<td>Ebrahim Mohamed Elaraby</td>
<td>12/15/1981</td>
<td>Eltantawy st.</td>
</tr>
<tr>
<td>1017</td>
<td>Ebrahim Mohamed Elaraby</td>
<td>7/5/1961</td>
<td>Azeza Elshenawy</td>
</tr>
<tr>
<td>1013</td>
<td>Eman Ahmed Elhodad</td>
<td>2/22/1982</td>
<td>Abdl Elsalam Aref</td>
</tr>
<tr>
<td>1025</td>
<td>Eman Ahmed Elhodad</td>
<td>2/22/1982</td>
<td>Abd Elsalam Aref</td>
</tr>
</tbody>
</table>

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 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 
', `>, <, %, +, (, ), *, $, #, ;, e, n, !, @, `, ^, &, &

b. Title or Salutation tokens are
Herr, Monsieur, Hr. Frau, Admiral, Admiral, Baron, Brig. Brother, Canon, Capt., Captain, Cardinal, Cdr, Cik, Col, Colonel, Count, Mr, Mrs., Miss, Dr., Chief, Dean, Doctor, Dra., Drs., Father, General, Jonkheer, Judge, Justice, Kolonel, Lady, Lic., Madame, Major, Master, Mi ss., 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, Squ., Ldr., Sr., Sr. D. and so on.

c. Ordinal forms are


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.
Appendix B

<table>
<thead>
<tr>
<th>Shortcut</th>
<th>Full form</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>(in dates) ante</td>
</tr>
<tr>
<td>brev.</td>
<td>abbreviation (of)</td>
</tr>
<tr>
<td>Argt.</td>
<td>argument</td>
</tr>
<tr>
<td>Arith.</td>
<td>arithmetic</td>
</tr>
<tr>
<td>Arrangem.</td>
<td>Arrangement</td>
</tr>
<tr>
<td>art.</td>
<td>Article</td>
</tr>
<tr>
<td>Bk.</td>
<td>Book</td>
</tr>
<tr>
<td>BNC</td>
<td>British National Corpus</td>
</tr>
<tr>
<td>Bord.</td>
<td>Border</td>
</tr>
<tr>
<td>cent.</td>
<td>Century</td>
</tr>
<tr>
<td>Cent.</td>
<td>Central</td>
</tr>
<tr>
<td>Chr.</td>
<td>Christian</td>
</tr>
<tr>
<td>Dict.</td>
<td>Dictionary</td>
</tr>
</tbody>
</table>

And so on in: Oxford English Dictionary

Appendix C

Selection of attributes

Formation of tokens

Clustering

REFERENCES


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.


15. (URL: http://datamining.csiro.au)
