JOURNAL OF COLLEGE OFEDUCATION ___ NO 6..........2010

Basic Steps to Get Data Quality for Data Mining

Dr. ZAKI .S. TOWFIK
Department of Computer Science
Collage of Science
University of AI-Mustansiriya - Iraq
Email- zeik_sead@yahoo.com
Email- zekisaeed@gmail.com

Abstract

The Data extracted from many sources will be integrated and then transform into suitable form. Tthese data may be includes many errors and noise or inconsistencies data. It is necessary to clean the data to get quality data before the data mined from errors and noise data. The cleaning is the first task before any data analysis. The resultant of cleaning analysis/model can be stamped for data quality which very impotent for data minig process because without data quality the algorithms of data nining can not work well or the result of algorithms not good.

Therefor this paper deal with Basics steps to clean data that extracted from many sources to get good quality data for data mining also reduce processing time, storage data and reducing costs and increasing profits, for this case an implementation for data selected from clinical chemical test for yarmook hospital education to detect and remove the errors or noise and or inconsistencies data.

Keywords: Knowledge Discovery, Data Preparation, Data Cleaning is Not Completed Yet,

Basics steps of Data Cleaning

خطوات أساسية لتحسين نوعية البيانات

لغرض تعدين البيانات

الخلاصة:

البيانات المستخرجة من مصادر المختلفة تتكامل وتحول إلى الصيغة المناسبة. هذه البيانات تحتوي على العديد من الأخطاء أو عدم تناسق في البيانات. لذلك من الضروري تنظيف إزالة الأخطاء من البيانات الموجودة للحصول على بيانات عالية جودة قبل إجراء عملية تعدين البيانات. تعدّ عملية التنظيف البيانات من هذه المهمة الأولى لأي تحليل البيانات وقد يترتب العملية تحليل البيانات ووضع نموذج جيد من البيانات. من دون إجراء عملية تنظيف البيانات تكون من الصعب تطبيق خوارزميات تعدين البيانات لان النتايج سوف تكون غير جيدة.

لذلك تم وضع خطوات أساسيات لتنظيف البيانات التي تستخرج من مصادر عديدة للحصول على بيانات ذات نوعية جيدة وذات قيمة عالية في عملية تعدين البيانات والتي تعمل على تخفيض الوقت اللازم عند عمليات تعدين البيانات وتخزين البيانات وتخفيض التكاليف وزيادة الأرباح نتيجة دقة البيانات الناتجة ، مع تنفيذ لبيانات لتنفيذ خطوات تم اختيار بيانات من استمارة الفحوص المختبرية للكيمياء السريرية لمستشفى اليرموك ألتعلمي لكشف الأخطاء وإزالتها أو عدم اتساق البيانات المرضى.

1. Introduction

A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. One of important and complex steps in quality data process is the cleaning data step, also called data cleansing or scrubbing. It deals with detecting and removing errors or noise and inconsistencies from data before data reach to data mining in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

Data Mining require and provide extensive support for data cleaning. They load and continuously refresh huge amounts of data from a variety of sources so the probability that some of the sources contain "dirty data" is high. Furthermore, data mining are used for decision making, so that the correctness of their data is vital to avoid wrong conclusions. For instance, duplicated or missing information will produce incorrect or misleading statistics[1].

Data is not integrated as for data mining but needs to be extracted from multiple sources, transformed and combined during query runtime. The corresponding communication and processing delays can be significant, making it difficult to achieve acceptable response times. The effort needed for data cleaning during extraction and integration will further increase response times but its mandatory to achieve useful query results[2].

Some research groups concentrate on general problems not limited but relevant to data cleaning, such as special data mining approaches. In this paper describe the ten basic step for cleaning data with algorithm of cleaning data and example of data which taken form yarmook hospital education for implementation.

2. Knowledge Discovery

Knowledge Discovery has been defined as the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data'. It is a process of which data mining forms just one part see figure 1.

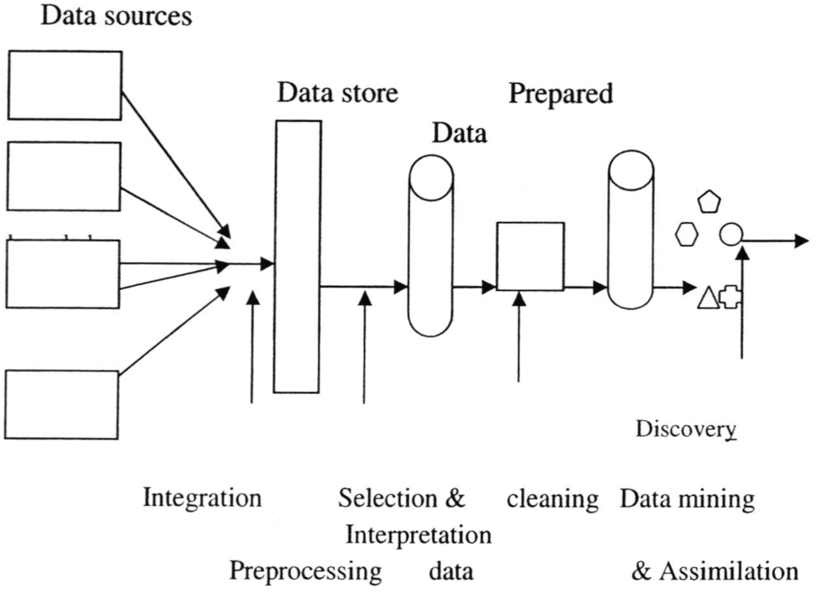

Figure 1 Shows process of cleaning data to complete knowledge discovery
in Data mining.

Data comes in, possibly from many sources. It is integrated and placed in some common data store. Part of it is then taken and pre-processed into a standard format. This 'prepared data' is cleaned from missing words or unknown then passed to a data mining algorithm which produces an output in the form of rules or some other kind of 'patterns'. These are then interpreted to give knowledge discovery—new and potentially useful knowledge[2].

This brief description makes it clear that although the data mining algorithms, which are the principal subject, are central to knowledge discovery they are not the whole story. The pre-processing of the data and the interpretation of the results are both of great importance.

3. Data Preparation

For many applications the data can simply be extracted from a database in the form many sources with many errors and noise data and may be in different forms. The data stored in suitable medium to prepare for translate it in standard form, perhaps using a standard access method such as Oriented Database (ODB). However, for some applications the hardest task may be to get the data into a standard form in which it can be analyzed[3]. For example data values may have to be extracted from textual output generated by a fault logging system or (in a crime analysis application) extracted from transcripts of interviews with witnesses. The amount of effort required to do this may be considerable[3,5].

4. Data Cleaning is Not Completed Yet

After the implement ten basic steps and analyst-specific checks are done, data cleaning is not completed until the noise in the data is eliminated. Noise is the idiosyncrasies of the data. The particulars, the "nooks" that are not part of the sought-after essence (e.g., predominant pattern) of the data with regard to the objective of the analysis/model. Ergo, the data particulars are lonely, not-really-belonging-to pieces of information that happen to be both in the population from which the data was drawn and in the data itself (what an example of a double-chance occurrence!) Paradoxically, as the analyst includes more and more of the prickly particulars in the analysis/model, the analysis/model becomes better and better, yet the analysis/model validation becomes worse and worse. Noise must be eliminated from the data[3].

The solve of this problem is to provide a procedure for eliminated noise from the actual records that define the idiosyncrasies of the data. Now, the analysis/model can be built with cleaned data that reliably represents the sought-after essence of the data, yielding a well conducted analysis and a well-fitted model or the errors con be translated to another file or report and then correct the error or noise by retuned the data in file(prepared data), that means the data go in the cycle until the data become quality data as shown in figure 2.

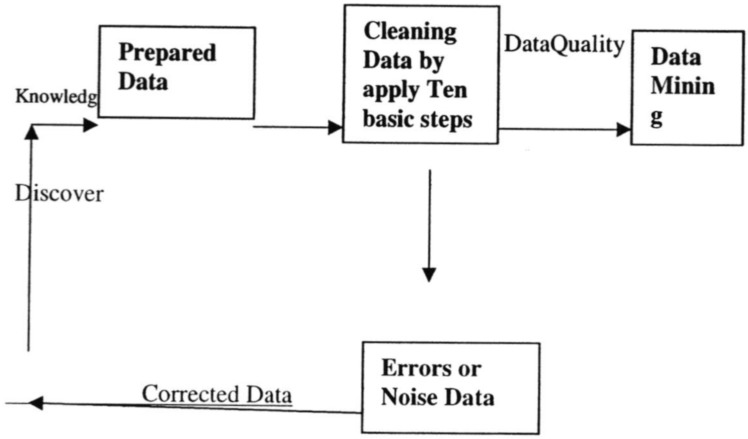

Figure 2 The Cycle of Data in Data Cleaning Process

5. Cleaning Process Method

There are many methods and techniques but these methods no clear in fitting the data or that can aid in the cleaning of errors or noise in primary species an species-occurrence databases. But need to automated methods. This paper looks in detail at a range of methods for cleaning data in cycling process Until the "Noise" is Eliminated under control of analyst data[4].

For this reason the analysis data are needed after the extracted, stored and prepared data able to determine errors and noise from the data themselves.

For this purposed ten basic steps for cleaning data to get quality data which it is very important for data mined and its algorithms can reach good result and its quality is merely a factor of fitness for use or potential use and is a relative term .

Below the ten basics of data cleaning and algorithm for these ten basics to get data quality[6,7,8].

5.1. Basics steps of Data Cleaning

1. Check frequencies of continuous and categorical variables for unreasonable distributions.

2. .Check frequencies of continuous and categorical variables for detection of unexpected values. For continuous variables, look into data "clumps" and "gaps."

3. Check for improbable values (e.g., a boy named Sue), and impossible values (e.g., age is 120 years young, and x/0).

4. Check the type for numeric variables: Decimal, integer, and date.

5. Check the meanings of missing formative values, e.g., "NA", the blank " ", the number "0", the letter "O", the dash "—", and the dot ". ".

6.Check for out-of-range data: Values "far out" from the "fences" of the data. [1]

7. Check for outliers: Values "outside" the fences of the data. [1]

8. Check for missing values, and the meanings of their coded values, e.g., the varied string of "9s", the number "0", the letter "O", the dash "—", and the dot ". ".

5.2 Algorithm

The algorithm to applied to basic steps for prepared data ST1,ST2, which are temporary files.

```
1. Missing values and missing type:
Begin
   For 1 to z1
         .Sort attributes values (key problem ) in ST1 with null value.
           If attribute value is null.
             THEN The number of records of null value is equal to ST1.
           send the data error to error file.
        3. Else Send the correct records to ST2.
           End if
      End FOR
2. Misspelling
   Begin
         Check the typing of the name
         IF the letters fist letter = number or small letter or "."or"," or "-
      " or "0"
            THEN send the record to the file error and
                  "Misspelling".
         IF there is duplicate in the letter
            THEN send the record to
            the file error and" Misspelling" error to file error. Send the
      records to ST2.
          End IF
      End if.
```

9. Check the logic of data, e.g., response rates cannot be 110%, and weigh contradictory values, along with conflict resolution rules, e.g., duplicate records of BR's DOB: 12/22/15 and 12/22/51.

```
4. Illegal values or outsideBegin
  IF Cardinality value not Greater than  OR Less than threshold
 THEN send the records of error to file error  Send the  records to
ST2.
 End IF.
  End.
5. DuplicateBegin  Sor attributes values in ST1.
```

```
 Compare the first record with other records  . IF  the process is
equal
THEN  purge the redundant record and send to file error.
    Send the correct records to ST2.
     End IF.
End.
6. Varying value representBeginSort attributes values in ST1.
 Comparing attribute value set of a column of one source in the
against that  of a column of another source.  IF compared attributes
are different in  value  THEN send the record error to error file send
the correct   records to ST2. End if.
End.
```

6. Implementation

To applied the some basic steps to data for cleaning to get quality data which it important for data mined. Selected sample of data from for the lab form for clinical chemical test which is special for yarmook hospital education as in the shown in table 1. Not that can applied all basic sets if have huge data.

These data putted in the file which is prepared (ST1) for procedure to detect the error or noise data from file that prepared for theses procedure see in figure 2.

Tabe(1). The data before program process

ide	Patient name	Sex	age	N.value Blood suger / mmol/l	Test1 value	N. value Blood uera/ mmol/l	Test2 value
1001	ali	M	45	3.6-6.1	3.9	3.3-7.5 \	3.7
1002	sammer	M	50	3.6-6.1	5.2	3.3-7.5	8
1003	Nadea	F	46	3.6-6.1	3.8	3.3-7.5	4.2
1005	Zad	M	60	3.6-6.1	7.3	3.3-7.5	6.1
	Rbab	F	55	3.6-6.1	4.8	3.3-7.5	4.5
1006	Tywfik	M	41	3.6-6.1	6	3.3-7.5	4.1
1007	Ahmmad	M	35	3.6-6.1	3.3	3.3-7.5	5.5
1008	abaz	M	30	3.6-6.1	3.6	3.3-7.5	3.5
1009	Diana	F	30	3.6-6.1	3.8	3.3-7.5	3

The program detects the error data from sources and loaded it in the error data to error or noise file(ST2) or report for repair data by control analyst data as shown in the Table2. Can do the procedure of cleaning data in cycle many time to ensure that all data is complete without errors. In Statues empty that means the record is correct.

Table(2) The data after detected by program

ide	Patient name	sex	age	N.value Blood sugar / mmol/l	Test1 value	N. value Blood urea/ mmol/l	Test2 value	Statues
1001	ali	M	45	3.6-6.1	3.9	3.3-7.5	3.7	
1002	sammer	M	50	3.6-6.1	5.2	3.3-7.5	8	Illegal values
1003	Nadea	F	46	3.6-6.1	3.8	3.3-7.5	4.2	
1005	Zad	M	60	3.6-6.1	7.3	3.3-7.5	6.1	
	Rbab	F	55	3.6-6.1	4.8	3.3-7.5	4.5	Missing values
1006	Tywfik	M	41	3.6-6.1	6	3.3-7.5	4.1	
1007	Ahmmad	M	35	3.6-6.1	3.3	3.3-7.5	5.5	
1008	abaz	M	30	3.6-6.1	3.6	3.3-7.5	3.5	
1005	Zad	M	60	3.6-6.1	3.8	3.3-7.5	3.5	Duplicate
1009	Diana	F	30	3.6-6.1	3.8	3.3-7.5	3	Illegal type

7. Conclusion and Future Work

In this paper, we have introduced basics steps for detecting data quality problems in high volume transaction processing. Described the procedure can remove the noise and error from huge data to get quality data for data mined. data cleaning is not completed until the error or noise eliminated, there for put proposed procedure will continue into cycle process until the elimination all detected errors and then the word data cleaning is Not completed yet is valid .

In the implantation the data that taken sample from form of lab form for clinical chemical test which is special for yarmook hospital education only two statues which are Blood sugar and Blood urea many errors are detects and its statues which can reaper by data analyst in cycle procedure. In future can make cycle procedure automatically and can make center computer which connected to all computer sections in hospital .

Reference:

[1] Thomas C. Redman, Data Quality: "The Field Guide, Digital Press", Boston, 2001.

[2] Hanne Riis Nielson and Flemming Nielson " principles of data minig" Springer-Verlag London Limited 2007.

[3] Dorian Pyle,"Data Preparation for Data Mining", Morgan Kaufmann, 1999.

[3] Tamraparni Dasu and Theodore Johnson, "Exploratory Data Mining and Data Cleaning", Wiley, 2003.

[4] Jeffrey W. Seifert "Data Mining: An Overview" CRS Report for Congress" 2004.

[5] Jeffrey W. Seifert " Data Mining and Homeland Security: An Overview: " CRS Report for Congress 2008.

[6] Dalcin, E.C. 2004. "Data Quality Concepts and Techniques Applied to Taxonomic Databases". Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine,

[7] English, L.P. "Improving data warehouse and business information quality": methods for reducing costs and increasing profits. John Wiley & Sons, Inc., New York 1999.

[8] Redman, T.C. "Data Quality: The Field Guide". Boston, MA: Digital Press 2001.

[9] M. Berry and G. Linoff, "Data Mining Techniques", John Wiley, 1997.

Tamraparni Dasu and Theodore Johnson, Exploratory Data Mining and Data Cleaning", Wiley, 2003.