JK TOOLS for Detection and Correction Erroneous Data from Database

[a]Jay Kumar M.Purohit, [b]S.B.Kishor, [c]Ajay S. Kushwaha
[a]Research Scholar, Gondwana University, Gadhchiroli, India, 4424605)
[b]HOD, Dept of Computer Science, S.P.Collage,Chandrapur,India,442401)
[c]HOD, Dept of Computer Application, BIT, Ballarpur)

Abstract

Correcting data from impurities is an integral part of data processing and maintenance. One important aspect of JK TOOLS is identification of the basic causes of the errors detected and using that information to improve the data entry process to prevent those errors from reoccurring. It discusses guidelines, methods and tools that can assist museums and herbaria to follow best practice in digitising, documenting and validating information. This paper is based on the concept about how to avoid dirty and faulty data to get populated in the databases as well as data files. Most of the databases are only providing the data validation techniques in the form of integrity constraints which is useful for valid data entry. But in most of the database such as M.S Excel which is also considered as good earlier version of the database fails to provide the running JK TOOLS and data validation techniques, In short in case of any spelling correction it is not possible to correct it automatically as in the case of M.S.Word where it is provided the facility of current JK TOOLS procedure by providing the list of related spelling, it also warns the user in the form of green and red wavy lines which is not available in the any of the database. Our research work is to provide such kind of JK TOOLS algorithms in the form of in build procedure.

Introduction

JK TOOLS is the process of determining the faulty data, inaccurate data from the database. The actual process of JK TOOLS may involve checking the values entered in the dataset of the database (tables),JK TOOLS differs from data validation it rejects the faulty data at the time of data entry. JK TOOLS like data validation performs validation & it performs very strong and strict validation. For ex:-

a) Rejecting home add without pin code.

b) Rejecting similar entries that matches existing records

JK TOOLS also improves the data quality by making short codes into certain meaningful form such as st->street, rd->road, ny->new york,nu-delhi->new Delhi .These process of standardizing of data is also known as **Harmonization of data.**

Research work carried out by the researcher in this paper is about the designing an algorithm which eliminates the dirty,errorneous data from the database or data files,textfiles.Since incorrect data works as infectious virus which spreads from files to files and results in great economic losses, great expenses. Algorithm is designed such a way that it useful for JK TOOLS on any type of data sources, b'coz clean data is essential requirement for quality data. In order to implement so designed algorithm in the form of real time software, Developer needs to pay attention, towards quality checking algorithm

at the data entry point as well as correcting the corrupted data files which is full of faulty and corrupted data. Here we are trying to focus on JK TOOLS in text.

JK TOOLS Framework

- Searching for errors
- Knowing & categorize errors
- Knowing the sources of errors
- Correcting or removing the errors
- Documenting the type & sources of errors
- Modify data entry procedure to reduce future errors
- Preventing the error rather than detection of error and removal of error

Objectives of JK TOOLS

- **Error Reduction**:- reducing errors and improving data quality
- **Error Understanding**:-understanding errors controls errors propagation & data quality can be further improved
- **Error Documentation**:-it involves correction or changes made to the error are documented
- **Error Preventition**:-preventition of error is always recommended later detecting and correcting the errors
- **Merge & Purge:-**it is required to merge the two or more Databases. it gives rise to new kind of errors.

Principles of Rules based correction and detect in of Errorneous data from database

- Data quality directly affects data warehouse acceptance.
- Source data is typically “dirty” and must be made reliable for warehousing.
- Data models are a good source of data integrity rules
- Both source and target data models yield integrity rules.
- Data correcting rules combine integrity rules with actions to be taken when violations are detected.
- Both source data and target data may be cleansed.
- Auditing reports data integrity violations, but does not fix them.
- Filtering discards data that violates integrity rules.

- Correction repairs data that violates integrity rules.
- Choice of techniques involves severity of the data quality problem, available data models, and commitment of time and resources.

  Whatever techniques are chosen, a systematic, rule-based approach yields better results than an unstructured approach

Integrity rules vs. JK TOOLS rules

- Integrity rules: refer to the way the data must conform to meet business rules.
- Correcting rules: combine the definition from the integrity rule with the action to be taken in the event of a violation

Without JKTOOLS Strategy

- DW will suffer from:
  - lack of quality
  - loss of trust
  - diminishing user base, and
  - loss of business sponsorship and funding Entity-Relationship vs. Dimensional Models

Steps in JK TOOLS Data Refinement Model

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

Parsing

- Parsing locates and identifies individual data elements in the source files and then isolates these data elements in the target files.

Input Data from Source File

Jaykumar M.Purohit ASST PROF.

Govt College of Engg

Bypass Road

Chandrapur(m.s) 442401

Parsed Data from in Target File

First Name:-Jaykumar Purohit

Middle Name:-Manoherlal

Last Name:-Purohit

Desig:-ASST PROF.

Firm:-Govt College of Engg

Location:-Bypass Road,Ballarpur

City:-Chandrapur (m.s)

Pin:- 442401

State:-Maharashtra

Parsed Data

First Name:-Jaykumar Purohit

Middle Name:-Manoherlal

Last Name:-Purohit

Desig:-ASST PROF.

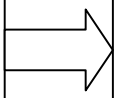
Firm:-Govt College of Engg

Location:-Bypass Road,Ballarpur

City:-Chandrapur (m.s)

Pin:- 442401

State:-Maharashtra



Corrected Data

First Name:-Jaykumar Purohit

Middle Name:-Manoherlal

Last Name:-Purohit

Desig:-ASST PROF.

Dept:-Computer Sci

Firm:-Govt College of Engg

Location:-, Bangali Camp

Street:- Bypass Road

City:-Chandrapur

Zip code:-4424

Zip code + Pin:- 442401

Matching

- Searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.

Standardizing

- Standardizing applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom business rules.

Types of Errors

- Here types of errors are considered in the college information system are as follows
- 1. Numeric values in place Non-numeric (Name, Gender, and City)
- 2. Non-numeric values in place of numeric (phone no, registration no, date)
- 3. Invalid or Redundant ID's
- 4. Invalid Gender.

ID Validation Algorithm

- Step 1 start
- Step 2.check for alphabet in input ID, Eliminate if occurs
- Step 3 Concatenate id's.with preceding Zeros (0) as per following rules
- 3.1. If length of ID is equals to 1 the replace ID="00"+ID
- 3.2. If length of ID is equals to 2 then replace ID="0"+ID
- 3.3. If length of ID is greater than 3 then take only 3 characters eliminates rest
- Step4 Change the ID as per the following rules
- 4.1. If student ID then ID="S"+ID
- 4.2. If Department ID then ID="D"+ID
- 4.3. If Course ID then ID="C"+ID
- 4.4. If Subject ID then ID="Sub"+ID
- Step 5. Return clean ID

Alphabetic Validation Algorithm

- Step 1 Start
- Step 2 Check the input String for all character of upto the null character is encountered
- Step 3 Check if char=’5’ OR char=’0’ OR char=’$’ OR char=’&’ OR char=’@’ OR char=’I’ OR char=’I’ OR char=’1’ OR char=’!’ Then
- Replace 0 by o
- Replace 5,$,& by S
- Replace @ by a
- Replace by ,!,1,I by 1
- Step 4 Return Formatted Strings
- Step 5 Stop

Sample Output Before Correcting

Sid	Sname	Gen	city	Co-no	Phno	Course id
S001	A.Rao	Mle	NAG	11OO23	9422907637	C1
S002	S.jyoti	FML	MUM	1122008	9420080013	C02

AfterCorrecting

Sid	Sname	Gen	city	Code no	Phn	Course id
S001	A.Rao	Male	Nagpur	110023	9422907637	C001
S002	S.jyoti	Female	Mumbai	1122008	9420080013	C002

CONCLUSION

In these paper researcher has proposed some of the JK TOOLS algorithms for databases and text files. It can detect errors, programmatically create valid values and refine the fields in the database. The information age has meant that collections' institutions have become an integral part of the environmental decision making process and politicians are increasingly seeking relevance and value in return for the resources that they put into those institutions. It is thus in the best interests of collections' institutions that they produce a quality product if they are to continue to be seen as a value-adding resource by those supplying the funding.

It is been observed that there are most of the primary species databases which collects and maintains the data about the plant and species as well as their locations and occurings.but often it is observed that some of the older database collection or recording about the plant and species are not found to be accurate about their location, hence in our research we tried to mentioned more accurate way to provide information about the plant and primary species data base. For Ex Natural growing plants in lake are mostly occurs in "Ramala Lake" 5km away from the road side at chandrapur city.

These way of referring data provide more precise way of referring the data, which was not earlier mentioned, which we tried to implement in our research. One of the most challenging things is that providing the species name in all possible different languages of different countries so that there will be no confusion about getting and providing the information about the species.

Best practice for database information in museums and herbaria and institutions maintaining survey and observational information means making the data as accurate and possible, and using the most appropriate techniques and methodologies to ensure that the data are the best they can possibly be. To ensure that this is the case, it is essential that data entry errors are reduced to a minimum, and that on-going JK TOOLS and validation are integrated into day-to-day data and information management protocols.

IV References:-

[1]. Arup kumar Bhattacharjee,Atanu Mallick,Arnab Dey .Sadananda B.JK TOOLS in Text Files

[2] Wikipedia Free Encylopedia

[3] R. Cody, "JK TOOLS 101," Proceedings for the Twenty-Seventh SAS User Group International Conference. Cary, NC: SAS Institute Inc

[4]. Dr. Mortadha M. Hamad and Alaa Abdulkhar Jihad, "An Enhanced Technique to Clean Data in the Data Warehouse". Computer Science Department. University of Anbar, Ramadi, Iraq.

[5]. Hasimah Hj Mohamed, Tee Leong Kheng, Chee Collin and Ong Siong Lee, "E-Clean: A JK TOOLS Framework for Patient Data". School of Computer Sciences. University Sains Malaysia Penang, Malaysia.