IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detection And Elimination Of Duplicate Files

[1]Mahesh Kumar, [2]Ms.Reetu
[1]MTech (CSE), [2]Asst. Prof. of CSE DEPTT
[1]Computer Science of Engineering
[1]RPS College of Engineering &Technology,Balana, Mahendergarh (Haryana)

Abstract: The problem of police investigation and eliminating pirated knowledge is one among the key issues within the broad space {knowledge of information} cleansing and knowledge quality in data warehouse. Many times, constant logical world entity might have multiple representations within the knowledge warehouse. Duplicate elimination is tough as a result of it's caused by many forms of errors like trade errors, and totally different representations of constant logical worth. Also, it's vital to sight associated clean equivalence errors as a result of an equivalence error might lead to many duplicate tuples. Recent analysis efforts have targeted on the difficulty of duplicate elimination in knowledge warehouses. This entails attempting to match inexact duplicate records, that area unit records that consult with constant real-world entity whereas not being syntactically equivalent. This paper chiefly focuses on economical detection and elimination of duplicate knowledge. The most objective of this analysis work is to sight actual and inexact duplicates by victimization duplicate detection and elimination rules. This approach is employed to enhance the potency of the info. The importance of information accuracy and quality has enhanced with the explosion of information size. This issue is crucial to confirm the success of any cross enterprise integration applications, business intelligence or data processing solutions. Police investigation duplicate knowledge that represent constant world object quite once during a sure dataset is that the beginning to confirm the info accuracy. This operation becomes additional sophisticated once constant object name (person, city) is drawn in multiple natural languages thanks to many factors together with orthography, trade and pronunciation variation, dialects and special vowel and consonant distinction and different linguistic characteristics. Therefore, it's tough to determine whether or not or not 2 grammar values (names) area unit various designation of constant linguistics entity. Up to authors' information, the antecedently planned duplicate record detection (DRD) algorithms and frameworks support solely single language duplicate record detection, or at the most bilingual. During this paper, 2 out there tools of duplicate record detection area unit compared. Then, a generic cross language primarily based duplicate record detection resolution design is planned, designed and enforced to support the big selection variations of many languages. The planned system style uses a lexicon supported phonetic algorithms and support totally different indexing /blocking Techniques to permit quick process. The framework proposes the utilization of many proximity matching algorithms, performance analysis metrics and classifiers to suit the variety in many languages names matching. The framework is enforced and verified through empirical observation in many case studies. Many Experiments area unit dead to check the benefits and downsides of the planned system compared to different tool. Results showed that the planned system has substantial enhancements compared to the well-known tools.

Keywords: Duplicate Record Detection, Cross Language Systems, and entity matching. Data Cleaning, Duplicate Data

I. Introduction

Data warehouses hold vast amounts of data that may be mined for analysis and decision-making. Data miners must not only evaluate data, but also prepare it in a format and state suitable for analysis. The actual data mining procedure is predicted to take only 10% of the time required for the entire knowledge discovery process. According to Jiawei, the previously time-consuming stage of preprocessing is critical for data mining. It's more than a laborious requirement: the preprocessing techniques utilized might have a significant impact on the results of the next stage, the actual use of a data mining algorithm.

According to Hans-peter, the role of the impact on the link between data preparation and data mining will continue to gain focus in the next years. Preprocessing will be a major concern and future trend in data mining in the next years. Data is combined or collected from numerous sources in a data warehouse. When combining data from numerous sources, the total amount of data grows, and data is duplicated. For the mining operation, the data warehouse could have terabytes of data. Data preparation is the first and most important phase in the data mining process. Data preparation is required to improve the accuracy of the mining result because 80 percent of mining efforts are spent on this.

As a result, data cleaning is critical in the data warehouse prior to the mining process. Because of data duplication and poor data quality, the outcome of the data mining process will be inaccurate. There are numerous known ways for detecting and eliminating duplicate data. However, the data cleaning process is very slow, and the time it takes to clear a big amount of data is very long. As a result, there is a need to minimize the time and speed of the data cleaning process, as well as to improve the data quality. When it comes to duplication detection, there are two factors to consider: accuracy and speed. The number of false negatives (duplicates you didn't categories as such) and false positives determines duplicate detection accuracy (non-duplicates which were classified as duplicates). A duplication detection and eradication rule is designed in this research study to manage any duplicate data in a data warehouse. Duplicate elimination is critical for determining which duplicates should be kept and which should be deleted. The primary goal of this study is to decrease the amount of false positives, speed up the data cleaning process, minimize complexity, and increase data quality. To achieve the goal, a high-quality, scalable duplicate elimination method was employed and evaluated on real datasets from an operational data warehouse.

Only one copy of exact duplicated records is kept in the duplicate removal process, while other duplicate records are deleted. To produce clean data, the elimination procedure is critical. Prior to the elimination procedure, the similarity threshold values for all of the records in the data set are calculated. For the elimination procedure, the similarity threshold values are crucial. Select all possible couples from each cluster and compare records inside the cluster using the selected attributes during the elimination phase. The majority of elimination processes merely compare records inside a cluster. Other clusters may occasionally have duplicate records with the same value as other clusters. With a modest increase in running time, this strategy can significantly minimize the likelihood of erroneous mismatches.

To find or detect duplicates and eradicate duplicates, apply the steps below.

i. Use the LOG table to find the threshold value.

ii. Determine the certainty factor

iii. Determine the data quality factor for each record. iv. Use the certainty factor, threshold value, and data quality factor to detect or identify duplicates.

v. Remove duplicate records depending on data quality, threshold value, amount of missing values, and field value ranges.

vi. Only keep one duplicate record with good data quality, a high threshold value, and a high certainty factor.

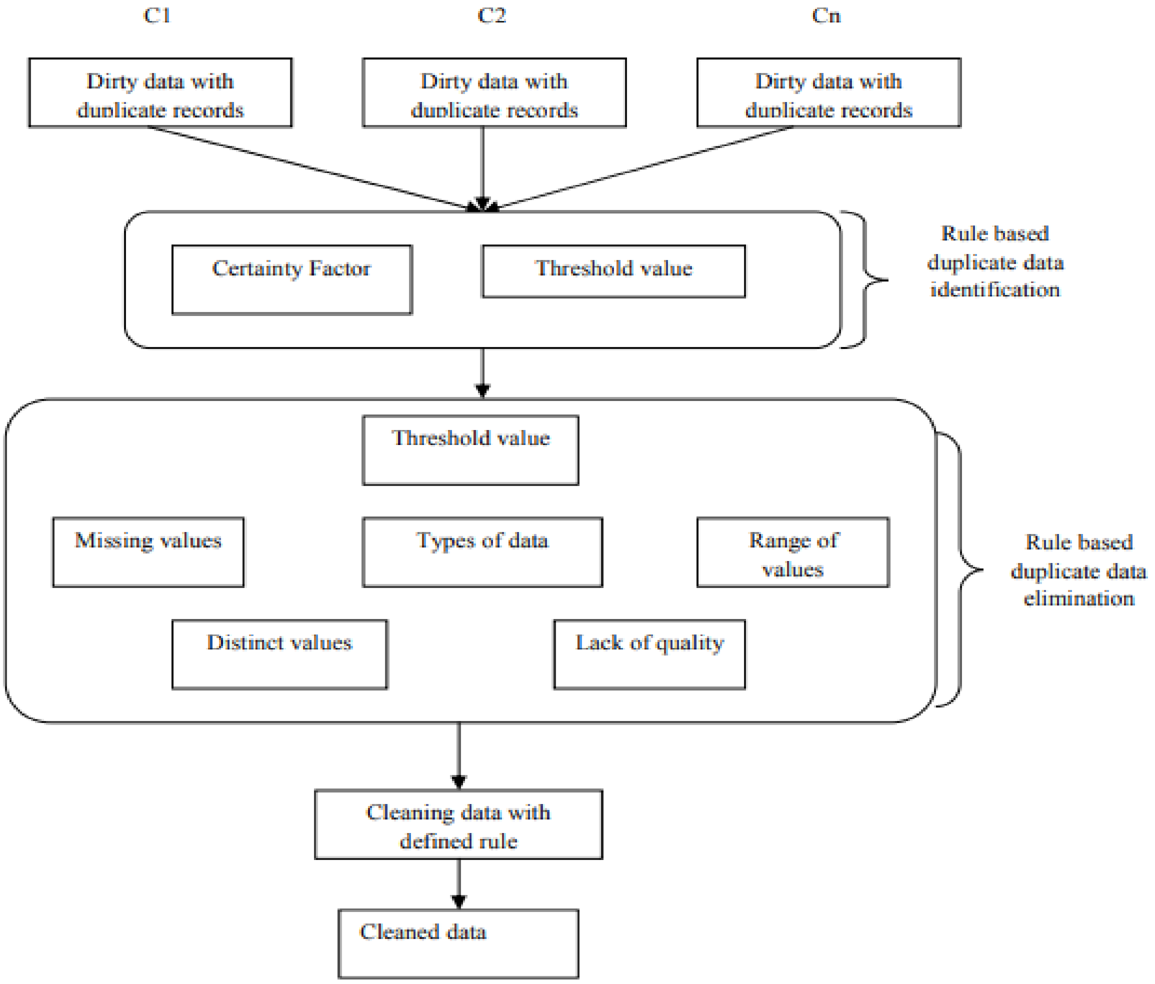
C1
C2
Cn
Dirty data with duplicate records
Dirty data with duplicate records
Dirty data with duplicate records
Certainty Factor
Threshold value
Rule based duplicate data identification
Threshold value
Missing values
Types of data
Range of values
Distinct values
Lack of quality
Rule based duplicate data elimination
Cleaning data with defined rule
Cleaned data

Figure 1. Framework for duplicate identification and elimination

Figure 1 depicts the framework for detecting and eliminating duplicate data. In this framework, there are two types of rules.

1) Identification of duplicate data;

2) Elimination of duplicate data the certainty factor and threshold value are used to identify or detect duplicate data in a duplicate data identification rule. Duplicate data removal rules are used to remove duplicate data using specific parameters and keep only one exact duplicate data.

LITERATURE REVIEW

Record In the categorization process, the pairs in the second, third, and fourth rows demonstrate disparities between the machine result and SME judgement. To improve machine classification results, the reasons for these disparities should be noted and transformed into rules. The following are possible reasons: R1: The machine did not recognise compound names that require a unique localised phonetic soundex for each language; R2: The machine swapped the first and last names, which is not acceptable in some languages; R3: The machine did not find dictionary entries for certain names when one of the records is a transliteration of the second; R4: The machine did not find dictionary entries for certain names when one of the records is a transliteration of the These reasons and SME comments that explain the disparities between machine and SME results are turned into rules and added to the language extension.

The confusion matrix is then used to classify the data. The TP, TN, FP, and FN measurements are then added together. Accuracy, Precision, and Recall will be among the other metrics computed. It's worth noting that converting the mismatch between the machine and the subject matter expert into a used defined rule and adding it to the language extension can help improve the situation.

Related Work:-

In the categorization process, the record pairs in the second, third, and fourth rows demonstrate disparities between the machine result and SME opinion. To improve machine classification results, the reasons for these disparities should be noted and transformed into rules. The following are possible reasons: R1: The machine did not recognize compound names that require a unique localized phonetic soundex for each language; R2: The machine swapped the first and last names, which is not acceptable in some languages; R3: The machine did not find dictionary entries for certain names when one of the records is a transliteration of the second. These SME comments and justifications for the disparities between machine and SME results are turned into rules and added to the language extension. In the confusion matrix, the data is then categorised. The TP, TN, FP, and FN metrics are then counted. Accuracy, Precision, and Recall will be among the other metrics calculated. It's worth noting that if the mismatch between the machine and the subject matter expert is converted into a used defined rule and added to the language extension, it can be improved.

DUPLICATE DATA IDENTIFICATION / DETECTION RULE

Duplicate record detection is the process of identifying different or multiple records that refer to one unique real world entity or object if their similarity exceeds a certain cutoff value. However, the records consist of multiple fields, making the duplicate detection problem much more complicated. A rule-based approach is proposed for the duplicate detection problem. This rule is developed with the extra restriction to obtain good result of the rules. These rules specify the conditions and criteria for two records to be classified as duplicates. A general if then else rule is used in this research work for the duplicate data identification and duplicate data elimination. A rule will generally be of the form:

```
if <condition >
then <action >
```

The action part of the rule is activated or fired when the conditions are satisfied. The complex predicates and external function references may be contained in both the condition and action parts of the rule. In existing duplicate detection and elimination method, the rules are defined for the specific subject data set only. These rules are not applicable for another subject data set. Anyone with subject matter expertise can be able to understand the business logic of the data and can develop the appropriate conditions and actions, which will then form the rule set. In this research work, the rules are formed automatically based on the specific criteria and formed rules are applicable for any subject dataset. In duplicate data detection rule, threshold values of record pairs and certainty factors are very important

REFERENCES

[1] Shintaro Yamamoto, Shinsuke Matsumoto,Sachio Saiki, and Masahide Nakamura Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan,Using Materialized View as a Service of Scallop4SC for Smart City Application Services (2014)

[2] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W. “Shared disk big data analytics with Apache Hadoop” (18-22 Dec. 2012)

[3] Kudakwashe Zvarevashe1, Dr. A Vinaya Babu, Towards MapReduce Performance Optimization: A Look into the Optimization Techniques in Apache Hadoopfor BigData Analytics (2014)

[4] Gartner: Hype cycle for big data, 2012. Technical report (2012)

[5] IBM, Zikopoulos, P., Eaton, C.:Understanding BigData: Analytics for Enterprise Class Hadoop and Streaming Data. 1st edn. McGraw-Hill Osborne Media,New York (2011)

[6] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., Tufano, P.: Analytics: The realworld use of big data. IBM Institute for Business Value—executive report, IBM Institute for Business Value (2012)

[7] Evans, D.: The internet of things—howthe next evolution of the internet is changing everything. Technical report (2011)