International Journal of Engineering & Technology, 7 (1.9) (2018) 171-172

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Survey of duplicate detection using progressive detection techniques

K. Venkatraman [1] *, A. Akila [2]

[1] *Research scholar Dept. of Computer science School of Computing Sciences Vels University*
[2] *Assistant Professor Dept. of Computer science School of Computing Sciences Vels University*
**Corresponding author E-mail: akila.scs@velsuniv.ac.in*

Abstract

Data is an important task in real world; the common data is represented and used in all the fields. The duplicate data is executed and displayed in scenario. The proposed work two types of techniques used first one Progressive Sort Neighbourhood Method (PSNM) and Progressive Blocking (PB). Progressive Sort Neighbourhood Method is used to deliver the exact input based output and the method will separate the input based keywords and check the similarity of the output data. The progressive blocking is to filter the irrelevant information, keywords based indexing and entry level filtering standard input is implemented based on user requirement.

Keywords: *Progressive Sort Neighborhood Method; Progressive Blocking; Duplicate Detection.*

1. Introduction

In day today life data is valuable one, data is commonly used to all members in various place in different situations. In environment like getting information server or web server, there exit more duplication of data.
The duplicate data sets available in database, the user based output is not a valid one. The data set reflect more relevant data's as to check the standard conditions and methods only to filter or extract the correct data the end user. As possible the way of standard values given in input time the end user query based or requirement based output displayed as quick as possible the thing to avoid the duplicate data's, so here I discuss or review how to avoid the duplication data and analyse the existing work to duplicate detection. In this existing work discussion about various merits, the highlighted point of the existing works is stated.

2. Literature review

The Duplicate record detection is set to clarify the irrelevant information, filtered data. The duplicate detection is a process to avoid the same output at the execution time. The duplicate detection to concentrate the value of data, keyword and index based filtering. Another filter is to check the data at input time. The duplicate detection methods detect the exact data in execution time [1].The progressive sorted neighbourhood method and progressive blocking algorithms increase the efficiency of duplicate detection for various environments with restricted performance time [2].
The PSNM and PB method executes correct and clear data. They reduce block duplicate values. The user friendly based ranking also implemented [3]. In today world, data are not pure some duplicate data accrued in merging multiple database depend on user queries. In the data cleaning, equation theory also verifies similar databases. The result is low cost at the user independent result from combining database [4].
The Pay go architecture executes the exact value as it is applied in various data in a large datasets [5]. The top k- set similarity joins algorithm export the pure & similar data the efficient algorithm that computes the correct data in a progressive manner [6].
System highlights that one of the most essential factors for effective and exact indexing for record linkage and de-duplication is the proper explanation of blocking keys. Because training data in the form of known true matches and non-matches is often not available in real world applications. It is commonly up to the domain and linkage experts to decide how such blocking keys are defined [7]. Spelling Correction is a process of detecting and sometimes providing suggestions for incorrectly spelled words in a text. Spell Checker may be stand-alone capable of operating on a block a text such as word processor, electronic dictionary. We get the exact inputs like a standard inputs get from user as address based on street, place, pin code [8].

3. Methodologies

a) Progressive Sort Neighborhood Method (PSNM)
Progressive Sorted Neighborhood Method Require: dataset reference D, sorting key K, window size W, enlargement interval size I, number of records N. Size p is calculation the partition size.
Array order size is calculated and the sort progressive method is applied, then for loop check size and data strength this type filter the data.
The comparison is executed using the compare (pair) function.
This function returns "true", a duplicate has been found and can be emitted.
PSNM evokes the look ahead (pair) method, progressively search for more duplicates in the current neighbourhood.
The search process not terminated early by the user, PSNM finishes when all intervals have been processed and the maximum window size W has been reached.
b) Progressive Blocking

Copyright © 2018 K. Venkatraman, Dr. A. Akila. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Progressive Blocking Requirement : dataset reference D, key attribute K, maximum block range R, block size S and record number Nprocedure PB (D, K, R, S, N), pSize calc PartitionSize(D).
The PB algorithm compares each record of the first block to all records of the second block.
Each block to check the duplicate data is available. The identified duplicate pairs'd' are then emitted.
The duplicate pairs to the current PB to later rank the duplicate density of this block pair with the density in other block pairs.

4. Analysis of existing result

In this review we check the data is not pure. More duplicate data created or available the database. It will create more redundant outputs. The alternate reason is no standard input is given in the entry level method. Sorted neighborhood approach also have some key based output, that key is to reduce the duplicate records but unless accuracy.
PSNM performs best on small and almost clean dataset. The spelling correction is a process of avoiding duplicate data's in a particular field (or) text boxes at the input time, but most dictionary indicate correct words. Some dictionaries display related data based duplicate or non-matching word or text. The process of record linkage or de-duplication is to get the quality & standard output from several or single database based on record matching.

5. Conclusion

The duplicate detection executes the exact data value. It may refer as to avoid the duplicate data based on user requirement, here we use PSNM, PB methods, the formation of PSNM & PB techniques implement large database on windows based system test then it will implement the online real world entity. In previous PSNM & PB methods only access the data sets, here we implement entry level filtering then use PSNM, PB methods it will give effective and efficient output to the end user. In future we will identify an approach to combine or group of datasets based search on duplicate values.

References

[1] Ahmed K. Elmagarmid, Vassilios S. Verykios, Member,"Duplicate Record Detection: A Survey". IEEE KDE, VOL. 19, NO. 1, JANUARY 2007.
[2] S. Ramya, C. Palaninehru ineering, "A Study of Progressive Techniques for Efficient Duplicate Detection". International Journal of Advanced Research in Computer Science and Software Engineering. Volume 5, Issue 11, November 2015. www.ijarcsse.com.
[3] Mohd Shoaib Amir Khan, "Progressive identification of duplicity".International Journal of Scientific and Research Publications, Volume 6, Issue 4, April 2016.
[4] Mauricio A. Hernandez, J.Stolfo, .Real World Data Is Dirty:Data Cleaning And The Merge/Purge Problem.
[5] Jayant Madhavan, Shawn R. Jeffery, Shirley Cohen, Xin (Luna) Dong,David Ko, Cong Yu, Alon Halevy,Google, Inc. "Web-scale Data Integration: You can only afford to Pay As You Go".
[6] Shawn R. Jeffery_ UC Berkeley Jeffery,Alon Y. Halevy "Pay-as-you-go User Feedback for Dataspace Systems".
[7] Top-k Set Similarity Joins Chuan Xiao Wei Wang Xuemin Lin Haichuan Shang
[8] Ritika Mishra1, Navjot Kaur2 "A Survey of Spelling Error Detection and Correction Techniques" International Journal of Computer Trends and Technology- volume4Issue3- 2013.
[9] Piotr Indyk1 A Small Approximately Min-Wise Independent Family of Hash Functions Received June 7, 1999.
[10] Uwe Draisbach Hasso Plattner, "A Generalization of Blocking and Windowing Algorithms for Duplicate Detection".
[11] Rupali Vairagade, Savitribai Phule "A Survey of Sorted Neighbourhood Indexing Technique for DeDuplication" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization). Vol. 3, Issue 12, December 2015.