ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

DATA EXPOSURE USING PSNM

[1]A.Kalaiarasi, [2]S.Jancy
[1]PG Student, Department of MCA, Sathyabama University, Chennai-600119, Tamil Nadu, India.
[2]Assistant Professor, Faulty of Computing, Sathyabama University, Chennai-600119, Tamil Nadu, India.
Email: [1]akalaiarasi21@gmail.com, [2]jancymtech11@gmail.com

Received on 14-02-2016 Accepted on 15-03-2016

Abstract

Detecting Duplication is one problem with alarming significance in a number of applications such as data mining, personnel data management, or management of client relationship. Detection of duplication is a process that identifies manifold depictions of the same actual physical entities. In the present day, duplicate diagnosis strategies involve processing very large datasets in a very short time, thus making it progressively difficult to maintain the datasets quality. In this study, we propose a progressive diagnosis algorithm method in addition to add Support Vector Machine (SVM) algorithm. It is proposed t use Progressive sorted neighborhood method (PSNM) for finding duplicate diagnosis approach in large datasetsand Progressive Blocking (PB) is performed on very dirty and large datasets. These two approaches improve the efficiency of duplicate diagnosis even in the case of very huge datasets. This suggested system enhances usage of Support Vector Machine algorithm to provide execution speed under limited time compared to other algorithms in the present system.

The existing progressive methods have to keep running for given time period and they are not capable of maximizing the efficiency for a specific time slot given. We have put forward three types of the progressive duplicate diagnosis methods known as hints. A hint can be defined as the probable good order of execution of comparisons for matching favorable record pairs before less favorable record pairs. Every cited hint produces static orders toward comparisons while missing the chance of dynamically adjusting the order of comparison during the run time, with relation to intermediate outcomes. For overcoming the said problem, it is suggested to make use of PB and PSNM along with add SVM algorithm. Their effective performance easily detects duplicated data and also compare in a short period of time, with high process speed. The technique proposed by us employs three categories: PSNM method for detecting duplicate dataset, PB Technique for blocking unauthentic users, and SVM for removing duplicate dataset under short period of time in the huge datasets. The said algorithms are able to double the efficiency with regard to time, compared to other traditional duplicate diagnosis. Assessments from experiments prove that the techniques we have proposed are more efficient and they function better the methods suggested earlier.

Keywords: Support Vector Machine algorithm (SVM)**,** Progressive sorted neighborhood method (PSNM), progressive blocking (PB), Dataset.

1. Background

The Role played by Databases is vital in the IT-based economy of the current period. Several systems and industries have to rely on databases accuracy for carrying out their operations. Data happens to be susceptible to different types of corruption like incorrect, missing, or inconsistent portrayals [1]. Real-life data are commonly found to be integrated from among multiple origins, and it is possible for the process of integration to lead to a range of errors [2]. Caused by careless data entry and data changes, certain errors like duplicate record occur. Finally, we suggest a method that solves progressive duplicate diagnosis problem found in real life with efficiency and gives result with accuracy.David Marmaros and Steven Euijong Whang [7] put forward that Entity Declaration (ER) as the trouble in identification of which particular records found in a database point to same entity. Practically, several applications have to resolve huge datasets with efficiency, but they may not need the result of ER to be accurate. For instance, the data about people found on the Web may be just too huge to be resolved completely using a rational amount of effort. Take for another example, real-time exercises may be unable to tolerate an ER processing which takes more than a stipulated amount of the time. This study analyzes how progress of ER can be maximized with a given amount of the efforts by making use of hints that furnish information about records that may potentially point to same real-life entity. Here, we suggest a group of strategies in order to build methods efficiently to use hints for maximizing number of duplicate records determined under certain restricted amount of effort. By making use of actual datasets, we describe the probable gains regarding our pay-when-you-go method compared to executing ER without the use of hints. An ER procedure is often found to be very expensive because of very huge datasets and estimate-intensive comparison of records. The suggested pay-when-you-go method for Entity Declaration (ER) has been given limited resources such as runtime and work, so we try to produce the maximum progress that is possible. [8] Ahmed K. Elmagarmid, Vassilios S. Verykios, PanagiotisG.Ipeirotis Duplicate records are not found to share common key and/or it is possible that they may contain errors which make the task of duplicate matching difficult. Errors may be introduced out of incomplete information, lack of sufficient standard formats, transcription errors, or even a combination of the said factors. In this study, we furnish a complete assessment of the research on detection of duplicate record. We cover similar metrics which are being used commonly for detecting similar such field entries. We also furnish a detailed group of algorithms for duplicate detection which can approximately detect the duplicate records present in a database. We include also several approaches to improve efficiency and capability to cater to approximate algorithms of duplicate detection. We end with our coverage of the present tools and a brief review about certain big and open challenges in the particular area. The issue that we took up for the study is known for over five decades as record connection or as record matching trouble among the statistics circle. Record matching aims at identifying records present in same or various databases which point to same real-life entity even in the case of the records not being identical. [9] The issue of identification of matching records approximately in databases proves to be an imminent step regarding data integration and data cleaning processes. Most of the present methods are found to rely on general or manually regulated distance metrics in the estimation of similarity of the possible duplicates. In this study, we have presented a structure to improve duplicate diagnosis by making use of trainable ranges of the textual similarity. We suggest employing acquirable text distance operations for every field in the database, and prove that such ranges are competent in adapting the particular concept of similarity which is found to be suitable for the given fields domain. We put forward two acquirable text similarity ranges appropriate for this operation: an extensive alternative of the acquirable string revise distance, and one innovative vector-space oriented measure which uses a Support Vector Machine (SVM) with regard to training. Results of experiments on a varied range of datasets prove that the framework proposed by us is capable of accurately detecting duplicates and is better when compared to other existing traditional strategies.

2. Proposed Work

This paper aims at detecting and removing duplicate data over the short possible time with regard to real-life entities. The processes in this study involve users to get registered first in the admin side and then uploading the data in the data storage. While uploading the data making use of PB and PSNM, we suggest two innovative progressive duplicate diagnosis algorithms, namely, Progressive Blocking (PB) that functions well in the case of very dirty and huge datasets, Progressive Sorted Neighborhood Method (PSNM) that can perform best in the case of almost clean and small datasets. In addition, this study has added Support Vector Machine (SVM) for reducing the time duration in duplicate diagnosis with regard to longer period of time consumption. This helps in speeding up the process of detection of matching data and duplicate data. The above said two methods help in greatly enhancing efficiency of duplicate diagnosis even in the case of datasets that are very large. PSNM classifies input data by making use of pre-described sorting key. It can only compare records which are present within a given window of records available in the classified order. The algorithm of PSNM differs from this in that it dynamically changes the order of execution of comparisons founded on intermediate results. Blocking algorithms allocate each entry to some fixed set of similar entries (blocks) and then they compare all the record pairs within these sets. Using SVM for huge dataset values assists the process by reducing the time consumed, thus avoiding the time-related drawback; it also offers good and reliable matching process. Progressive Blocking happens to be an innovative method which builds over an equidistant obstructing approach and the consecutive expansion of the blocks. In the end, the output given is found to be accurate and the method is efficient in the real-life entities.

2.2 Registration and Login Page:

In this study, the first step is the registration process. It is required to register any user after uploading the data. The process of registration happens to be an individual process with regard to progressive duplicate diagnosis algorithm. The process of registration involves certain verification or authentication details and after getting registered successfully, it is possible to go to the login page of the database. For accessing the login page, user ID and secure password are necessary and after logging in successfully, any data can be uploaded. Data may be considered with regard to size as it can be available in any of the various sizes possible. This data may be authentic or duplicate in the database. Next, the data will be read in the database. All these processes are handles in the admin side.

2.3 Overall Architecture

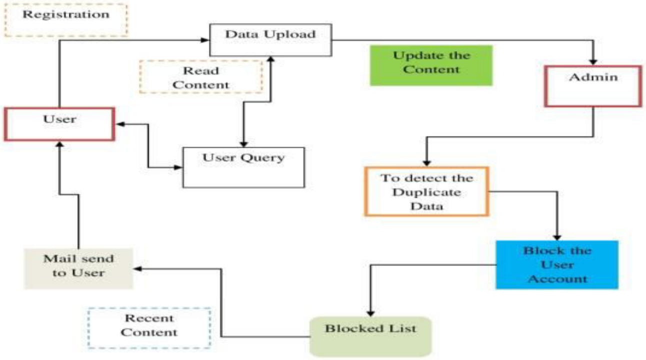


Figure-1: Overall Architecture.

2.4 PSNM Algorithm and Duplicate Detection

while employing PSNM algorithm for detecting duplicate data, first upload any given data for checking, data that was previously uploaded will be checked, all the processes will run on admin side. When uploading the data in the database, it will check all details of the data such as how we upload, size of data, and considering these, data matching will be done first. This process will continue over all the uploaded data and finally it will conclude whether duplicate data is present in it or not. This study suggests 75% matching when duplicate data is available. Progressive Sorted Neighborhood Method (PSNM) can be use for sorting content in required order.

2.4.1 Algorithm 1 Progressive Sorted Neighborhood:

Require: dataset reference D, sorting key K, window size

W, enlargement interval size I, number of records N

```
procedure PSNM (D, K, W, I, N)
    pSizecalcPartitionSize(D)
    pNumdN=(pSize□ W + 1)e
    array order size N as Integer
    array recs size pSize as Record
    order sort Progressive (D, K, I, pSize, pNum)
    for currentI 2 to dW=Ie do
        for currentP 1 to pNumdo
            recs load Partition (D, currentP)
            for dist 2 range (currentI, I, W) do
                for i   0 to jrecsj□ dist do
                    pair   hrecs [i]; recs[i + dist]i
                    if compare (pair) then
                        emit (pair)
                        lookAhead(pair)
```

2.5 PB Algorithm and Blocked Untruth Account

Progressive Blocking (PB) is used for comparing data with the previous content (all existing content) after uploading a given content. It is possible to block the accounts of untrustworthy users who can possibly change the given content. Duplicate content with relation to real life entities are avoided.

2.5.1 Algorithm 2 Progressive Blocking

Require: dataset reference D, key attribute K, maximum block range R, block size S and record number N

```
procedure PB(D, K, R, S, N)
    p Sizecalc Partition Size(D)
    b Per Pbp Size=Sc
    b Numd N=Se
    pNumdbNum=bPerPe
    array order size N as Integer
    array blocks size bPerP as hInteger, Record [ ]i
    priority queue bPairs as hInteger,Integer,Integeri
    bPairs   fh1; 1; _i ; ... , hbNum; bNum; _ig
    order   sortProgressive(D, K, S, bPerP, bPairs)
    for i   0 to pNum□ 1 do
        pBPsget(bPairs, i _ bPerP, (i+1) _ bPerP)
        blocks   loadBlocks(pBPs, S, order)
        compare (blocks, pBPs, order)
    while bPairs is not empty do
        pBPsfg
        bestBPstakeBest(bbPerP=4c, bPairs, R)
        for bestBP 2 bestBPs do
            if bestBP[1] □bestBP[0] < R then
                pBPspBPs[ extend(bestBP)
        blocks   loadBlocks(pBPs, S, order)
        compare(blocks, pBPs, order)
        bPairsbPairs[ pBPs
```

```
procedure COMPARE(blocks, pBPs, order)
    for pBP 2 pBPs do
        hdPairs, cNumicomp(pBP, blocks, order)
        emit(dPairs)
        pBP[2]   jdPairsj / cNum
```

2.6 SVM Technique and Reduce Duplicate Data

In this approach, matching process is fast and the time needed for detecting process is high. Data can be uploaded after checking for any duplication. SVM method is used for detecting quickly whether duplicate data is present or not. We suggest enhancing the process of matching by using SVM technique, as it helps in detecting duplicate under a very short time. Matching process involves checking various sized data. Consider for example, when 2kb data is being uploaded in the database, it first checks order-wise pairing such as 2, 4, 6 kb, and so on. All the uploaded data get matched and when half the data equals the above data, then it is considered as duplicate data. As SVM contains benefits in matching all the data in a short period of time, it becomes possible to handle information about duplicates in large dataset easily.

2.7 Mailling Process and User Verified

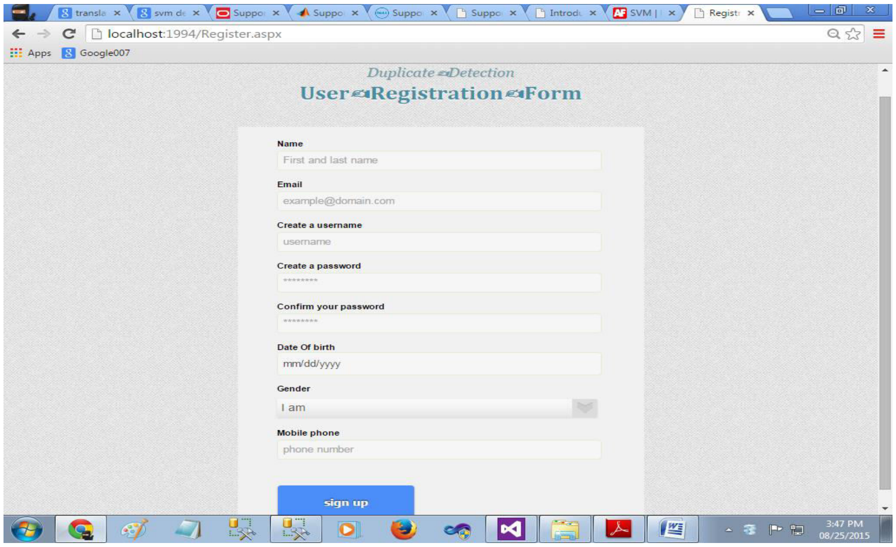
In this study, if user has any query, it has to be sent to the admin side and admin will reply to user. In case any user account gets blocked, the admin can send email to such blocked user. Ultimately, the list of blocked users will be shown on the user and admin sides. If any of the user account has been blocked, it is possible to get it verified.

3. Result and Discussion

The process that identifies multiple portrayals of the same real life entities is known as duplicate diagnosis. Duplicate detecting techniques have to process increasingly huge datasets in very short time. Hence, it becomes difficult to maintain the quality of the dataset. There are certain algorithms that significantly enhance the efficiency of identifying duplicates, when there is limitation on execution time.

Progressive duplicate diagnosis helps in identifying most of the duplicate sets in the early stage of the detection process. In order to reduce the overall time required for finishing the complete process, try reducing average time subsequent to finding a duplicate. In this analysis, SVM method is used to enhance the process of data matching as well as detection of duplicate data under short period of time in the case of huge dataset. The inference and review define two classifications that are formed from the result of the experiment. We have considered three models, namely, mailing process, content upload, and registration.

3.1 User Registration:



Duplicate Detection

User Registration Form

Name
First and last name

Email
example@domain.com

Create a username
username

Create a password
••••••••

Confirm your password
••••••••

Date Of birth
mm/dd/yyyy

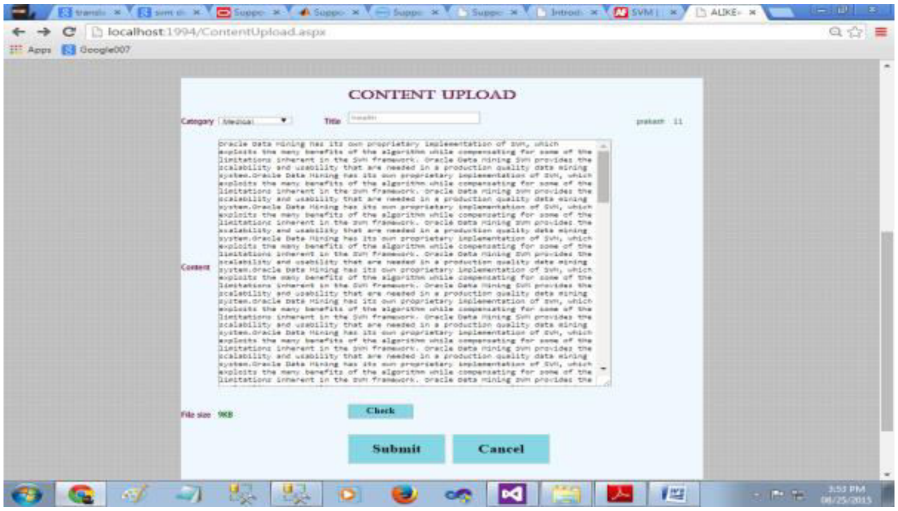
Gender
I am

Mobile phone
phone number

Figure-2: User Registration Form.

The First process is the registration for uploading the content in database. The above Figure 2 shows the registration form, which contains username, mail id, password and other verification details.

3.2 Content Upload



CONTENT UPLOAD

Category Medical

Title Health

prakash 11

Content

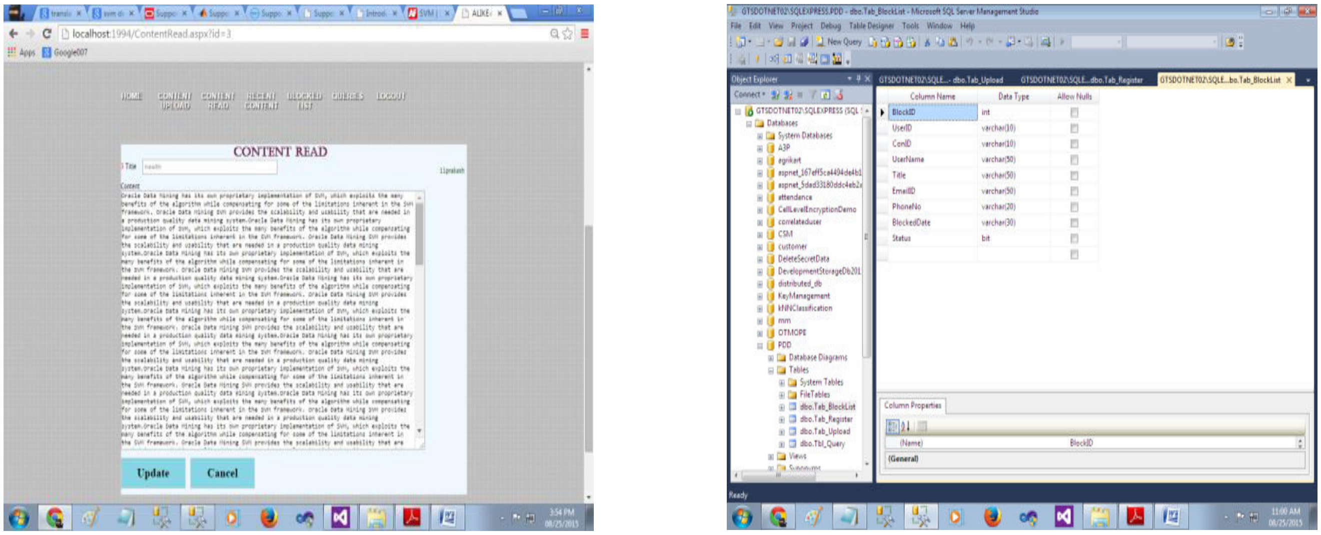
```
oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
scalability and usability that are needed in a production quality data mining
system.Oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
scalability and usability that are needed in a production quality data mining
system.Oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
scalability and usability that are needed in a production quality data mining
system.Oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
scalability and usability that are needed in a production quality data mining
system.Oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
scalability and usability that are needed in a production quality data mining
system.Oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
scalability and usability that are needed in a production quality data mining
system.Oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
scalability and usability that are needed in a production quality data mining
system.Oracle Data Mining has its own proprietary implementation of SVM, which
exploits the many benefits of the algorithm while compensating for some of the
limitations inherent in the SVM framework. Oracle Data Mining SVM provides the
```

File size 9KB

Figure-3: Upload content in database.

The above Figure3shows the successful login of the authorized user and uploading data in database, this data may be duplicate or true data. Only after registration the content can be uploaded by authorized user. Then the user can login into the account to upload content, and Read the verified content.

3.3 Duplicate Detection and Mailing


localhost:1994/ContentRead.aspx?id=3
HOME
CONTENT UPLOAD
CONTENT READ
RECENT CONTENT
BLOCKED LIST
QUERIES
LOGOUT
Update
Cancel
GTSDOTNET02\SQLEXPRESS.PDD - dbo.Tab_BlockList - Microsoft SQL Server Management Studio
File Edit View Project Debug Table Designer Tools Window Help
New Query
Object Explorer
GTSDOTNET02\SQLE...- dbo.Tab_Upload
GTSDOTNET02\SQLE...dbo.Tab_Register
GTSDOTNET02\SQLE...bo.Tab_BlockList
Column Properties
Ready

CONTENT READ

Title: health

11prakash

Content

Oracle Data Mining has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. Oracle Data Mining SVM provides the scalability and usability that are needed in a production quality data mining system.Oracle Data Mining has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. Oracle Data Mining SVM provides the scalability and usability that are needed in a production quality data mining system.Oracle Data Mining has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. Oracle Data Mining SVM provides the scalability and usability that are needed in a production quality data mining system.Oracle Data Mining has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. Oracle Data Mining SVM provides the scalability and usability that are needed in a production quality data mining system.Oracle Data Mining has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. Oracle Data Mining SVM provides the scalability and usability that are needed in a production quality data mining system.Oracle Data Mining has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. Oracle Data Mining SVM provides the scalability and usability that are needed in a production quality data mining system.Oracle Data Mining has its own proprietary implementation of SVM, which exploits the many benefits of the algorithm while compensating for some of the limitations inherent in the SVM framework. Oracle Data Mining SVM provides the scalability and usability that are

GTSDOTNET02\SQLEXPRESS (SQL
- Databases
  - System Databases
  - A3P
  - egrikart
  - aspnet_167eff5ca4494de4b1
  - aspnet_5ded33180ddc4eb2a
  - attendance
  - CellLevelEncryptionDemo
  - correlateduser
  - CSM
  - customer
  - DeleteSecretData
  - DevelopmentStorageDb201
  - distributed_db
  - KeyManagement
  - kNNClassification
  - mm
  - OTMOPE
  - PDD
    - Database Diagrams
    - Tables
      - System Tables
      - FileTables
      - dbo.Tab_BlockList
      - dbo.Tab_Register
      - dbo.Tab_Upload
      - dbo.Tbl_Query
    - Views
    - Synonyms

Column Name	Data Type	Allow Nulls
BlockID	int	□
UserID	varchar(10)	□
ConID	varchar(10)	□
UserName	varchar(50)	□
Title	varchar(50)	□
EmailID	varchar(50)	□
PhoneNo	varchar(20)	□
BlockedDate	varchar(30)	□
Status	bit	□
		□

(Name) BlockID

(General)

Figure-4: Duplicate detection and block list details.

Figure 4 show the read content and block list details, block list details ismaintained in both side between user and admin. Block list details display to untruth user and after sent mail to truth user mail id. True user get id after login page and read original content in database.

4. Conclusion and future Enhancement:

We have proposed PB and PSNM along with SVM approach regarding speed in time to access duplicate data. Time consumption issue happens to be very vital in the database processes. In this study, we have suggested SVM as the enhancing technique for future in matching process and quick detection in PSNM. The method put forward includes first registering by making use of the registration form, then logging in for uploading the data, and then checking whether there is data duplication or not by employing PB and PSNM. SVM, when used for matching process forms an easy and fast way for detecting duplicate data. I t also overcomes the problem of time duration. In using PB approach for the detection of duplicate data, the method involves blocking the account subsequent to sending email to the particular mail ID. The introduced method proves to be accurate and it gets better results when compared with the present SNM strategies that are available. To conclude, the proposed method includes efficiency and quality and produces accurate results. Future enhancement in accuracy and high quality is possible by employing SVM and PB strategies which will produce reliable result and also accuracy and quality in huge databases

5. References:

1. N. Swartz. Gartner warns firms of ’dirty data’. Information Management Journal, 41(3), 2007.

2. X. L. Dong and D. Srivastava, “Big data integration”, 2013. *PVLDB*, 6(11):1188–1189.

3. For big-data scientists, ’janitor work’ is key hurdle to insights. http://www.nytimes.com/2014/08/18/technology/forbig-data-scientists-hurdle-to-insights-is-janitor-work.html.

4. H. Park and J. Widom. Crowdfill: June 22-27, 2014,” Collecting structured data from the crowd”. In International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, , pages 577–588, 2014.

5. S. Ramya and C. Palaninehru, November 2015 , “A Study of Progressive Techniques for Efficient Duplicate Detection” ,Volume 5, Issue 11, ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com.

6. P. Indyk, MAY 2015.“A small approximately min-wise independent family of hash functions,”in Proc. 10th Annu. ACM-SIAM Symp.Discrete Algorithms, 1999, pp. 454–456.1328 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5.

7. S. E. Whang, D. Marmaros, and H. Garcia-Molina, May 2012 “PAY-AS-YOU-GO Entity Resolution,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 5,pp. 1111–1124.

8. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, Jan. 2007 “Duplicate record detection: A survey,” IEEE Trans. Knowl. Data Eng., vol. 19,no. 1, pp. 1–16.

9. Mikhail Bilenko and Raymond J. Mooney, August, 2003,”Adaptive Duplicate Detection Using Learnable String Similarity Measures”,Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discoveryand Data Mining (KDD-2003), Washington DC, pp.39-48.

10. Wan-Lei Zhao, Chong-Wah Ngo¤, Hung-Khoon Tan, Xiao Wu, “Near-Duplicate Keyframe Identification with Interest Point Matching and Pattern Learning.

Corresponding Author:
A.Kalaiarasi*,
Email: *akalaiarasi21@gmail.com*