Data Dashboard-Integrating Data Mining with Data Deduplication

Vitasta Abrol
Department of Information Technology
MIT-COE, Kothrud, Pune, India

Jyoti Malhotra
Department of Information Technology
MIT-COE, Kothrud, Pune, India

ABSTRACT

Many applications deal with huge amount of data and that scattered data needs to be transformed into something relevant and meaningful. To make sense of such data is the need of many applications and areas of technology. The data that is already present is very huge, noisy and has a complex structure. We are working on the idea of integrating data mining with data deduplication. Data Dashboard is a tool which can take complex data involving various dimensions and simultaneously uses data deduplication algorithms that help in removing redundancy in the data up to 95%. Thus it provides high reliability, low disk space and high throughput. The tool then uses this data mined information and deploys it on dynamic platform such as web which provides ease to user to access huge database.

Keywords

Data mining, Data Deduplication, Clustering, Hashing, API

1. INTRODUCTION

Let us now start with the concept of Data mining [1]. It is a field which incorporates features of both computer technology and information statistics and basically aims on extracting desired information from required set of data. It also attempts to explore some fascinating and interesting patterns in large databases. In short, this field uses the interesting concepts of intelligence, neural techniques, statistics, and database systems. Data mining is an interesting process that can extract some information from a huge data set and then converts it into a form that can be understood or used for some purpose in future.

Mining of data can be referred as a data processing technique which includes collection of raw data, distillation of desired or useful information from that raw data and later stages concentrate on processing or cleaning of that distilled data. If we have to sum it in some words, we can say data mining is discovery of data or exploring something new in the old data.

The varied features of data mining process also include detection of interesting patterns in data records. This feature is of extreme importance in marketing. Discovered patterns can be used to analyze market conditions and finding out the needs of customers. For example: -

Interesting patterns were found in Wal-Mart Sales data. Their sales data inferred that people, who bought coke, bought some kind of potato chips also. So to boost the sales they decided to keep coke and some eatables together. Thus we can conclude that pattern extraction can be of extremely highly utility.

'**Data Dashboard**' is a tool for Comprehensive Metrics (Standards of measurement by which efficiency, performance, progress or quality of a plan, process, or product can be assessed.).The most important purpose of this tool is to represent the information in graphical format using the process of Data Mining and also uses Data Deduplication techniques for removing redundant data. For this tool, the prerequisites are listed as-

- User should be familiar with the FileMaker Pro tool which is used as a database for this tool.
- User should also have the knowledge about the Web Siebel Interface which is used as database for Software Progress Report (SPRs) related data.
- User must have knowledge of Java, JSP and MySQL.
- Configured environment for running the tool.
- The user should know tool's purpose and complete product understanding.

We are dealing with high dimensional data in our tool. Let us explore bit about that form of data.

2. HIGH DIMENSIONAL DATA

Clustering high-dimensional data [2] is the cluster analysis of data having dimensions that can vary between tens to thousands.

High dimensional data can actually be defined as data having a large number of attributes. These attributes can be named as dimensions of a particular data entry.

This type of data is commonly encountered these days in variety of areas. It can be seen in the area of medicine, since any molecule or microarray is able to produce huge measurements. It can also be seen in the database of any school or college, where a data entry such as student will have a large number of attributes associated with it. Here we are extending our tool to deal with such kind of data.

Now let us proceed to the architecture of tool.

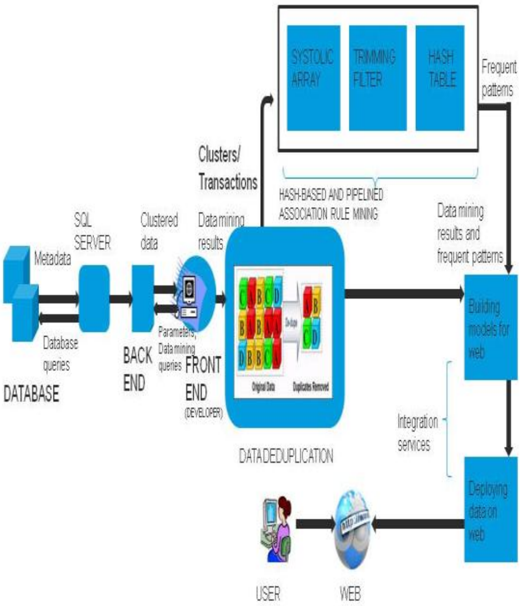


Fig. 1: Architecture of data dashboard.

3. ARCHITECTURE OF DATA DASHBOARD

Figure 1 shows the architecture of data dashboard. The architecture shows how data mining is linked with data deduplication. The data mining is performed on the initial state of database. The query language which is commonly used for the relational databases is SQL, since it allows efficient manipulation, retrieval and extraction of the data which is stored in the tables. It also assists in the calculation of aggregate functions such as, sum, max, min, average and count. For instance, an SQL query to select the test cases grouped by category would be:

SELECT count (*) FROM Test Cases WHERE type=TC_Name GROUP BY category.

Data mining algorithms which use relational databases are more versatile than those data mining algorithms which are specifically written only for flat files. The former data mining algorithms take advantage of dealing with the structure inherent to the relational databases**.** The architecture is explained as follows-

The tool Data Dashboard works in various steps which are mentioned below-

Level 1- In level 1, capable or desired useful information is extracted from a huge data set by using various data mining algorithms. A database can be used as raw data. The useful data extracted is sent to another level for processing or cleaning.

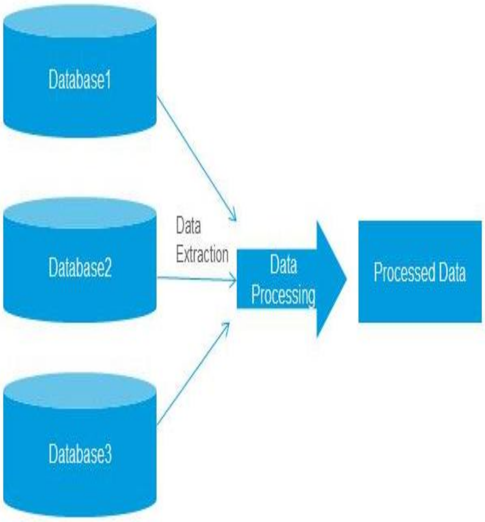


Fig. 2: Processes in level 1

Level 2- The most informative data is selected by eliminating irrelevant and redundant ones. Data Deduplication algorithms are used for this level. It can provide efficiency of redundancy detection up to 95%.

After Deduplication Techniques we apply *HAPPI* (HASH-based and pipelined) architecture[3] which has can do hardware-enhanced association rule mining. Various associations among the data are discovered using this architecture.

Level 3-Next level is to deploy the data on web which can be accessed by users .Various Interfaces or Application interfaces (APIs) for other languages and systems are used for deploy data on a dynamic platform.

Level 4- Integration of *DATA DASHBOARD* with various applications. This tool Data dashboard can efficiently be used for extracting potential or required information in variety of applications.

4. CLUSTER ANALYSIS

Cluster analysis [5] is a type of data mining process that basically works on grouping a set of objects **.**The objects are grouped on the basis of the relations or on basis of any mathematical calculations. The closely associated objects form one cluster. Working upon the same objective various clusters are formed and all the objects are converted into clusters. It is an important task of explorative mining of data and can be efficiently used in variety of fields that include pattern analysis, image analysis, retrieval of information and many more such as bioinformatics.

Clustering when used for mining high-dimensional data can be described as making clusters of data having large dimensions. The dimensions may vary from a few tens to many thousands. Such high-dimensional data can often be seen in fields such as medical field, where DNA and RNA technology are able to give us a huge number of measurements at once. In that case number of dimensions produced is really huge.

There are many clustering algorithms available. We will discuss one clustering algorithm and compare the results with another clustering algorithm to find out which algorithm will work better.Initailly we start with density based clustering algorithm.

4.1 Density Based clustering algorithm

Density based clustering algorithm is an efficient algorithm which helps in finding nonlinear shapes or structures based on the concepts of density. Most widely used density based algorithm is known as Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

Algorithmic steps of DBSCAN clustering algorithm:

Let X = {x1, x2, x3, xn} be a set of data points. DBSCAN requires two parameters: ε (epsilon) and the minimum number of points required to form a cluster (minimum Points).

1) We start with a random starting point which has not been initially visited.

2) Then the neighborhood of that point is extracted using ε. (Neighborhood means all those points which are within the ε distance from that point).

3) Then clustering process starts only if there is sufficient neighborhood around that particular point and after that point is marked as visited. Else if there is less neighborhood this point is labeled as noise.

4) If a point is discovered to be a part of the some cluster then its ε neighborhood also becomes the part of that particular cluster. The above algorithm is repeated from step 2 for all ε neighborhood points. This is repeated again and again until all points in the particular cluster are determined.

5) A new unvisited point is again selected, retrieved and processed which leads to the discovery of a further cluster or noise.

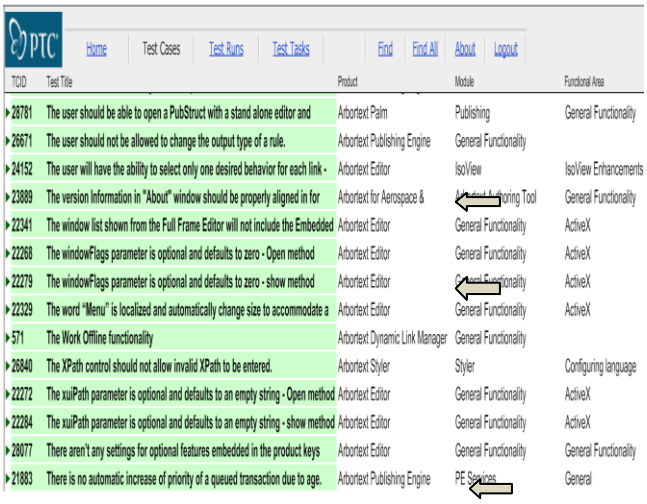
6) This process is repeated again until all points are marked as visited or they form some cluster or noise.

4.1.1 Advantages

- Does not need initial idea and specification of number of clusters.
- It is able to identify noise data also while clustering.

4.1.2 Disadvantages

- This algorithm is not efficient in case of varying density clusters.
- Does not work well in case of high dimensional data also.



TCID	Test Title	Product	Module	Functional Area
28781	The user should be able to open a PubStruct with a stand alone editor and	Arbortext Palm	Publishing	General Functionality
26671	The user should not be allowed to change the output type of a rule.	Arbortext Publishing Engine	General Functionality	
24152	The user will have the ability to select only one desired behavior for each link -	Arbortext Editor	IsoView	IsoView Enhancements
23889	The version Information in "About" window should be properly aligned in for	Arbortext for Aerospace &	Arbortext Authoring Tool	General Functionality
22341	The window list shown from the Full Frame Editor will not include the Embedded	Arbortext Editor	General Functionality	ActiveX
22268	The windowFlags parameter is optional and defaults to zero - Open method	Arbortext Editor	General Functionality	ActiveX
22279	The windowFlags parameter is optional and defaults to zero - show method	Arbortext Editor	General Functionality	ActiveX
22329	The word "Menu" is localized and automatically change size to accommodate a	Arbortext Editor	General Functionality	ActiveX
571	The Work Offline functionality	Arbortext Dynamic Link Manager	General Functionality	
26840	The XPath control should not allow invalid XPath to be entered.	Arbortext Styler	Styler	Configuring language
22272	The xuiPath parameter is optional and defaults to an empty string - Open method	Arbortext Editor	General Functionality	ActiveX
22284	The xuiPath parameter is optional and defaults to an empty string - show method	Arbortext Editor	General Functionality	ActiveX
28077	There aren't any settings for optional features embedded in the product keys	Arbortext Editor	General Functionality	General Functionality
21883	There is no automatic increase of priority of a queued transaction due to age.	Arbortext Publishing Engine	PE Services	General

Figure 3- Results after performing density clustering algorithm.

Figure 3 shows the results and the clusters formed after density clustering algorithm. After we achieved results for Density Based Clustering, it was seen that it wasn't much efficient to group the entries properly. Some random entries appeared in between the groups. The arrows in above figure point to those random entries in between clustered data. Also it wasn't efficient while dealing with different and large dimensions of the data. It couldn't group all the attributes correctly for all the entries. It failed while dealing with high dimensional data. Thus we moved onto using another algorithm for clustering and to check whether it was able to deal efficiently with the high dimensional data. We thus deployed K-means clustering algorithm. [6]

4.2 K-means clustering algorithm

In data mining, k-means clustering [6] is an algorithm of cluster analysis which works on the rule of finding centroid and the nearest mean. It partitions datasets into different clusters. Each unit of dataset belongs to one specified cluster having the nearest mean. This results in a division of the data sets into cells. A cluster centroid can actually be defined as the mean or median of the points in its cluster. The centroid chosen for finding the clusters is actually the mean or median of different points and the "nearness" of the clusters or the points is determined by using a distance or similarity function.

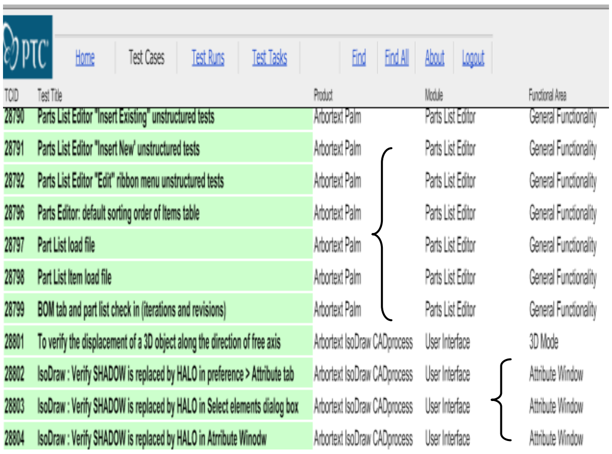
Basic steps in K-means Algorithm for finding K clusters-

1. Select any K points from the total data set as the initial centroids.

2. Assign a centroid which is closest to that particular point. Do this for all points.

3. Compute the centroid again so as to verify that the chosen cluster centroid is accurate. Do this for each cluster.

4. Repeat the above steps 2 and 3 till the point centroids don’t change (or change very little).



TCID	Test Title	Product	Module	Functional Area
28790	Parts List Editor "Insert Existing" unstructured tests	Arbortext Palm	Parts List Editor	General Functionality
28791	Parts List Editor "Insert New' unstructured tests	Arbortext Palm	Parts List Editor	General Functionality
28792	Parts List Editor "Edit" ribbon menu unstructured tests	Arbortext Palm	Parts List Editor	General Functionality
28796	Parts Editor: default sorting order of Items table	Arbortext Palm	Parts List Editor	General Functionality
28797	Part List load file	Arbortext Palm	Parts List Editor	General Functionality
28798	Part List Item load file	Arbortext Palm	Parts List Editor	General Functionality
28799	BOM tab and part list check in (iterations and revisions)	Arbortext Palm	Parts List Editor	General Functionality
28801	To verify the displacement of a 3D object along the direction of free axis	Arbortext IsoDraw CADprocess	User Interface	3D Mode
28802	IsoDraw : Verify SHADOW is replaced by HALO in preference > Attribute tab	Arbortext IsoDraw CADprocess	User Interface	Attribute Window
28803	IsoDraw : Verify SHADOW is replaced by HALO in Select elements dialog box	Arbortext IsoDraw CADprocess	User Interface	Attribute Window
28804	IsoDraw : Verify SHADOW is replaced by HALO in Atrribute Winodw	Arbortext IsoDraw CADprocess	User Interface	Attribute Window

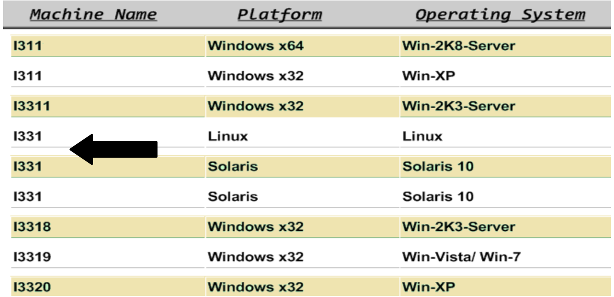
Figure 4- Results after performing k-means clustering algorithm.

Figure 4 shows the results and the clusters formed after k means clustering algorithm. After deploying k-means clustering we could conclude that it worked much efficiently with high dimensional data as it was able to cluster/group all the attributes of the given entries at the same time. In the above figure we can see clearly that all the attributes or dimensions are clustered nicely defining visible boundaries between those clusters.

5. DATA DEDUPLICATION

Data deduplication [7] is a technology becoming increasingly popular these days which works on reducing the undesired data by eliminating redundant copies of data. Multiple copies of same data can be created during system backup or e-mail attachments distributed to multiple users or when we are sharing documents, music and projects etc. Data Deduplication works on this kind of data and identifies the duplicates present in the system. We refer to the reduction in data footprint size due to deduplication. Various categories of data deduplication algorithms include file based, block based or hash based algorithms. The advantages include improved storage efficiency and cost savings, as well as bandwidth minimization for less-expensive and faster offsite replication of backup data.

The increasing volume of digital world and the rate at which information available in digital media is growing can definitely act as a bottleneck. A major problem area is the presence of replicas or redundant information in the system.



Machine Name	Platform	Operating System
I311	Windows x64	Win-2K8-Server
I311	Windows x32	Win-XP
I3311	Windows x32	Win-2K3-Server
I331	Linux	Linux
I331	Solaris	Solaris 10
I331	Solaris	Solaris 10
I3318	Windows x32	Win-2K3-Server
I3319	Windows x32	Win-Vista/ Win-7
I3320	Windows x32	Win-XP

Figure 5. - Duplicate entries arising in the system after data mining and clustering.

We are concentrating on deploying hash based data deduplication for removing redundancies from the mined data.

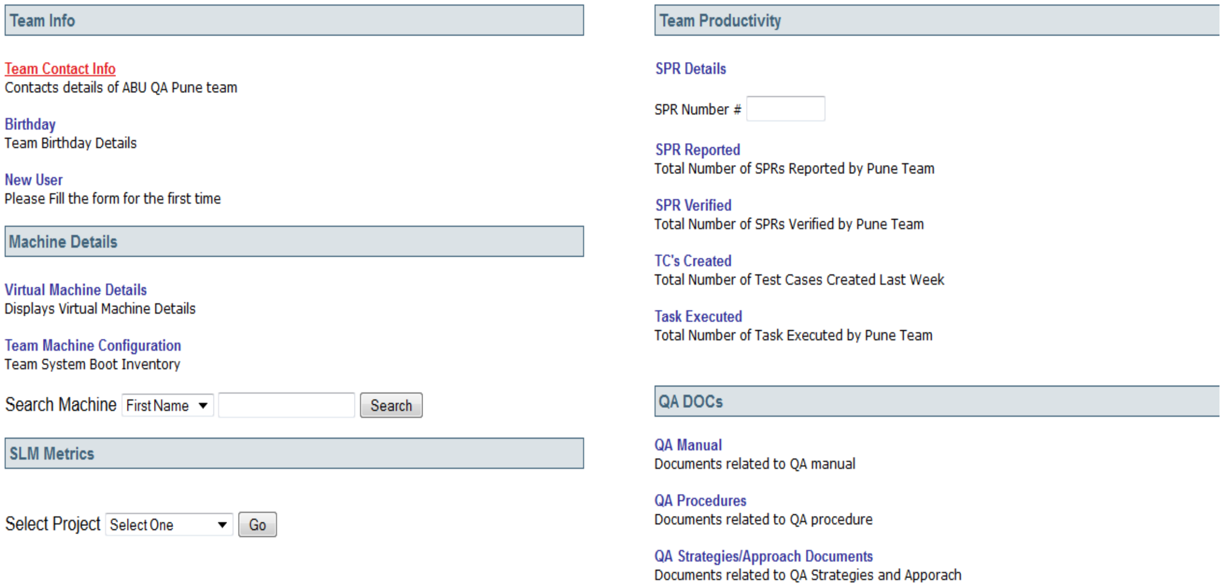
In hash based technique, each data entry will have a unique hash tag which is calculated by identifying the attributes of each data entry. The hash of each data entry is compared with other records in the system and only unique hash tags are stored in the system. Thus unnecessary data is eliminated from the system.

In hash based algorithm we slice data into chunks. Since we are dealing with high dimensional data having large number of attributes it becomes important to slice the data into chunks. After slicing the data into chunks, we generate hash value for each chunk. Thus each entry gets a unique hash value. If the hash value for two entries matches exactly, then only one entry is stored in the system. This is an easy technique to remove duplicate data while dealing with high dimensional data.

6. DEPLOYING DATA ON WEB

After we get processed data which is free from duplicates and fits all the requirements, the next step is to deployed mined data into a dynamic platform such as web which can be easily accessed by user.

Using Application Programming Interfaces (APIs) to deploy data on a dynamic platform- These programming structures are then used to define data mined information as an XML object that can easily be transferred between different environments. The user gets a feel as if he is accessing a website which stores all the data and is able to provide him all the information he desires. This capability enables analysts to create data mining information and deploy them in their native forms in the data warehouse where they can be used by applications. An application-programming interface (API) can be described as a set of instructions that can be used for accessing a software application that is deployed on a dynamic platform.



Team Info

Team Contact Info
Contacts details of ABU QA Pune team

Birthday
Team Birthday Details

New User
Please Fill the form for the first time

Machine Details

Virtual Machine Details
Displays Virtual Machine Details

Team Machine Configuration
Team System Boot Inventory

Search Machine First Name ▾ Search

SLM Metrics

Select Project Select One ▾ Go

Team Productivity

SPR Details

SPR Number #

SPR Reported
Total Number of SPRs Reported by Pune Team

SPR Verified
Total Number of SPRs Verified by Pune Team

TC's Created
Total Number of Test Cases Created Last Week

Task Executed
Total Number of Task Executed by Pune Team

QA DOCs

QA Manual
Documents related to QA manual

QA Procedures
Documents related to QA procedure

QA Strategies/Approach Documents
Documents related to QA Strategies and Apporach

Fig. 6- Output generated by Data Dashboard.

Fig. 6 shows the mined data displayed on a dynamic platform. User can easily access the data. The data has been mined for an IT company. It shows Details of the employees, the teams and the projects.

7. EXPERIMENTAL RESULTS

Experimental Results concentrate on how the data is being displayed and in what forms what can have data visible to us. It also shows internal data of each of the modules that are displayed in figure 6.

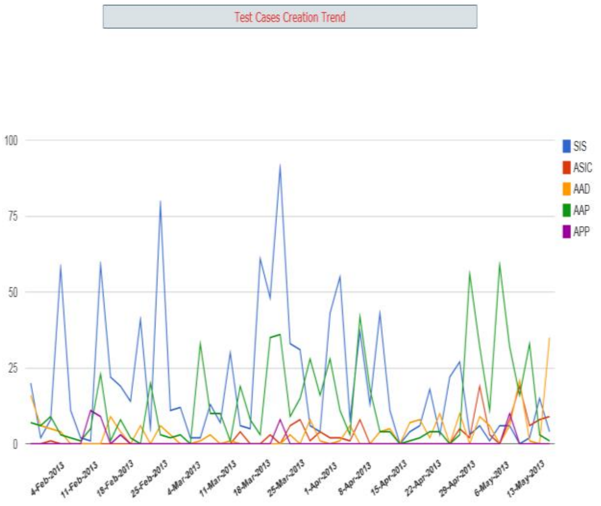
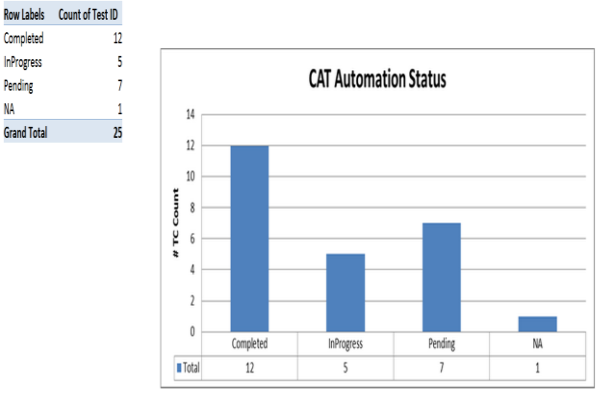


Fig. 7- Data mined information being displayed in graphical format on a dynamic platform.

Figure 7 shows data mined information displayed in graphical format. Displaying information in this form helps user to easily interpret the data and information regarding different projects which are in process. It shows the number of test cases created for a particular project within the given time frame.



Row Labels	Count of Test ID
Completed	12
InProgress	5
Pending	7
NA	1
Grand Total	25

Fig. 8- Data mined information being displayed.

Figure 8 shows the data mined information displayed for a particular project in the company. From the huge database the information regarding a project is mined and displayed as above to provide ease of access to the user.

The status for this project is displayed in terms of number of test cases which are completed or drafted, number of test cases which are pending, number of test cases which are in progress and number of test cases which cannot be completed due to some functional issues.

8. CONCLUSION

Several systems require data to be consistent since they have to offer high-quality services. These systems may be affected largely by the presence of duplicate data in their repositories.

A Data Mining tool having the capability of dealing with high dimensional data and simultaneously using efficient data deduplication techniques for providing high reliability, low disk space and high throughput is being developed. Its use is to derive useful knowledge from databases that are far too large to be analysed by hand. The tool deploys data mined information into a dynamic platform such as web which provides ease to user to access huge database.

Data mining is performed by accessing the data from two large databases which contains data in millions. UI testing is done with the help of the Qualitia testing tool. Functional testing is done with the help of the Selenium testing tool.

9. REFERENCES

[1] Ming-Syan Chen, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, VOL 8, NO.6.

[2] Michael Steinbach, Levent Ertöz, and Vipin Kumar "The Challenges of Clustering High Dimensional Data".

[3] Mamatha Nadikota, Satya P Kumar Somayajula,Dr. C. P. V. N. J. Mohan Rao ,CSE Department,Avanthi College of Engg &Tech ,Tamaram,Visakhapatnam,A,P..,India "Hashing and Pipelining Techniques for Association Rule Mining" International Journal of Computer Science and Information Technologies, Vol. 2 (4), 2011, 1448-1452.

[4] Ying-Hsiang Wen, Jen-Wei Huang, and Ming-Syan Chen,"Hardware-Enhanced Association Rule Mining with Hashing and Pipelining" IEEE Transactions on knowledge and data engineering, VOL. 20, NO. 6, JUNE 2008.

[5] Alissar NASSER, Denis HAMAD, Chaiban NASR, KAIST Ky Ho Park CORE Lab. KAIST "K-means Clustering Algorithm in Projected Spaces".

[6] R. Agrawal and al, "Automatic Clustering of High Dimensional Data used for Data Mining Applications," Proc. ACM SIGMOD conf., pp. 94- 105, 1998.

[7] N. Mandagere, P.and S. Uttamchandani. "Demystifying Data Deduplication" presented In Companion '08: Proceedings of ACM/IFIP/USENIX Middleware '08, pages 12-17,2008