Estd. 1964
Celebrating
50
& beyond
BHARATI VIDYAPEETH

BHARATI VIDYAPEETH
BVIMIT

BHARATI VIDYAPEETH'S

Technical Magazine

2017-2018

Table Of Content

International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

A Review of Cleanliness Mission "Swachh Bharat Abhiyan"- A Survey done for Thane District

Supreeta Desai, Prof. Sudeshna Roy
MCA Department, Bharati Vidyapeeth Institute of Information Technology and Management, CBD Belapur, Navi Mumbai, Maharashtra, India

ABSTRACT
Living within the lap of nature is an ecstasy. The need of the hour is to conserve and safeguard the aura and ecological life balance of nature. This menace may be fought against solely by creating awareness in the society.

"Cleaning and organizing is a practice and not a project"

Swachh Bharat Abhiyan is a mission towards protecting our environment from getting filthy. It was a major step taken by the Government of India and the objective is to eliminate open defection, conversion of insanitary toilets to pour flush toilets and eradication of manual scavenging. Cooperation of people is mandatory to make Swachh Bharat Abhiyan mission successful.

The dusty and stained walls have been transformed into beautiful art piece like that of thane station area. The dump yard which was filled with garbage and of course the unbearable stink have been cleaned. The garbage which was overflowing has been picked up. On the other hand there are residents with poor hygiene which needs to be improved. People talk about litter around them but are not taking any measures to dispose it. It should not just be the responsibility of the 'safaai kaamgar' who clean the particular area.

Keywords: *Swachh Bharat Abhiyan, Swachh Bharat Mission, Swachhata Abhiyan,Clean India, Swachh Bharat Abhiyan Clean India*

I. INTRODUCTION

Mahatma Gandhi was a Prodigious person in the life of every individual in India, whatever Gandhji had uncovered has timeless dimension to it. The ideas are concerning even more so today. His vision was clear cut: a clean mind, a clean body and clean surroundings.

Back in 1999 the government had structured the gramin sanitation program and launched "Nirmal Bharat Abhiyan" which was so not recognized by people of India and was unsuccessful to achieve its target. However failure had its drawbacks as the campaign had minimal people participation and the awareness level was very low.

On 2nd October 2014 our prime minister Narendra Modi launched "Swacch Bharat Abhiyan" campaign in India which aims to eliminate open defecation through construction of household owned and community owned toilets. Nirmal Bharat Abhiyan was wholly and solely had its focus on open defecation whereas now Swacch Bharat Abhiyan have addressed the drawbacks.

India has evolved on many fronts like Business, science and technology, Cloud architect, Health science and many more over the decades since independence in 1947. Our per capita income has been rising at current prices during 2017-18 is estimated to have attained a level of Rs 1,12,835 as compared to the estimates for the year 2016-17 of Rs 1,03,870, showing a rise of 8.6 percent . Today the literacy rate in India has been improved a lot; the most literate state is Kerala with 93.92%. However, on the

Contrary, India has the largest numbers of malnourished people in the world. Studies show that malnourishment is not only lack of proper nutrition but also access to hygiene, safe drinking water and food. Many water related diseases like cholera, diarrhea, malaria, typhoid and filariasis erupt every year in India due to poor quality drinking water and sanitation. Access to safe water and sanitation are crucial for a healthier lifestyle.

India's 1.32 billion people live in large number of rural as well as urban habitations. The cities such as Mumbai, Bangalore, Delhi, Ahmadabad and Hyderabad are considers as the most populated cities in India. The population density is around 412 people per square kilometer, which ranks 31st in the world. Around 70% of India's population live in rural area so about one-third population i.e. 30% population live in urban area. So looking at the statistics we need to determine the sanitation program differently in both the rural as well as urban area.

II. LITRATURE SURVEY:

1. List of facilities provided by swachhata abhiyan.

As of today Swachh Bharat Abhiyan has cover more than 7 .5 crore households, 3.8 lakh open defecation free villages, more than 4,465 open defecation free villages in namami gange, 395 open defecation free districts and 17 open defecation free states/UTs as of June 2018.

A. Swachh bharat mission (Urban area)

The mission had its precedence i.e. to bring behavioral changes in people regarding healthy and hygienic lifestyle, eliminating solid waste and scavenging, conversion of unsanitary toilets to pour flush toilets and completely eradicating open defecation. Public toilets have been build in various location around like the bus stops, railway stations, tourist places, markets and also in slum areas the government is progressively planning to build more toilets. Rs 4000/- is given to every household for construction of a toilet by the government of India where 2000/- will is given as first installment after verification and 2000/- will be given after under construction toilet photographs are sent as a proof to ministry, additionally, 1300 per household will be given as an incentive. Swachh Bharat for Urban area has constructed around 34 lakh toilets.

According to the guidelines of Swachh Bharat Mission (Urban) tentative basic cost for community toilets is Rs. 65,000/- per seat which has been revised with additionally Rs. 39,200/- per seat and for public toilets is Rs. 75,000/- per seat which has been revised with additionally Rs. 12,800/- per seat.

B. Swachh bharat mission (Gramin or Rural area)

Nirmal bharat abhiyan which was initially called as total sanitation campaign was planned to make india an open defecation free in gramin area through proper management of solid and liquid waste.Swachh bharat gramin has build over 5.3 crore toilets. In rural India, 3.8 lakh villages, have been declared open defecation-free.

Swachh bharat gramin has been allocated Rs.13,948 crore in 2017-2018. Total Rs.10,000-Rs.12,000 per unit fund allocation has been given for rural toilets and Rs.35,000 has be given for school toilets. For anganwadi toilets,the funds provided is Rs. 8,000 and community toilets is Rs.2 lakh. The mission is been carried out with involvement of every gram panchayat i.e. village council, panchayat samiti and Zila parishad. Also school Children and teachers have put efforts in making this mission successful. The main Mission is to contribute in construction of individual household latrine in rural development.

2. Promotion or advertisement.

Bollywood celebrities play powerful impact on today's youth. The government of India hence collaborated with several celebrities for this purpose with an intent to carry out an Open-Defecation Free (ODF) India by 2 October 2019. Various personalities who have promoted the cause include Amitabh Bachchan, Anushka Sharma, Shilpa Shetty, Virat Kohli, Priyanka Chopra, Sachin Tendulkar and Salman Khan. While some of these personalities like vidya balan, Amitabh Bachchan featured in the ad campaigns; others were seen encouraging the motive by picking a broom. Around Rs 530 crore was spent for the marketing the Swachh Bharat Abhiyaan in three years.

3. NGO's which promotes swachhata abhiyan.

The NGO's are the non-governmental organizations who work independently of any government whose whole idea is to handle issues that are either social or political. Various NGO's like SWaCH pune(Solid Waste Collection and Handling), The Ugly Indian and Waste warriors.

III. OBJECTIVE:

1. To understand and identify the programs or facilities by government through swacchata abhiyan.
2. To know the awareness of facilities.
3. To identify the use of facilities.

IV. HYPOTHESIS:

H0: There is no correlation between awareness and utilization of facilities.

H1: There is positive correlation between awareness and utilization of facilities.

V. Research Questionnaire:

Thane made a forward leap from 116 to 40 according to a recent article by times of India.

Sr No.	Questions	Options
1.	Are you aware of the Nirmal Bharat Abhiyan?	1. yes 2. no
2.	Are you aware of the Swachh Bharat Abhiyan?	1. yes 2. no
3.	Are you interested in contributing to the Swachh Bharat Abhiyan	1. Strongly Disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly Agree
4.	Is 24 hour water available in /for the toilet?	1.yes 2.no 3.sometimes
5.	Do you prefer using Public Toilet?	1.yes 2.no
6.	Is there any Open Defecation spot /excreta in an open place?	1.yes 2.no
7.	Do you still find plastic being used at around you in market instead of a eco friendly bag	1.yes 2.no
8.	Are you using plastic bags instead of eco friendly bags ?	1.yes 2.no
9.	Do you know about the Swachh Bharat Abhiyan app?	1. yes 2. no 3. no idea
10.	Have you used the Swachh Bharat Abhiyan app?	1. yes 2. no 3. no idea
11.	Do you think that hoarding and advertisment are enough to spread the awarness about Swachh Bharat Abhiyan?	1. yes 2. no
12.	Do you know about wet or dry garbage	1. yes 2. no 3. no idea
13.	At home do you maintain separate Wet and Dry Garbage?	1. yes 2. no 3. no idea
14.	Do you find less littering on road?	1. yes 2. no 3. no idea
15.	In your perception would the swatch Bharat Abhiyan change the mindset of the people?	1. yes 2. no 3. no idea
16.	Rate the cleanliness of the public toilets in your area	1 to 5

VI. RESEARCH METHODOLOGY:

➢ Primary Data:

Survey forms are distributed to the local residents of thane for data collection. Survey forms contain Questionnaire. The reasons were:

1. To determine whether or not the queries chosen are going to be relevant in addressing the set objectives of the analysis and conjointly to update form and discard pointless queries.
2. To rectify errors before the specific questionnaires go out to the particular respondents.
3. To calculate the intermediate response rate within the use of numerous medium for information collection.

We have prepared semi-open ended questions. These questions allocated have no restrictions on how research participants could respond to the questions. Participants reply to the given questionnaire according to their own opinion with closed ended questions.

➢ Secondary Data:

➢ Secondary research methodology defined as Data which were gathered from miscellaneous sources, inclusive of reference materials like dictionaries, archival sources, textbooks, journals/articles, review and online sites.

In secondary research methodology, we have refer various application on the playstore listed below:

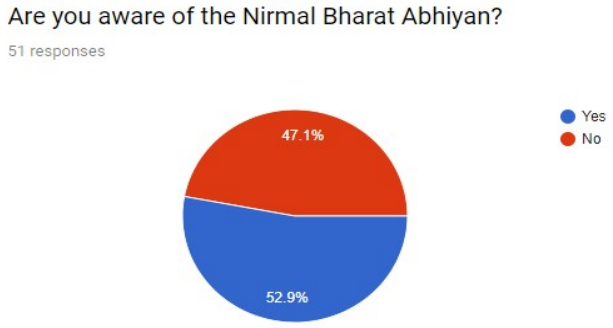
1. Swachhata-MoHUA
2. Swachh Bharat Abhiyaan
3. Swachh Bharat Clean India App
4. My Clean India

Also, Radio stations and TV channels promote 'swachhta hi seva' campaign.

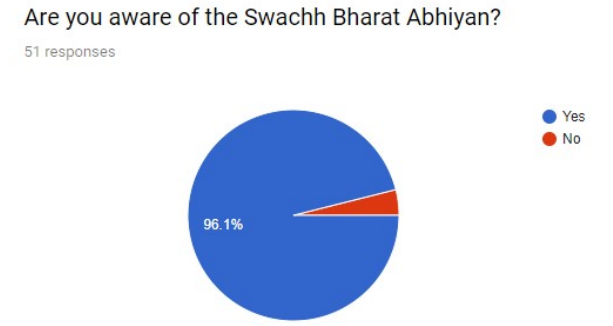
➢ **Sample Size**

The population under-study which consists of local residents in thane district of 18.9 lakhs as of july 2018, it is remarkably impossible to interview such chaotic amount of population. So a part of population is referred as sample for the survey. 51 members from Lodha Paradise in Thane district were surveyed which included 23 female and 28 male.

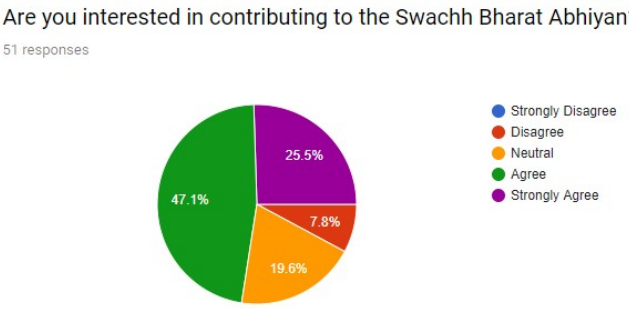
VII. Data Analysis and Interpretation



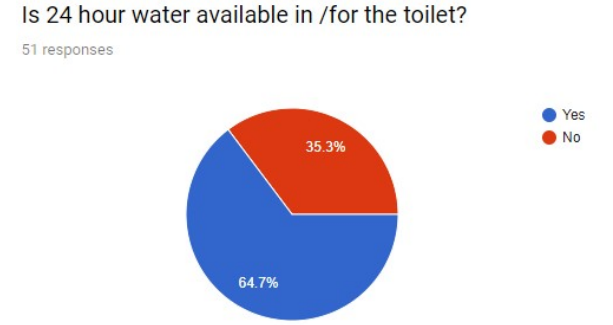
Interpretation: The above Result shows that 52.9% people are aware about nirmal bharat abhiyan and the rest were unaware about it.



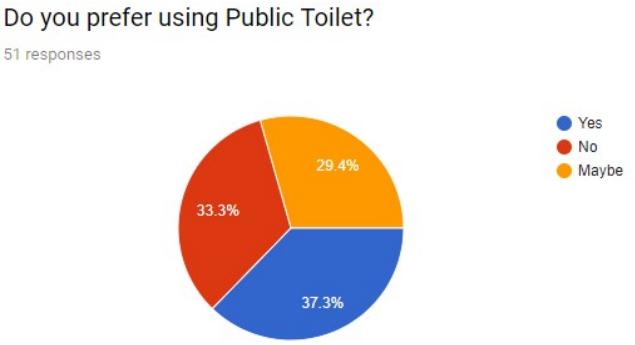
Interpretation: The above Result shows 96.1% people are aware about swachh bharat abhiyan.



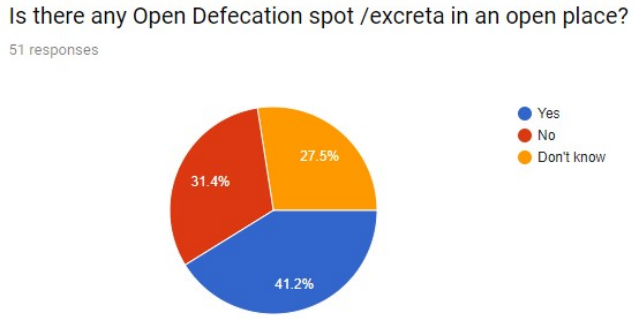
Interpretation: According to our survey 25.5% strongly Agree, 47.1% Agree, 19.6% Neutal, 7.8% disagree to show contribution to Swachh Bharat Abhiyan.



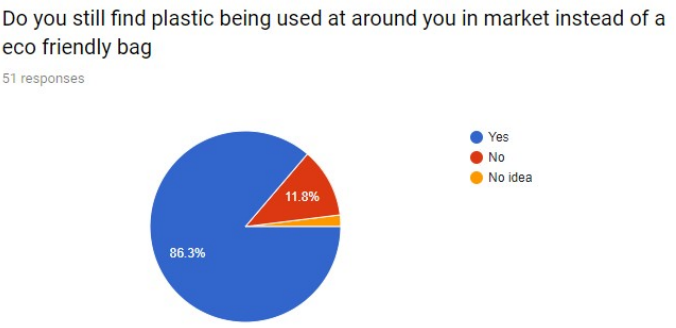
Interpretation: The water facility is 24 hours in the residence according to 64.7% of the people who have been surveyed.



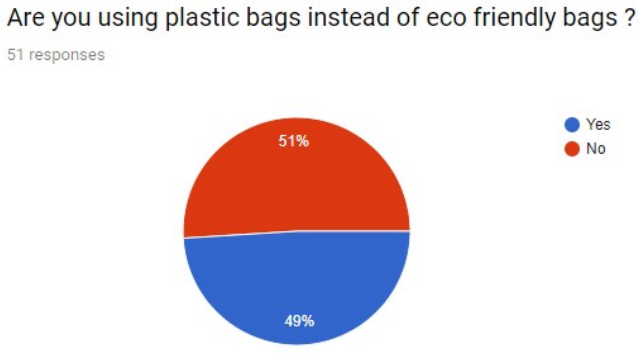
Interpretation: 37.3% prefer using public toilets whereas 33.3% disagree with it and some are on the neutral side.



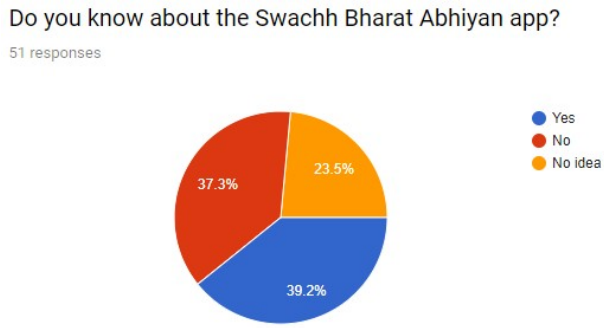
Interpretation: 41.2% have given a positive outlook whereas 31.4% disagree.



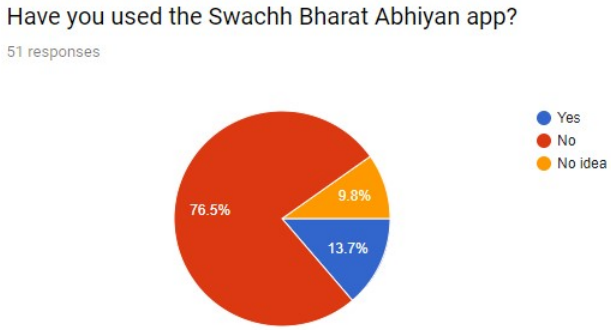
Interpretation: Recently the government has banned plastics but still 86.3% have noticed of plastic still been used.



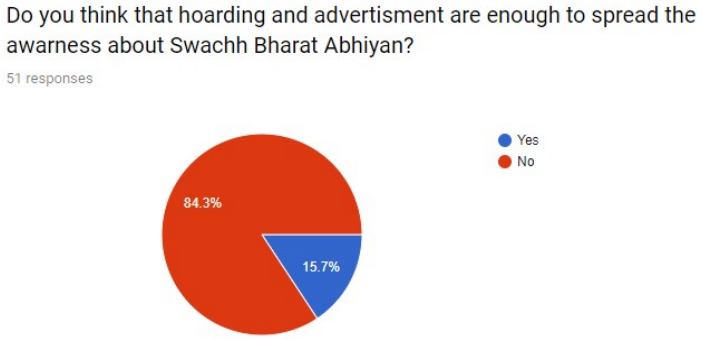
Interpretation:49% people use plastic bags where as 51% show a positive response and are using eco friendly bags instead of plastic.



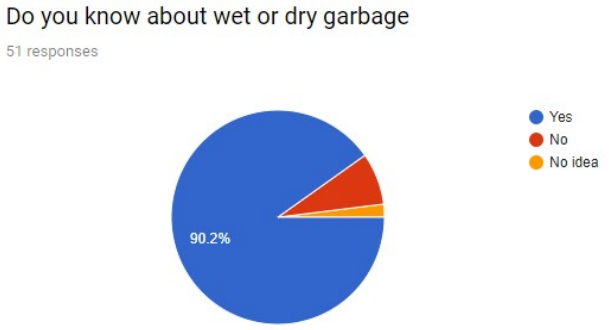
Interpretation: swachh bharat app can also be used to directly connect with the ‘safai karamcharwala’ to keep their surroundings clean yet 37.7% and 23.5% are not using and have no knowledge of it.



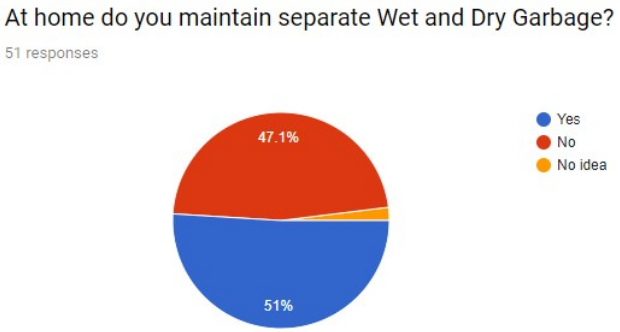
Interpretation: 76.5% do not use the SBA app but 13.7% have used the SBA.



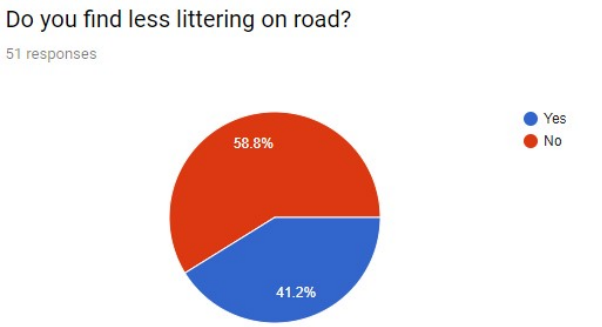
Interpretation: 84.3% still think that hoarding and advertisement are just not enough to create awareness among the people.



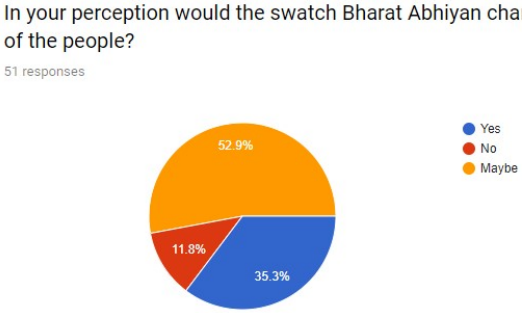
Interpretation: 90.2% have knowledge about wet or dry garbage. Managing separate dry or wet waste is mandatory for recycling purpose.



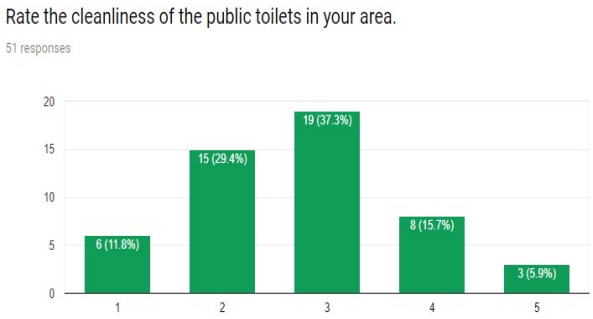
Interpretation: Even after 90% people having knowledge about wet and dry waste people still do not maintain separate dustbins.47.1% still do not use separate dustbins whereas 51% use separate dustbins.



Interpretation: 41.2% find less littering on road whereas 58.8% find more littering in their surroundings.



Interpretation: only 35.3% agree that SBA has change the mindset of the people. Whereas 52.9% are hoping that SBA might change the mindset of the people. 11.8% still disagree.



Interpretation: The above histogram shows the cleanliness of public toilets which has 19 people have rated it 3 which is average and 15 people have rated as below average. Only 3 people have rated as excellent.

VIII. Result and discussion

This chapter analyses the responses given by the locals of thane district through the administration of structured questionnaire conducted. In order to make interpretation and analysis easier, tables are presented using Chi-Square test for the above analyzed data.

Table I

Usage of SBA	Count	Percentage
Strongly Disagree	1	1%
Agree	24	24%
Disagree	3	3%
Neutral	10	10%
Strongly Agree	13	13%
Total	51	100%

Table I represents usage of SBA interview on various areas from thane district.

Table II

Awareness of SBA	Count	Percentage
Yes	48	48%
No	2	2%
Total	51	100%

Table II represents the awareness of SBA. Out of these, 48% represent people who are aware of SBA.

Table III

Rate of Cleanliness	Count	Percentage
1	6	6%
2	15	15%
3	19	19%
4	8	8%
5	3	3%
Total	51	100%

Table III represents the rate of cleanliness with respect to their public area who are interested to Contribute for SBA.

IX. Conclusion:

1. It is necessary to educate young children to make cleanliness as habit and not a responsibility.
2. People should use the SBA app to complain or inform about any garbage or litter in surrounding areas.
3. Heavy fine should be demanded from the people who litter or spit in public areas.

If these points are followed by people then there might be 'Acche Din' very soon.

X. Reference:

1. https://www.researchgate.net/profile/Aparna_Nayak/publication/280939151_Clean_India/links/560433ab08aeb5718feee475.pdf
2. http://questforequity.org/contents/Papers/No%20Swachh%20Bharat%20without%20Annihilation%20of%20Caste.pdf
3. https://en.wikipedia.org/wiki/Swachh_Bharat_Abhiyan
4. http://www.swachhbharaturban.in/sbm/home/
5. https://www.indexmundi.com/india/demographics_profile.html
6. https://www.researchgate.net/publication/279201808_Swachh_Bharat_Mission_A_Step_towards_Environmental_Protection
7. http://www.indiacelebrating.com/essay/swachh-bharat-abhiyan-essay/
8. https://www.readycleaningserviceskc.com/single-post/2016/05/28/Cleaning-and-Organizing-is-a-Practice-Not-a-Project
9. https://timesofindia.indiatimes.com/city/thane/mindset-change-is-the-key-to-clean-campaign/articleshow/61520562.cms
10. https://timesofindia.indiatimes.com/city/mumbai/swachh-ranking-navi-mumbai-in-top-10-big-leaps-for-mum-and-thane/articleshow/64716990.cms
11. https://www.livemint.com/Politics/wLpq03qE6gkVL9FlCQMzgL/Govt-asks-radio-stations-TV-channels-to-boost-Swachhta-Hi-S.html
12. https://www.slideshare.net/AnirudhMehta24/empirical-study-on-measuring-attitude-and-perception-of-people-towards-swachh-bharat-abhiyan
13. https://www.slideshare.net/aditieedeshpande/business-research-methods-project
14. https://www.hindustantimes.com/mumbai-news/thane-clean-up-drive-to-end-on-a-high-note/story-hEcwcMQQ0oMkQ8UwysUdRL.html

IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Cluster-Then-Predict and Predictive Algorithms (Logistic Regression)

A. Bhattacharjee[1*], J. Kharade[2]

[1*] Bharti Vidyapeeth's Institute of Management and Information Technology, Mumbai University, Navi Mumbai, India
[2]Bharti Vidyapeeth's Institute of Management and Information Technology, Mumbai University, Navi Mumbai, India

[*]*Corresponding Author:* anikb48@gmail.com , *Tel.: +91-91673-87781*

Available online at: www.ijcseonline.org

Received: 26/Jan//2018, Revised: 06/Feb2018, Accepted: 21/Feb/2018, Published: 28/Feb/2018

***Abstract*—** Stock market is playing a vital role as investments option and investors make short-term investments as well as long-term investments. But here the main question arises "Where to invest?" and "when to invest?" even if an investor is aware about where to invest, it is still unpredictable whether or not stocks will have good future returns over time. To eliminate this dilemma predictive algorithms were introduced that will help investors in making investments by predicting which stocks will have positive expected returns. However, predicting stock returns with predictive algorithms alone is not enough. Clustering algorithms are widely used to cluster the stocks that have related returns over time. Using Cluster-Then-Predict approach we are going to prove that it provides more accurate results than the original predictive (Logistic Regression) model.

***Keywords*—** Predictive Algorithms, Regression, Classification, Clustering, Logistic Regression, Stock Returns, Cluster-Then-Predict

I. INTRODUCTION

Machine learning is described as the data which is obtained by knowledge extraction. Machines do not require to be programmed directly instead it's trained to make decisions driven by the data. Rather than writing a code for every specific problem, data is provided to the algorithms and logic is developed on the basis of that data. When the machine improves itself depending on its previous experiences it can be concluded that machine has truly learned on its own [1].

Different databases such as RDBMS, multimedia databases, ORDBMS etc use Data mining. Data mining is used on wide applications like stock prediction, market analysis, stock forecasting etc. Item sets that are frequently used in data mining have an important role to find out the correlations between the database fields.

The objectives of the research paper are:

1) To study Predictive Algorithms
2) To study Cluster-Then-Predict
3) To compare Predictive Algorithms with Cluster-Then-Predict

The purpose of carrying out this study is to spread the awareness about the methodologies that can help in solving real-world problems and can derive better results than the traditional approaches. In this study, Section I contains the introduction of the study which has been discussed already, Section II contain the related work of the study that describes the previous research works, Section III discusses thea two different methodologies in detail. Section IV contains the comparison of results between the two approaches. Section V contains the conclusion and the future scope.

II. RELATED WORK

In paper [2] presents analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. SEER public datasets has been used in this project. The preprocessed dataset consists of 151, 886 records which have 16 fields from the SEER. Three data mining techniques like Naïve Bayes, back propagated neural network and C4.5 decision tree algorithm are investigated and compared with the achieved prediction performance. it is It was concluded that C4.5 algorithm has a much better performance than other two techniques.
In paper [3] they try to help investors with the better period for the selling and buying stocks within the stock market on the knowledge of past historical experiences. It is analysis that past investigations help to predict future in data analysis. In this paper they used decision tree algorithm which is one of the best data mining techniques. They also explained that forecasting stock return is one the important topic to be learnt for prediction of data.
In paper [4] they explained the main content of the study is to predicting the changes of balance of the users of balance of Yuebao. The methods of prediction adopts first clustering, then predicting. First of all, according to the user's balance of account information, the user's basic information and operating characteristics of the user, this paper made the

classification for users. And then the amount of the user's balance of each class are predicted, so that authors can greatly guarantee the loss of information in the forecast process, this can greatly to ensure the accuracy of data prediction, and the empirical data analysis of the results is also proved that the forecasting model can well describe and forecast the change of the balance data, which can get more excellent results than the direct forecasting.

Algorithm	Usage	Pros	Cons
Linear Regression	Predicts a continuous outcome (price, salary, etc)	1) Simple, well recognized 2) Can work on small and large datasets	Consider a linear relationship
Logistic Regression	Predicting a categorical outcome (Yes/No, True/False, etc)	1) Calculates probabilities that is used to asses confidence of the prediction	consider a linear relationship
CART	It can Predict a categorical outcome or continuous outcome	1) Handles datasets has no linear relationship 2) Easy to explain and interpret	1) Does not work well on small datasets
Random Forests	Predicting a categorical outcome or continuous outcome	1) Has better accuracy over CART	1) Parameter tunings are needed 2) Cannot easily interpret unlike CART

Table 1. Comparison of predictive algorithms

III. METHODOLOGY

3.1) Predictive Algorithms (Logistic Regression)

Logistic Regression is commonly applied to a group of independent variables to predict an outcome which is binary. This binary outcome can be represented in either in 0 or 1/ True or false. Dummy variables often represent this binary outcome. Mostly, it is used for solving Classification problems. Logistic regression is considered as an exceptional case of linear regression in which the target variable or the dependent variable is categorical

For the dependent variable it uses the logit function, it fits the data that consists the log of odds using the possibility of the occurrence of the event.

In this example the value of 'y' can be 0 to 1 it is represented by the equation [5] odds = probability of event occurrence divided by the probability of the event.

$$Odds = \frac{P}{1-P}. \quad (1)$$

$$\ln(Odds) = \ln\left(\frac{P}{1-P}\right). \quad (2)$$

$$\log(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + (\beta_1 \times 1) + \cdots + (\beta_k \times k). \quad (3)$$

Where, P is the probability of interested characteristic presence. As there is a binomial distribution of dependent variable implemented, there has to be a link function which will be best fitted for the distribution, which is the logit function. The equation in the above selects the parameters to maximize the odds of receiving the sample values instead of reducing the sum of squared errors (as seen in the ordinary regression). [6]

Points to be considered before implementing logistic regression:

- It handles non-linear relationships efficiently between the target variable and predictors. It can take various relationships types because it uses a log transformation which is non-linear for predicting the ratio of odds.
- To remove overfitting as well as underfitting, all significant variables should be included. A better method to assure this practice is by using a step-by-step method to compute the logistic regression.
- It needs samples which are large in size. Since, maximum likelihood that calculations are less accurate at low sample sizes in comparison to the simple least square.
- It does not have multi-collinearity i.e. independent variables need not be inter-related with each other. However, there are still options to consider interaction impacts of categorical variables during modeling and computation.
- It will be known as Ordinal logistic regression when the values of dependent variable are ordinal. [7]

3.2) Cluster-Then-Predict

Current section explores the method of clustering. Clustering is an classification method which is part of unsupervised learning that focuses on creating groups of clusters, in a way that entities which belong to the same cluster are very identical and entities in other clusters are quite distinct [8]. Analysis of clusters is one of the oldest topics in the data mining field. It is the initial step in the direction of exciting knowledge discovery. Clustering is a procedure of grouping data entities into a group of different classes, called clusters. Now entities within a class have high similarity to one

another whereas entities in separate classes are more different.

Analysis of clusters has been used majorly in wide areas, that includes data analysis, bioinformatics, pattern recognition machine learning and text mining. In Industries, clustering can help entrepreneurs discover interests of their customers based on purchase patterns and characterize groups of the customers. In geology, the expert can apply clustering to identify regions of houses that are in a city and lands. Clustering can also used in classification of documents that are on Web for the discovery of information.

In K-Means clustering similar kind of data are clustered for data prediction. In K-Means clustering, each data point is assigns itself to its nearest cluster and using Euclidian distance formula the data points are clustered. Using this, it will it improves the clusters and improve the Euclidian distance formula. This improvement is based on normalization. Two new features are added because of this improvement. Based on normalization, the first feature will calculate normal distance metrics. Because of the majority voting the second feature will cluster the data points. The proposed technique will be implemented in R.

Steps for implementing K-Means clustering:

1. Define number of clusters k
2. Put each point to a cluster randomly
3. Calculate the cluster centroids
4. Again assigning each point to the nearest cluster centroid
5. Re-Calculate the cluster centroids
6. Repeat steps 4 and 5 until no improvement is made

IV. COMPARISON OF PREDICTIVE ALGORITHMS WITH CLUSTER-THEN-PREDICT

The StockCluster.csv dataset contains monthwise stock returns that has been provided by the NASDAQ stock exchange. This dataset has been taken from infochimps which provides access to many datasets. The data of stock price in this dataset has 12 variables and 11580 observations as shown in Figure 1. The dataset is loaded in R.

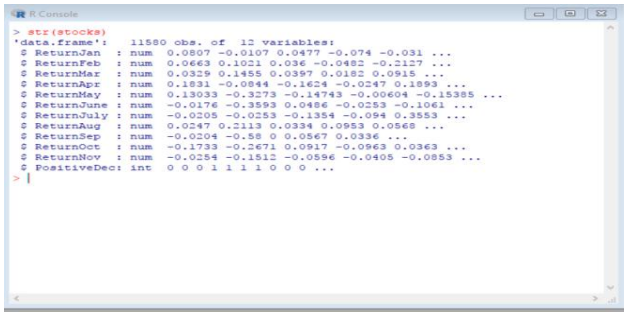


Figure 1: List of variables in the Dataset

```
> # Logistic regression model
>
> library(caTools)
>
> set.seed(144)
>
> spl = sample.split(stocks$PositiveDec, SplitRatio = 0.7)
>
> stocksTrain = subset(stocks, spl == TRUE)
>
> stocksTest = subset(stocks, spl == FALSE)
>
> StocksModel = glm(PositiveDec ~ ., data=stocksTrain, family=binomial)
>
> PredictTrain = predict(StocksModel, type="response")
> PredictTest = predict(StocksModel, newdata=stocksTest, type="response")
> # to test the accuracy of the model on test set
>
> table(stocksTest$PositiveDec, PredictTest > 0.5)

    FALSE TRUE
  0   417 1160
  1   344 1553
> (417+1553)/(417+1160+344+1553)
[1] 0.5670697
```

Figure 2: Test set accuracy of Logistic Regression model

In Figure 2, using this dataset the Logistic Regression model is created. This model gives the accuracy of 57% on the test set over the threshold of 0.5 which obviously beats the baseline model.

```
> # to find which cluster has largest observations
>
> set.seed(144)
>
> km = kmeans(normTrain, centers = 3)
>
> table(km$cluster)

   1    2    3
3157 4696  253
```

Figure 3: Assigning each observation to the cluster

In Figure 3, it can be observed that the K-Means clustering algorithm created 3 clusters and assigned each observation to the corresponding cluster based on the seed value that has been specified in the R console. The table command in R console showed the number of observations that each cluster has which helps the investigator, to find out the cluster that has the most number of observations.

```
> AllPredictions = c(PredictTest1, PredictTest2, PredictTest3)
>
> AllOutcomes = c(stocksTest1$PositiveDec, stocksTest2$PositiveDec, stocksTest3$PositiveDec)
>
> table(AllOutcomes, AllPredictions > 0.5)

AllOutcomes FALSE TRUE
          0   467 1110
          1   353 1544
> (467+1544)/(467+1110+353+1544)
[1] 0.5788716
```

Figure 4: Test set accuracy using Cluster-Then-Predict

As shown in Figure 4, there is a modest improvement on the accuracy of the test set over a threshold of 0.5. Using Cluster-Then-Predict approach we got the accuracy of 58% which is anytime better than the baseline model but also, it slightly gives better results than the traditional Logistic Regression model.

V. CONCLUSION AND FUTURE SCOPE

In this paper we came to know that Cluster-Then-Predict methodology can provide more accurate results than the simple implementation of the predictive algorithms. During the small experimentation conducted on the dataset, we

witnessed that, there was not a lot of improvement over the implementation of the single predictive algorithm (Logistic Regression). The reason behind that is predicting stock returns is a ridiculously hard problem considering that fact we can see a good increase in accuracy. The main objective is to make investors feel more confident that they will have positive returns on the stocks that they have invested in. We can improve our overall accuracy on this dataset by implementing other predictive algorithms that are new in the field of machine learning (e.g XGBOOST, LightGBM, CatBoost etc) which can provide better accuracy than the traditional predictive algorithms. Furthermore, to improve the results even further it can be recommended to implement clustering before applying any of the Predictive Algorithms mentioned above.

REFERENCES

[1] W. Huang, Y. Nakamoria and S. Wang, " *Forecasting stock market movement direction with support vector machine*", Computers & Operations Research, Vol. 32, pp. 2513 – 2522.

[2] A. Bellaachia, E. Guven, "*Predicting Breast Cancer Survivability Using Data Mining Techniques*", In the Proceedings of the 2010 Department of Computer Science The George Washington University, Washington DC, pp. 20052, 2010

[3] S. Gour, "*Developing Decision Model by Mining Historical Prices Data of Infosys for Stock Market Prediction*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.10, pp. 92-97, 2016.

[4] Y. X. Lu, T. Zhao, "*Research on time series data prediction based on clustering algorithm*", In the Proceedings of the 2017 American Institute of Physics Conference, United States, pp. 1864-020152, 2017.

[5] S. S. Sathe, S. M. Purandare, P. D. Pujari and S. D. Sawant, "*Stock Market Prediction Using Artificial Neural Network",* International Education and Research Journal. Vol. 2, Issue 3, pp. 2254-9916, 2016

[6] M. Mittermaye, "*Forecasting Intraday Stock Price Trends with Text Mining Techniques*", In the Proceedings of the 2004 37th Hawaii International Conference on System Sciences, , pp. 0-7695-2056-1/04, 2004.

[7] K. S. Kannan, P. S. Sekar, M. M. Sathik and P. Arumugam "*Financial Stock Market Forecast using Data Mining Techniques*", International Multi Conference of Engineers and Computer Scientists, Vol 1, I,IMECS 2010, March 17-19,2010, Hong Kong. pp. 2078-0966.

[8] Swati Joshi, Farhat Ullah Khan and Narina Thakur, "*Contrasting and Evaluating Different Clustering Algorithms: A Literature Review*", International Journal of Computer Science and Engineering, Vol. 2, Issue.4, pp. 2347-2693, 2014.

Authors Profile

Mr. A. Bhattacharjee pursed Bachelor of Science (Information Technology) from University of Mumbai, Mumbai, India in 2015. He is currently pursuing Masters Of Computer Applications from University of Mumbai, Mumbai, India.

Dr Jyoti . Kharade, Bachelor of Science,Master of Computer Application from Shivaji University, M.Phil from Bharati Vidyappeth deemed University and Ph.D from SNDT University. She is currently working as Associate Professor in Bharati Vidyapeeth's Institute of Management and Information Technology, University of Mumbai, since 2004. She is a member of CSI. She has published more than 27 research papers in reputed international journals including conferences. Her main research work focuses on e-Governance, Data Mining. She has 16 years of teaching experience.

IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 10-16
www.iosrjournals.org

Generating Association Rules for Social Media Analysis

Deepika Jaiswal[1], Shilpa Singh[2], Suhasini VijayKumar[3]
[1] *MCA Student, University of Mumbai*
[2] *MCA Student, University of Mumbai*

Abstract: *Association Rule Mining is a technique for identifying correlation between different data sets. It is also called as Market Basket Analysis, as this was original application area of association rule mining. Also the main aim of association rule is to identify association between items that occur together from a random sampling of all possibilities from data set. Social media mining has the process to represent, analyze, and extract patterns, trends from raw data of social media. These patterns and trends are useful to companies, governments and not-for-profit organizations, as these parties can usethose patterns and trends to design their strategies or to make new programs. This paper focus about generating association rules on these social media raw data which will help organization for analyzing trends very efficiently. It will discuss about various algorithm and techniques used in generating the rules with its procedure also will be compared from other algorithm to identify the best algorithm.*

Keywords: *Apriori Algorithm, Association Rule Mining, Data mining, Social Media Mining*

I. Introduction

A web based service where people can create a public/semipublic profile on some particular domain also which can connect and communicate within that particular network is known as Social media network [1]. A collection of people or a group of individuals which have similar way of contact or interaction like friendship is known as social network. This network can be represented in a graph format which has nodes indicating individual or group connected via a link which represented as line joining them according to their relationship. The graphs can either be depicted as directed or undirected graphs dependingupon the type of relation between those link nodes.

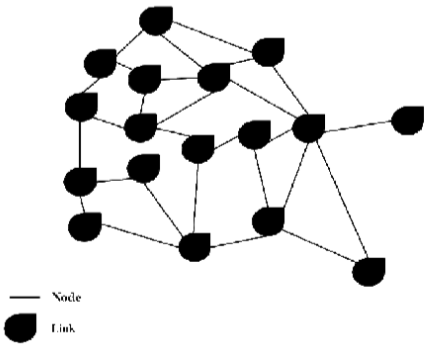


Fig. 1. Social Network showing nodes and links

Extraction of association rule which is said by [10]implies the importance of text mining field which involves identifying various association relationsfrom thewords found insome text data set. These association rules when applied on large amount of transaction database records will help in various decision making process.

There are three dominant issues in social network mining namely: size, noise and dynamism which can be handled by data mining techniques. The hugenature of social network arrangement datasets needs well programmedinformation dealing and analyzes it within a stipulated logical amount of time.Social network proves to be a good platform to mine significant patterns from large data set using data mining techniques [1].
This paper focus about generating association rules on these social media raw data which will help organization for analyzing trends very efficiently. It will discuss about various algorithm and techniques used in generating the rules with its procedure also will be compared from other algorithm to identify the best algorithm.

II. Literature Review

Social media analysis and Online social network are very popular in examine field. The majoreffort in social network research is put forth on social link [13],social connection prediction[14]. Various people across the globe also job on: Personality prediction for micro blog users [15], Using social media to expect real-world outcome[16], Predicting friendship concentration [17,18], Sentiment analysis and opinion mining [19].
Some more interesting areas of research focuses on esteem predicting social media depending on Comment mining[20], Predicting patterns of diffusion processes in social network[21].
Some research is also ongoing in identifying significant users using learning based approach or Page Rank Algorithm or adaptations of the same[21,22].

Social Media sites generate a large amount of data every minute which is shown in Fig 2.1 [11]. As per Technorati 1.2 million new post and approximate 75000 new blogs which gives opinion on a product or service is produced everyday. It's difficult for traditional methods to handle this huge data constantly generated from this site which makes it necessary to develop tools which are capable of analyzing these data. The data generated can be effectively mined by data mining techniques [3].

Fig 2.1: Estimated data generated every minute

The method of mining association rulefocus on discovery huge item sets, which are group of items that are of same view together in a sufficient number of dealings. Usage number of association rule can be identifying if the database is large so for minimizing association rule minimum support and confident are consider, both are specifying by the user which helps us to create valuable rules from the database. The various association rule mining algorithm are

1. **AprioriAlgorithm:** It was developed by Agarwal and Srikant which was intended to operate on database which contain transactions. This algorithm use "bottom-up" technique, where all therecurrent subsetis extended one item at once and groups of suchcandidates are tested against the data. The algorithm terminates when no additional successful extensions are originated. The following Fig 2.2 shows apriori algorithm:

$$
\begin{aligned}
&\textbf{Apriori}(T, \epsilon) \\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
&\quad k \leftarrow 2 \\
&\quad \textbf{while } L_{k-1} \neq \emptyset \\
&\qquad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\} \\
&\qquad \textbf{for transactions } t \in T \\
&\qquad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
&\qquad\quad \textbf{for candidates } c \in C_t \\
&\qquad\qquad count[c] \leftarrow count[c] + 1 \\
&\qquad L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\} \\
&\qquad k \leftarrow k + 1 \\
&\quad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

Fig 2.2: Apriori Algorithm

2. **FP (Frequency Pattern)-Growth Algorithm:** In this algorithm during the first pass, the algorithm counts the number of occurrence of objects in the dataset and saves it to 'header table'. During second pass, it constructs the FP-Tree by inserting those instance. Objects of each instance in the tree are arranged in reverse order of their lowest frequency in dataset, such that the fp-tree gets processed quickly.Objects in the tree are removed if they do not follow minimum threshold coverage. If there are many instance sharing most repeated items, FP-Tree provides high density close to the root of the tree. The only main difference in apriori and fp tree is that it does not generate separate candidate set rather it appends the header table from below by finding all instanceidenticalto the given condition. The following figure Fig 2.3shows the example of FP-Tree

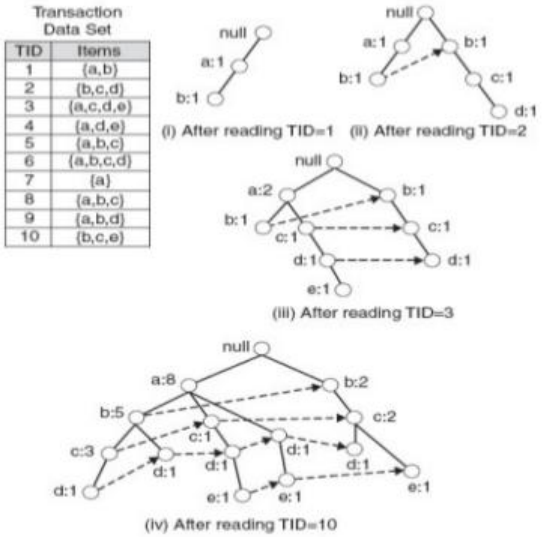


Fig 2.3: Example of FP Growth

The various association rules applications are market base data analysis, customer relationship management (CRM), web usage mining, bioinformatics and intrusion detection. There are two ways of computing usefulness in association rule mining that is subjective and objective. Objectives measures involve statistical analysis of the data such as support and confidence.

Support is given by the ratio of occurrence in the dataset which is denoted by Sup(X) and confidence is given by Confidence(Y=>X) = Support(YUX)/Support(Y).

III. Methodology

To generate Knowledge discovery, we are using raw data collected from social media sites. The proposed methodology is divided into three phases such as Text preprocessing, Association rule mining and Knowledge discovery phase shown in Fig3.1. In text preprocessing phase we first tokenize the input document

in to tokens. Then the tokens are filtered by removing stop word as they do not carry any meaningful information. Normally token contains many suffixes and it is required to remove all the suffixes to achieve better result in knowledge discovery. Then the tokens are indexed using TFIDF (Term Frequency Inverse Document Frequency) values. The Association rule mining phase generates association rules based on weighting scheme TF-IDF that is depending on the users requirement the high frequency keywords are selected to generate association rules. The last phase is to generate Knowledge discovery using those generated association rules.

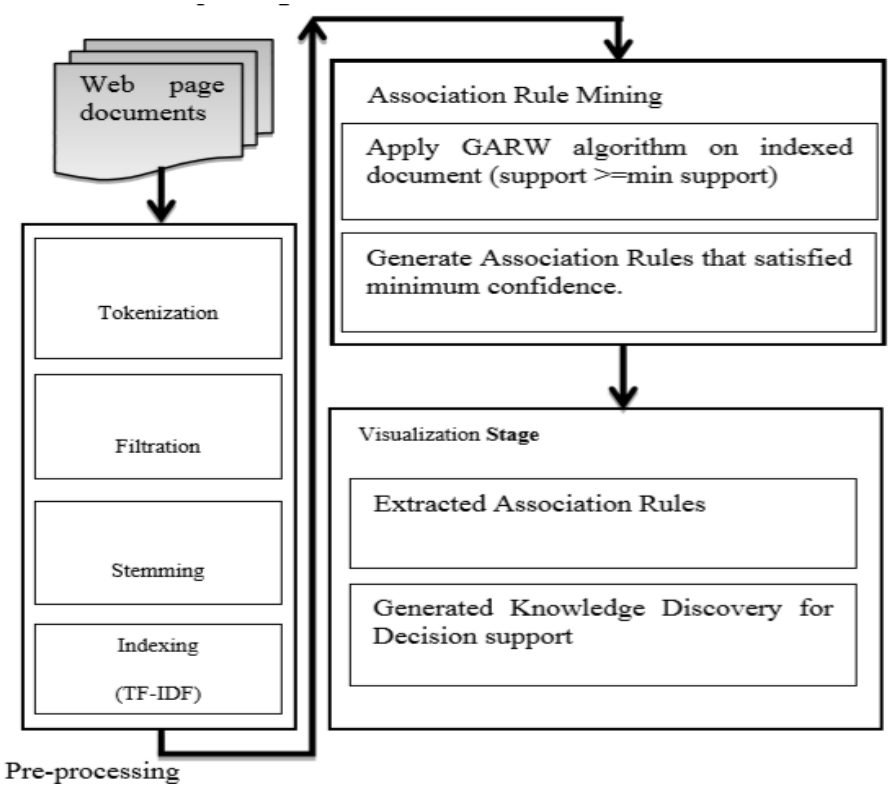


Fig 3.1: Block Diagram of Proposed System

3.1 Text Preprocessing Phase:

Since data on the web is in unstructured format also usage amount of data is generated everyday on social media, thus mining such large document it is necessary to preprocess the input document and store into structured format which can be further use for preprocessing and generating association in it. The text preprocessing phase involves the following phase shown below:

Tokenization:

Normally web page contain information in unstructured format. This create a problem for text mining and it also consume memory and time to process. So tokenization is the process of splitting the text into words. The main aim of this process is to convert unstructured document into structured document.

Filtration of keywords:

To generate accurate knowledge from the collection of raw data the user needs to find out relationship between all the keywords but this task is very tedious if we do not remove redundant and inefficient data. Finding relationship becomes very difficult if that document is not filter well, thus filtration of raw data can be done by removing stop words (does not have meaningful information) and suffixes (attached with the same word but with different form).

Stemming:

Stemming is a process in which variant form of same are reduced to common form. Stemming replace all the match suffix from the keywords with replacement character and words. The reasons for using stemming are it changes the meaning of term even main route word is same, ambiguous association rule are generated, it make data complex and occupy extra memory.

Indexing:

After all the above process the weight scheme Term Frequency, Inverse Document Frequency(TF-IDF)is use to allocate weight to distinguish expressions in the document. Frequency is the count that represent how many

times of keywords has occurred in that document whereas inverse document frequency is the count that represent total number of document that contains the keywords atleast once. We have use this weighting scheme to select higher frequency keyword for generating association rule. With apriori algorithm the only disadvantage is that it consider all the keywords without knowing importance of those keywords for generating association rule.

3.2 Association rule mining stage:

Association rule is of the IF-THEN structure, but it can predict attribute combination, and they are not intended to be used together as a set. For each rule IF antecedent THEN consequent we count its support and confidence matches with user specified values. Support is the possibility that a randomly selected instance will fulfill both the predecessor and successor, and confidence is the conditional possibility that a randomly selected instance will fulfill the consequent given that the instance fulfill the antecedent. In our developed system we have generated only those association rules which satisfy criteria such as support, confidence, and TFIDF value of keywords. The algorithm called as GARW (Generating Association Rule using Weighting Scheme) is found to be better than that of conventional Apriori algorithm. The GARW algorithm works in similar way of Apriori but with some additional steps to resolve problem of Apriori and to generate relevant association rules.

GARW Algorithm

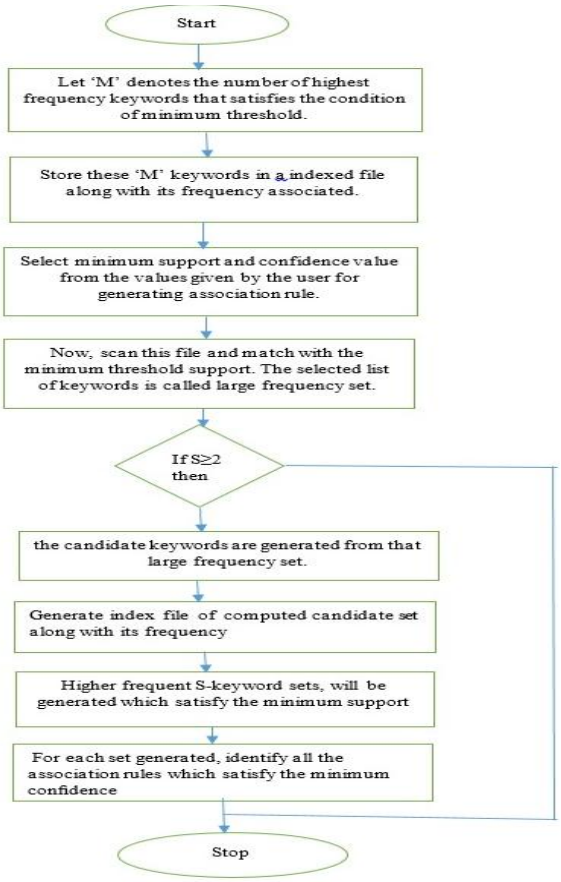


Fig 3.2: GARW Algorithm

3.3 Visualization Stage:

Thus implementing all the above phases on raw data collected from social media generates a collection of association rules between the highest occurring keywords. Now from this generated association rules only the rules which are applicable on the area of interest depending upon the organization's programme is selected and analysis is done .This extracted association rule can be represented as graph, table or in a textual format helping in the decision making process.

IV. Result and Discussion:

We have implemented this system using R language on some dummy text files. The system involved phases such as text preprocessing where text file was filtered by removing all the stopwords and stemming the keywords that is reducing the same word to its common form. Later an indexed file was created containing list

of keywords along with the frequency associated to it. Using this indexed file a word cloud was created by the system depicting the association between those keywords. The following Fig 4.1 shows the word cloud generated by the system showing keywords having higher frequency and Fig 4.2 shows the word dengue is having frequency greater than 12 which is generated by the system as a bar chart.

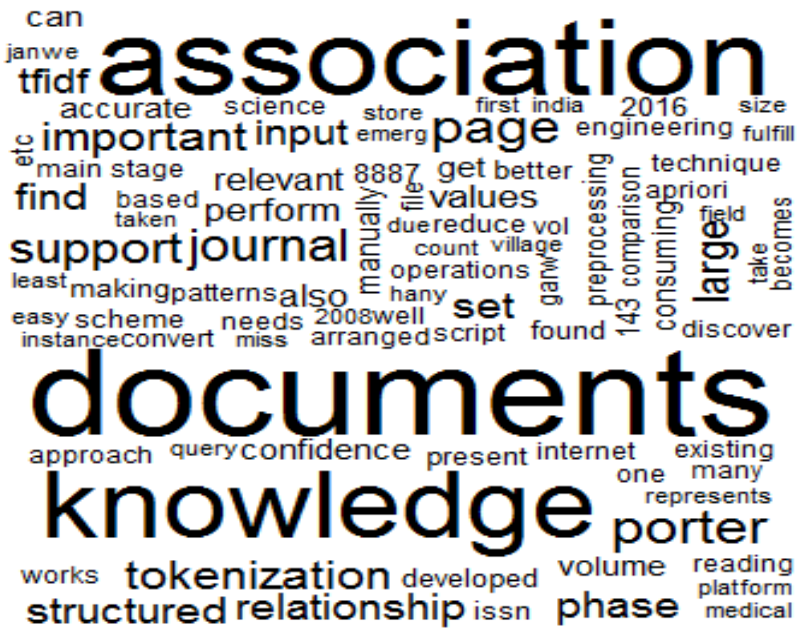


Fig 4.1: Word cloud created using highest frequency words

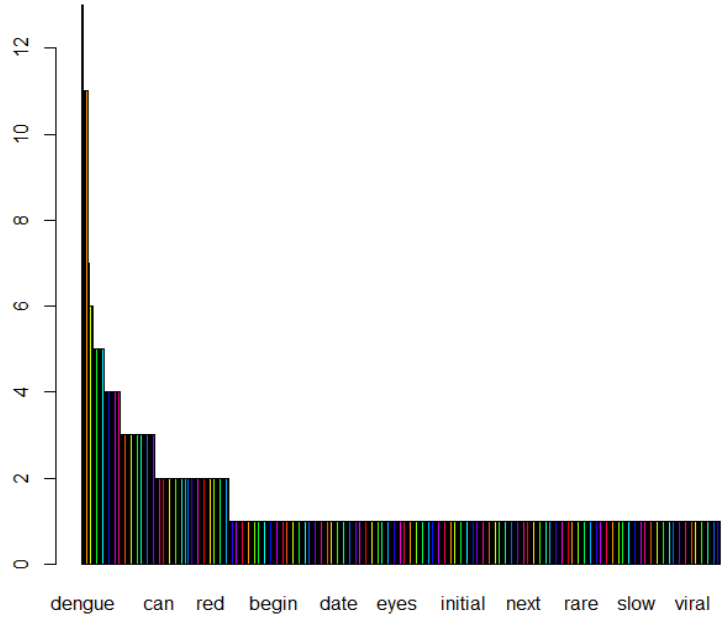


Fig 4.2: Bar plot showing highest frequency words

Hence this system can also implemented by collecting the user opinion or suggestion upon some idea or programme proposed by the organization over social media into a text file and applying the above methodology for the generation of association rule. This will help organization to identify the user support and for making strategic decisions.

V. Conclusion

As per our review there are many algorithms by which we can generate the association rule such as Apriori algorithm, FP-Growth Algorithm and GARW Algorithm. The best and effective algorithm among all three is GARW algorithm because it is based on weighting scheme having faster performance by reducing execution time whereas Apriori generates separate pair for each set, which consumes huge amount of memory and FP-growth algorithm appends the pair to the tree using frequency. Thus in this paper we proposed a system which uses GARW Algorithm having faster execution and better visualization of the extracted association rules.

References

[1]. Mariam Adedoyin-Olowe, Mohamed Medhat Gaber and Frederic Stahl A Survey of Data Mining Techniques for Social Network Analysis

[2]. Mariam Adedoyin-Olowe, Mohamed Medhat Gaber and Frederic Stahl A Survey of Data Mining Techniques for Social Media Analysis

[3]. Anu Sharma, Dr. M.K Sharma & Dr. R.K Dwivedi Literature Review and Challenges of Data Mining Techniques for Social Network Analysis

[4]. Frequent Pattern Mining[Book]

[5]. Said A. Salloum , Mostafa Al-Emran , and Khaled Shaalan Mining Social Media Text: Extracting Knowledge from Facebook . In Proceedings of the International Journal of Computing and Digital Systems ISSN (2210-142X) Int. J. Com. Dig. Sys. 6, No.2 (Mar-2017)

[6]. Chang Zhang , Yanfeng Jin, Wei Jin1, Yu Liu1 Study of Data Mining Algorithm in Social Network Analysis In 3rd International Conference on Mechatronics, Robotics and Automation (ICMRA 2015)

[7]. Anurag Agrahari , Prof D.T.V. Dharmaji Rao Association Rule Mining using RHadoop . In Proceedings of the International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 04 Issue: 10 | Oct -2017

[8]. Dr. R Nedunchezhian, K Geethanandhini Association Rule Mining on Big Data – A Survey . In Proceedings of the International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 5 Issue 05, May-2016

[9]. M. Vedanayaki A Study of Data Mining and Social Network Analysis In Proceedings of Indian Journal of Science and Technology, Vol 7(S7), 185–187, November 2014 ISSN (Print) : 0974-6846 ISSN (Online) : 0974-5645

[10]. Rakesh Agrawal ,Tomasz Imielinski, Arun Swami Mining Association Rules between Sets of Items in Large Databases

[11]. Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acmsigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.

[12]. Tepper, A.: How Much Data Is Created Every Minute?[INFOGRAPHIC]. 2012, http://mashable.com/2012/06/22/datacreated-every-minute/. Retrieved on 16/102013 at 19.00.

[13]. Liben-Nowell, D.; Kleinberg, J. The Link-prediction Problem for Social Networks. J. Am. Soc. Inf. Sci. Technol. 2007, 58, 1019–1031.

[14]. Utz,S.;Jankowski,J. Making"Friends"inaVirtualWorldTheRoleofPreferentialAttachment,Homophily, and Status. Soc. Sci. Comput. Rev. 2015, doi:10.1177/0894439315605476S.

[15]. Asur,S.;Huberman,B.A. PredictingtheFuturewithSocialMedia. In Proceedings of the 2010IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology–Volume 01; IEEE Computer Society: Washington, DC, USA, 2010; pp. 492–499.

[16]. Ahmad,W.;Riaz,A.;Johnson,H.;Lavesson,N. PredictingFriendshipIntensityinOnlineSocialNetworks. In Proceedings of the 21st Tyrrhenian Workshop on Digital Communications: Trustworthy Internet; Springer: Berlin/Heidelberg, Germany, 2010.

[17]. Nia, R.; Erlandsson, F.; Johnson, H.; Wu, S.F. Leveraging social interactions to suggest friends. In Proceedingsofthe2013IEEE33rdInternationalConferenceonDistributedComputingSystemsWorkshops (ICDCSW), Philadelphia, PA, USA, 8–11 July 2013; pp. 386–391.

[18]. Petz, G.; Karpowicz, M.; Fürschuß, H.; Auinger, A. Reprint of: Computational approaches for mining user's opinions on the Web 2.0. Inf. Process. 2015, 51, 510–519.

[19]. Jamali, S.; Rangwala, H. Digging Digg: Comment Mining, Popularity Prediction and Social Network Analysis. In Proceedings of the International Conference on Web Information Systems and Mining, (WISM 2009), Shanghai, China, 7–8 November 2009; pp. 32–38.

[20]. Jankowski, J.; Michalski, R.; Kazienko, P. The Multidimensional Study of Viral Campaigns as Branching Processes. In Social Informatics; Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7710, pp. 462–474.

[21]. Weng, J.; Lim, E.P.; Jiang, J.; He, Q. TwitterRank: Finding Topic-Sensitive Influential Twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining; ACM: New York, NY, USA, 2010; pp. 261–270.

[22]. Hotho, A.; Jäschke, R.; Schmitz, C.; Stumme, G. Information Retrieval in Folksonomies: Search and Ranking. In The Semantic Web: Research and Applications; Springer: Berlin/Heidelberg, Germany, 2006; pp. 411–426.

IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 01-05
www.iosrjournals.org

Integrating Text Mining with Image Processing

Anjali Sahu[1], Pradnya Chavan[2], Dr. Suhasini Vijaykumar[3]
[1](MCA, Student, Mumbai University)
[2](MCA, Student, Mumbai University)

Abstract: *Image Processing is the most stimulating subjectsof research. This paper presents the idea of recognizing the text in digital image using Optical Character Recognition (OCR) and then how this extracted text can be used for deriving information using Text Mining. In this paper we will discuss generic techniques and approaches that can be used to develop applications on integrating Text Mining with OCR. In this paper we will discuss various methodologies and we have categorized into five types such as Text Collecting, Text pre-processing, Text analysis, Visualization, and Model Evaluation. In the first step we have discussed howOptical Character Recognitioncan be integrated with Text mining.*

Keywords: *OCR (OpticalCharacter Recognition), Part-of-Speech, Tagging, Stemming, Tokenization*

I. Introduction

There are many fields like bio-medical, Search Engine, Geographic text search which requires applications which has both the features of OCR for reading text from image and then the text has to be further used for deriving high-quality information using Text Mining.For now we are considering a search application which processes results of OCR into various methods of text mining which we are currently analyzing in this paper. In this type applications the user first has to provide pdf document with images so it minimizes the number of documents to be checked.We have to first preprocess the retrieved text from the text collection step so that real time text mining tools can be used so that less precise algorithms can be used which save over computing time.

II. Literature Review

One of the example of this type of application which is currently available, in the market is KNIME Text Processing Feature[11] which does parsing of texts available in formats PDF documents and then the frequent words can be computed, keywords can be extracted, and can be visualized in a format such as tag of clouds. One of the research work on this type application is done by Brigitte Mathiak and Silke Eckstein on Five Steps to Text Mining in Biomedical Literature [1] which tells about data can be gathered for Biomedical and on that how text mining can be done. There is another existing proceeding paper on Text mining based journal splitting [2] which tells about how text from journal and magazines can be gathered using OCR then using text mining algorithms which identifies the important information from the huge text.

III. Methodology

There are various methods to implement this type of applications. In our research we have systemized the methods which are already there to contribute this growing field. [1]
This methods can be further divided into five distinct steps:
3.1 Text collecting,
3.2 Text pre-processing,
3.3Text analysis,
3.4 Visualization,
3.5Model Evaluation.
The objective of this paper is therefore to analyze the different methods applicable to the above listed five steps.

3.1 TEXTCOLLECTING

In this paper we have used OCR for text collecting which is built in python which can scan any number of pdfs documents. This pdfs consist of images from which the text has to be extracted. We have correlation method of Continuous character Recognition in OCR for text collecting process. [6]

Steps to design OCR is as follows:

3.1.1Preprocessing

3.1.2Segmentation

3.1.3Feature extraction
3.1.4Classification

3.1.1 PREPROCESSING

In this stage the raw image is taken and first it is converted to gray scale image after that it is converted into binary image this operation is called as thresholding. This process of image transformation is called as Image Digitization. Now to reduce the noise from image various techniques like morphological operations are used to connect unconnected pixels, to remove isolated pixels, to smooth pixels boundary.

Figure 1. Generated Threshold Image

3.1.2 SEGMENTATION

The segmentation stage takes in an image and separates the different parts of an image, like text from graphics, lines of a paragraph, and characters of a word. Then the character is segmented image is normalized to 32* 32 or 64*64 matrix. And the array of all the alphabet matrix is stored in Flattened.txt and classifications.txt File which are the input models.

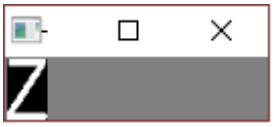


Figure 2. Segmented Image

3.1.3. FEATURE EXTRACTION

The feature extraction stage is used to extract the most relevant information like alphabet recognition of specific font type from the text image which helps us to recognize the characters in the text. The selection of a stable and representative set of features is the heart of pattern recognition system design

3.1.4 CLASSIFICATION

The classification stage uses the features extracted in the previous stage to identify the text segment according to preset rules.. And using the two Flattened.txt and classifications.txt File the input character is recognized.

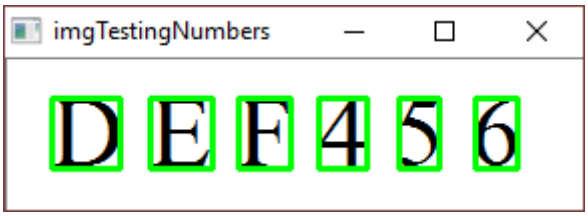


Figure 3. Recognized Text from input image

After all the above stages of text collecting using OCR the unstructured text data is retrieved which is provided as input for Text preprocessing stage.

3.1 TEXTPRE-PROCESSING

In the whole procedure of text mining pre-processing of the text is the most time taking process. The Text pre-processing is done to extract the interesting andtrivial knowledge from unstructured text data. Text pre-processing involves twoapproaches such as Tokenization and Part-Of-Speech for Tagging, orbag-of-words approach withword stemming and the application of a stop word list.As in the first approach first tokenization is done in which the stream of text is broken or dividedinto wordsor other significant elementswhich are called as

tokens. The objective of tokenization is to explore the words in the sentence. Then Part-Of-Speech Tagging is done in which words are tagged as per the grammatical context of the word in the sentence, hence we are dividing the words according to pronouns, adverbs, etc.

The second approach focuses on the words and their statistical distributions rather than the order of words. So this type of approach is called as bag-of-words approach. Now to use this unordered words first we have to provide theindextotext into a data vector which generates an index.Suppose the generated index is very largethen the words which are to each other grammaticallyare mapped to one word using stemming algorithms. And further reduce the index words index list is further compiled and the words which occur very often is removed from the list. There are various Stemming Algorithms present out of which the algorithm which optimizes the performance of statistical data analysis is selected for stemming after which the vector space representation measures are implemented on word list.

3.2 TEXT ANALYSIS

This step depends on the preprocessing and the type of data representation model chosen for preprocessing. For now we have considered the vector space representation model in which the data is analyzed standard data mining techniques, such as support vector machine, artificial neural networks etc. This techniques can be used in Weka software package.

Text analysis is one of the most varied and optimized step out of five steps. In this step text mining tools are used to produce the result of queries which can be too much information or can be too little.To provide assistance to user by showing concept in search result already found some research has been conducted in client based search application. Unsupervised clustering, Clustering via k-means and hierarchical clustering these methods are used for task clustering. Variation of these method are also exist , in order to save computing time dimension reduction or Monte-Carlo simplification are used to see how much variation can be applied without trading too much quality for time.

3.3 VISUALIZATION

It is useless to extract information that no one sees, so to visualize the results obtained lots of possibilities have been invented. For user to look up information he needs simply just make a table. On the other end user may navigate on three dimensional worlds, results of hypertext is the classical option for the visualization of query. He may just click on link if the user is interested in details were complexity can be hidden and how data to be shown to user is other issue of visualization. Here user is confronted with pure result how and why results were retrieved without the Meta information. It is important for user wants to know what it was exactly that made one result superior to another when results contain some kind of evaluation.

For this problem, solution is transparency. Transparency means the reason for decision, it does not means algorithm that made the decision is explained. Here in the results we can highlight the search keywords. The fact is quite complicated as it is simplification the highlighting is just symbolic for real events. Hiding their reasons in neural networks or complex weighting scheme this approach not all data mining algorithm supports. This situation may be improve by further research.

3.4 MODEL EVALUATION

The diverse forms of cross validation and test sets is the classic methods of evaluation. In order to optimize their parameters supervised machine learning relies in them. For unsupervised machine learning automatic evaluation is more unusual, as there are no standard for evaluation, but it is possible to do evaluation on mixed queries.

This mixed query is of two keywords which is linked to each other vaguely and queries are compared with resultant clusters with single keywords. It also brings unusual evaluation criteria user feedback. Algorithm may improve itself over time by learning which clusters the user chooses.

IV. Implementation

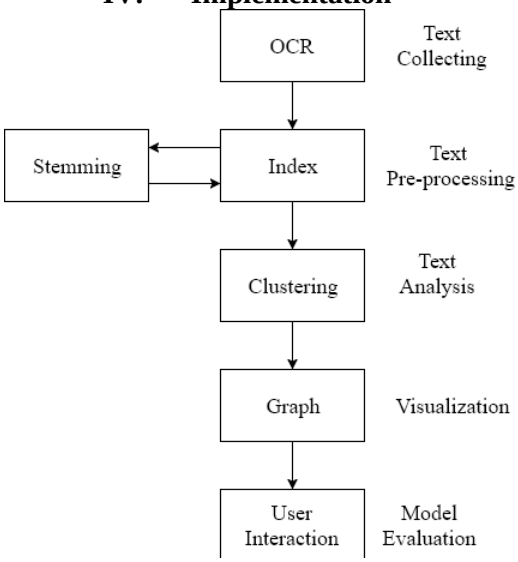


Figure 4. Steps for Implementing Text mining with Image Processing Application

The workflow of implemented example for the Integrating text mining with Image Processing is shown in Fig 4. For now the project is at the beginning stage so far we have completed with the first stage that is OCR which we have developed in Python which then has to be integrated with text mining.

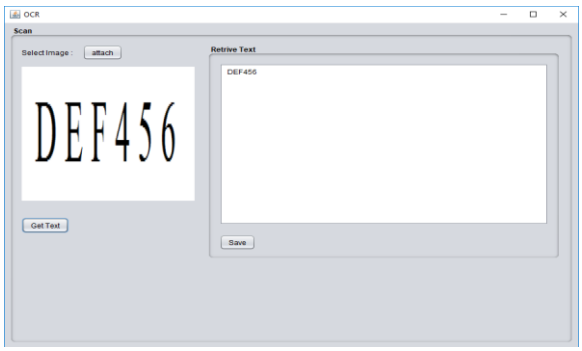


Figure 5. Screeshot of Implemented OCR Application

The implementation of text gathering is done through OCR application developed in Python and Java Fig. 5 shows text retrived.A dataset of text is first collected through OCR first transformed into index using bag-of-words approach in the preprocessing step. Then the index is stemmed to perform clustering. Clustering is done by creating matrix out of index.

V. Conclusion

We have described, various methods which can be used to develop applications based on Integrating Text mining with Image Processing. We also tried to implement this methods and develop a application which we have completed till text collecting stage that is OCR till now.So this paper will any one who is trying to develop application on Text mining with Image processsing.

References

Journal Papers:

[1]. Brigitte Mathiak and Silke Eckstein,*Five Steps to Text Mining in Biomedical Literature,* https://www.researchgate.net/publication/249954119_.

[2]. Xiaofan Lin, *Text-mining based journal splitting,* https://www.researchgate.net/publication/220860270.

[3]. Ian Lewin,*Retrieving Hierarchical Text Structure fromTypeset Scientific Articles - a Prerequisite for EScienceText Mining,* https://www.researchgate.net/publication/244495915.

[4]. Dr S.Vasavi, Srikanth Varma.Ch ,Anil kumar.Ch ,Santosh.D.M ,Sai Ram.S, *Book Search by Capturing Text from Digital Images Using Optical Character Recognition,* S.Vasavi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2377-2379.

Proceedings Papers:

[5]. D.S. Chan, *Theory and implementation of multidimensional discrete systems for signal processing*, doctoral diss., Massachusetts Institute of Technology, Cambridge, MA, 1978.
[6]. G.S. Lehal and Chandan Singh, *A Gurmukhi Script Recognition System*, Proceedings of the International Conference on Pattern Recognition (ICPR'00), 1051-4651/00, 2000.
[7]. Anil K. Jain and Sushil Bhattacharjee, *Text Segmentation Using Gabor Filters for Automatic Document Processing*, Pattern Recognition and Image Processing Laboratory. Michigan State University. E. Lansing,MI 48824-1027. USA.*
[8]. AJ Palkovic, *Improving Optical Character Recognition*, Villanova University, United States.

Books:

[9]. Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, Fred Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*(New York: Springer2005)

Websites:

[10]. Wikipedia, *Text-mining*, https://en.wikipedia.org/wiki/Text_mining.
[11]. Dr. Killian Thiel Killian, Dr. Michael Berthold, *Technical Report The KNIME Text Processing Feature*, https://www-cdn.knime.com/sites/default/files/.

Proceedings of 2nd National Conference on Emerging Trends: Innovations and Challenges in IT, 29 -30, June 2017

Educational Data Mining: A Critical Study

Priya Chandran
BVIMIT, Navi-Mumbai
priyaci2005@gmail.com

Sanakhatun S. Shaikh
BVIMIT, Navi-Mumbai
sana.shaikh.vashi@gmail.com

Abstract-Educational data mining (EDM) is an emerging research area, where different techniques are used to explore huge data coming from educational system. EDM is concerned about the effective use of educational data to improve and optimize learning process and the methods developed using this approach can be used for predicting and analyzing students learning behavior. Various factors affecting teaching-learning process can be analyzed and effectively applied using the models built using EDM. Data mining methods enable institutions to use their huge data, collected through various activities, to uncover and understand hidden associations and patterns. Data mining models are built using these patterns and associations to predict student's behavior and hence resources can be effectively allocated to attain outcome. This paper introduces and investigates state of the art schemes carried out in this field and their relevance.

Keywords: *Educational Data Mining, EDM, E-learning, data mining, Learning Management System, Knowledge Discovery Database.*

I. Introduction

In today's world, technology increasingly growing and the more technology enhancing, the concept of e-learning resources, educational software are exponentially growing and this produces the huge collection of datasets. This information is very helpful to discover the student's behavior and trends. It is not feasible to analyze the data collected from large repositories and data warehouses manually; it is effective for small databases but also becomes the bottleneck for large data. In this situation, data mining technique is used in education sector. Data mining provides the ability to analyze and predict data from multiple sources with different dimensions to their users [2]. Due to availability of e-learning facilities and use of Learning Management System (LMS) increased the wide usage of data mining techniques in educational field.

Data mining

Data mining is the process of identifying patterns and correlations within huge repositories to analyze and predict the outcome [6]. It is also known as Knowledge Discovery Database (KDD) process and this technique extracts knowledge which is hidden in huge data base. The knowledge is extracted in the form of hidden patterns and associations. In recent environment, data mining is most important and effective process for analyzing, predicting, visualizing the useful knowledge that can helpful for making strategic and knowledge driven decisions. Data mining is emerging into the education field and is known as Education Data Mining (EDM). Predicting student's performance is the most popular and traditional application of data mining in education and is used to estimate the hidden value of student's performance and knowledge [10]. Data mining is an emerging trend and is applied into almost every domain including business, medicine, banking, education, and also database, data warehouse systems, visualization, statistics, machine learning, algorithms and many other application domains [5]. Different data mining techniques like classification, clustering and association are used for knowledge extraction and these techniques are broadly classified into supervised and unsupervised algorithms. Each of these techniques uses different

approaches to build the model depending on the applications.

Steps involved in data mining are depicted in fig 1. In data mining, the collected data are preprocessed and appropriate mining technique is applied to identify patterns and associations hidden in the database. Also the strategy chosen in decision making process is based on the knowledge.

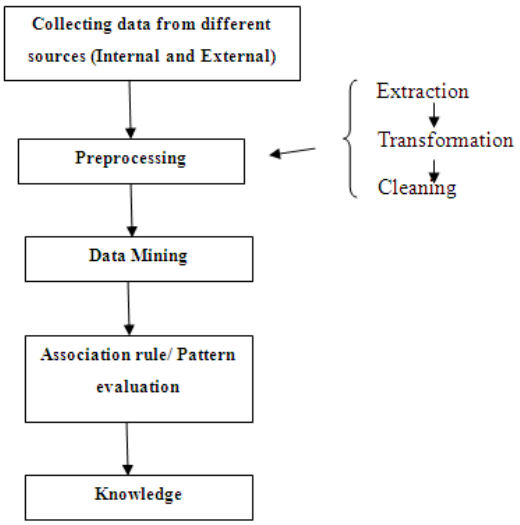


Fig1: Data mining steps

II. Educational Data Mining (EDM)

Data mining is applied for very large datasets using the supervised and unsupervised algorithms [3]. EDM is the emerged area for research as the technology improved and is applied into education field. Different types of data originated from different sources is analyzed to identify behavior of the students and used for further research on various areas of education data mining [10]. EDM is an emerging application of data mining which concerned to generate specific type of data that is extracted from various educational environment to predict behavior and the pattern of student learning process and often to predict faculty's behavior [10][12]. The technique called clustering is an important aspect in educational data mining which is used to group the students according to their behavior and performance [13].

Traditional education system has several issues. Some of them are: identifying needs of the student, prediction of quality of learning, training of student and faculty interactions [2]. Because traditionally, data related to student's learning was collected through various techniques such as interviews, surveys, attendance, classroom activities, focus group which was not effective and also a tedious process to gather the student's learning pattern [8].

Data generated in schools, colleges, universities, learning institutions providing traditional and modern forms and methods of teaching, as well as informal learning shall be analyzed with the help of EDM [10]. The above said issues regarding traditional education system are overcome by EDM in order to enhance students learning experience and profit is exponentially increased in business perspective [2]. C. Romero et al. [11] discussed application of DM in LMS. The authors described data mining process of e-learning data, step by step, as well as how to apply the main data mining techniques such as statistics, visualization, classification, clustering and association rule mining on MOODLE data.

III. EDM Techniques

Prediction and analysis of student performance, learning, need, and interest is very important in education environments .When the interest, need, performance, etc. of the students are extracted from huge repositories, it is important to use right data mining technique in education which helps to showcase when students are failing, how they can improve their studies, and also it can allows faculties

to better help them with material that they don't understand [6].

EDM also play an important role while identifying the suitable colleges, universities, schools which is helpful for students to find a perfect fit which is not an easy task. Meaningful patterns and associations can also be identified from large repositories of educational data, and their results are used for optimizing teaching-learning process. EDM framework is shown in fig 2. Data generated from various activities is preprocessed for applying data mining techniques. Internal and external data gathered from various sources like:

- Academic information
- Course details
- Placement and alumni details
- Social media data
- Faculty data

Using this vast source of data, student's performance in all aspects can be evaluated. Usage of social media as a source of data helps in identifying their personal interests. Scio-economic analysis can also be considered.

Preprocessing techniques like transformation and cleaning is applied on this generated data and data mining techniques like classification, clustering or association is applied related to the type of knowledge to be extracted. For example, association algorithm is used for deriving association rules and there by predicting the associations in educational data.

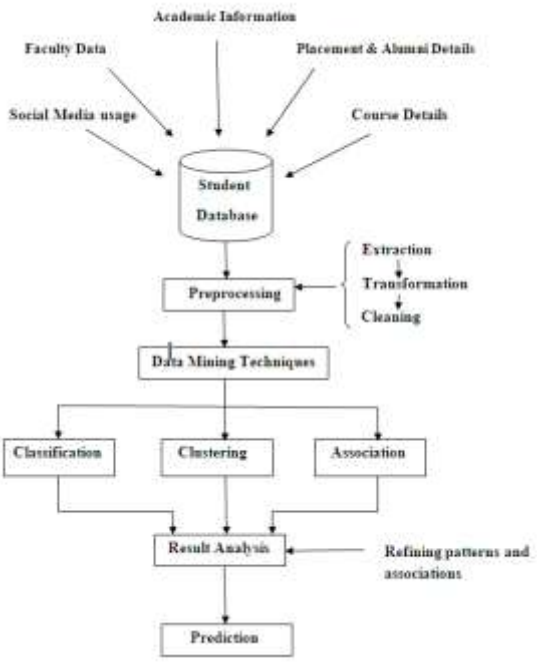


Fig 2: EDM Framework

Data mining techniques are broadly classified into:

A. Association Rule

Association rule algorithm is usually used to find frequent item sets finding among large data sets [14]. Group of one or more items is known as item set [14]. These associations are often used in the retail sales community to identify items that are frequently purchased together. This technique is mainly used for market basket analysis. In Education Data Mining, association rule is used for analyzing and predicting the student and teachers behavior [9]. The association among different students activity can be identified using this data mining technique.

B. Classification

Classification algorithm can be used to categorize student's performance into predefined classes. Decision tree, neural network and statistics are some of the commonly used classification methods.

C. Clustering

Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. Clustering is also referred to as unsupervised learning or segmentation. Clustering technique of data mining in EDM is used to group students on the basis of their learning and interaction patterns [10]. Typically, some kind of distance measure is used to decide how similar instances are. Once a set of clusters has been determined, new instances can be classified by determining the closest cluster.

Romero et al.[11] categorized techniques used in EDM into the following categories:

- Statistics and visualization
- Web mining
- Clustering, classification, and outlier detection
- Association rule mining and sequential pattern mining
- Text mining

There are various techniques are available in data mining under predictive and descriptive model. Classification, Regression, Time Series Analysis, Prediction are the techniques involved in predictive model and Clustering, Summarization, Association rule, Sequence Discovery are the techniques used under descriptive model. These are the tasks or techniques that can be used to predict and find the solution of the problem identified. These descriptive and predictive data models or approaches are used to discover hidden information [9]. Some of the examples are, EDM technique can improve the teaching and learning processes in the classroom, identify at-risk students, customize teaching processes, and also provide recommendations to both faculty and students [4]. Similar to above example, there are many other examples of applications or tasks in educational sector that are solved through Data mining. Comprehensive analysis of online and offline behavior pattern of students can predict the attainment of course outcome and accordingly teachers can plan the strategies to improve teaching-learning process. This process also helps to identify most popular courses, types of students join specific courses, dropout and placement ratio etc.

IV. Conclusion

Recent advances in the use of technology in educational field create huge data. Data mining can be applied on this data, for understanding and predicting student's performance, and is known as educational data mining. This prediction can be taken into account for improving the effectiveness of teaching-learning process. In this paper, we have discussed various educational data mining techniques and how this technique can be effectively used in education sector. It is observed that, the model built using EDM techniques incorporate students overall behavior and pedagogy to analyze the knowledge and teaching and learning outcomes.

References

1. Abu Saa, Amjad. "Educational Data Mining & Students' Performance Prediction." International Journal of Advanced Computer science and Applications 7.5 (2016): 212-220.
2. Bhise, R. B., S. S. Thorat, and A. K. Supekar. "Importance of data mining in higher education system." IOSR Journal Of Humanities And Social Science (IOSR-JHSS) ISSN (2013): 2279-0837.
3. Bhullar, Manpreet Singh, and Amritpal Kaur. "Use of data mining in education

sector." Proceedings of the World Congress on Engineering and Computer Science. Vol. 1. 2012.

4. Dheenathayalan, K., J. Ramsingh, and V. Bhuvaneswari. "ICICA 2014." https://www.safaribooksonline.com/library/view/data-mining-concepts/9780123814791/xhtml/ST0090.html
5. https://www.sas.com/en_us/insights/analytics/data-mining.html
6. http://www.onlineuniversities.com/blog/2012/08/10-ways-data-mining-will-transform-higher-ed/
7. Kashyap, Geeta, and Ekta Chauhan. "Review on Educational Data Mining Techniques." International Journal of Advanced Technology in Engineering and Science, 3 (11) (2015): 308-316.
8. Overview of Predictive and Descriptive Data Mining Techniques Pradnya P. Sondwale
9. Romero, Cristobal, and Sebastian Ventura. "Data mining in education." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3.1 (2013): 12-27.
10. Romero, Cristóbal, Sebastián Ventura, and Enrique García. "Data mining in course management systems: Moodle case study and tutorial." Computers & Education 51.1 (2008): 368-384.
11. Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40.6 (2010): 601-618.
12. Suman, Pooja Mittal P., and M. Pooja. "A Comparative Study on Role of Data Mining Techniques in Education: A Review." International Journal of Emerging Trends & Technology in Computer Science 3.3 (2014): 65-9.
13. Tair, Mohammed M. Abu, and Alaa M. El-Halees. "Mining educational data to improve students' performance: a case study." International Journal of Information 2.2 (2012): 140-146.
14. Verma, Sushil Kumar, R. S. Thakur, and Shailesh Jaloree. "Pattern Mining Approach to Categorization of Students' Performance using Apriori Algorithm." International Journal of Computer Applications 121.5 (2015).