A Data mining Technique for Analyzing and Predicting the success of Movie

K Meenakshi[1], G Maragatham[2], Neha Agarwal[3] and Ishitha Ghosh[4]

[1,2,3,4] Department of Information Technology, SRM Institute of Science and Technology

E-mail : **meenakshi.k@ktr.srmuniv.ac.in**

Abstract. In real world prediction models and mechanisms can be used to predict the success of a movie. The proposed work aims to develop a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. An attempt is made to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making [the success of the movie] is without risk, because the decision maker [movie makers and stake holders] has all the information about the exact outcome of the decision, before he or she makes the decision [release of the movie]. With over two million spectators a day and films exported to over 100 countries, the impact of Bollywood film industry is formidable We gather a series of interesting facts and relationships using a variety of data mining techniques. In particular, we concentrate on attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed. The paper additionally reports on the techniques used, giving their implementation and utility. Additionally, we found some attention-grabbing facts, such as the budget of a movie isn't any indication of how well-rated it'll be, there's a downward trend within the quality of films over time, and also the director and actors/actresses involved in the movie.

1. Introduction

Given the low success rate of movies, models and mechanisms can be used to predict the success of a movie. It will help the business significantly. Various stakeholders such as actors, producers, directors etc. can use these predictions to make more informed decisions. They can make the decision before the movie release. Historical data of each component such as actor, actress, and director, composer that influences the success or failure of a movie is given due to its weightage.. This proposed work aims to develop a model based upon the data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. The system is used to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making (the success of the movie) is without risk, because the decision maker (movie makers and stake holders) has all the information about the exact outcome of the decision, before he or she makes the decision (release of the movie). With over two million spectators a day and films exported to over 100 countries, the impact of Bollywood film industry is formidable. In particular, we concentrate on attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed. The proposed system reports on the techniques used, giving their implementation and usefulness. The important issue involved in the prediction system is, IMDb is difficult to perform data mining upon, due to the format of the source data. We also found that, the budget of a film is no indication of how well-rated it will be, there is a downward trend in the quality of films over time. Another important factors are the director and actors/actresses involved in a film.

Extensive description and data is available for Hollywood movies (IMDB, Rotten Tomatoes etc.). The IMDB site predominantly contains data for Hollywood movies. A structured database or a central repository of Indian movie data is difficult to find. Four billion tickets for Bollywood films are sold annually. It affects producers, distributors, actors, rentals agencies and Bollywood fans.Indian movie industry produces the maximum number of movies per year. However, very few movies taste success. The stakeholders concerned depend on film critics for movie reviews. However, the movies are reviewed after their release.

IMDB rates movies according to true Bayesian estimate.

$$WR = (v \div (v+m)) \times R + (m \div (v+m)) \times C \qquad (1)$$

where:
R = Mean Rating

v = votes for the movie

m = minimum votes required to be listed in the Top 250 (currently 25000)

C = the mean vote across the whole report (currently 7.0)

Indian Hindi Cinema industry popularly known as Bollywood has reached staggering proportions in terms of volume of business (184.3 billion), manpower employment (over 6 million workers), movies produced (more than 100 in a year) and its reach (exported to more than 100 countries worldwide). With so much at stake and highly uncertain nature of returns, it is of commercial interest to develop a model which can predict success of a movie. This however, is not an easy work, since movies have been described as experience goods with very less shelf life; it is difficult to forecast demand for a movie. There are number of parameters that may influence success of a movie like – time of its release, marketing gimmicks, lead actor, lead actress, director, producer, genre, music director – being some of the factors.

2. Related Works

2.1 Anatomy of Movie success predictor

With over two million Audience a day and films exported to over 100 countries, the impact of Bollywood film industry is formidable. From the first Indian film “Raja Harishchandra by Dhundhiraj Govind (Dadasaheb) Phalke in 1913 to 1981, India produced over 15000 feature films[1]. Since then it has produced, at least another 15000 at a rate of more than 1000 films a year (1091 in 2006, 1146 in 2007 and 1325 in 2008) in 26 languages. Literature survey has revealed only two studies which have attempted to predict the success of movies. While one study uses Bayesian belief network to predict the success, the other one uses neural network for the same. Lee and Change in their study using. Bayesian Belief Network for predicting box office performance concluded that Bayesian Belief Networks were better in predicting the success as compared to neural networks [2].

Machine learning has also been used for predicting movie success by using algorithms like RF and SVM . Although the use of RF and SVM within the movie domain seems to be fairly limited, the two algorithms have been applied and evaluated in many applications for the purpose of regression as well as classification. Within recent study Verikas et al. (2011) have surveyed a number of large as well as small scale comparisons on data mining and machine learning, all of which include the RF algorithm, specifically issuing its prediction performance in comparison to other algorithms as well as the use of the variable importance estimates available from RF. Among the previous applications and algorithm comparisons included by Verikas et al. (2011) are several large scale studies such as Meyer et al. (2003) and Statnikov et al. (2008), evaluating RF and SVM among other algorithms over a number of 33 and 22 datasets respectively.

2.2 Staggering proportions and uncertain returns

Indian Hindi Cinema industry popularly known as Bollywood has reached staggering proportions in terms of volume of business (184.3 billion), manpower employment (over 6 million workers), movies produced (more than 100 in a year) and its reach (exported to more than 100 countries worldwide). With so much at stake and highly uncertain nature of returns, it is of commercial interest to develop a model which can predict success of a movie. This however, is not an easy work, since movies have been described as experience goods with very less shelf life; it is difficult to forecast demand for a movie. There are number of parameters that may influence success of a movie like time of its release, marketing gimmicks, lead actor, lead actress, director, producer, writer, music director being some of the factors.

2.3 A straightforward approach

The approach to viewer rating prediction is very straightforward and can be easily applied by anyone. The movie prediction requires more experience, because of its dependency on the user to be able to interpret the different visualizations of the data in the right way and infer the right conclusions from it. Visual Analytics will become more and more integrated in our lives for the great possibilities it enables. The human mind is very intelligent for extracting information from visualizations and thus can analyze data quicker and in a more complex way.

3. Proposed Implementation

The methodology has 4 major components, these are 1. Data Collection 2. Data Cleaning
3. Data Transfer 4. Data Analysis and Prediction. Cleaning the dataset and discarding the irrelavant data from the IMDB dataset as well as through detailed study of the dataset, we get the attributes that can affect the prediction of success of a movie. The textual data is transferred to numeric data and converts it into CSV format. Different decision tree algorithms are then studied so that the best suited algorithm according to our problem could be determined. The best suited algorithm should be the one which gives the most accuracy and least error. The test data when tested according to the algorithm should give as accurate results as possible.

3.1 Data Collection

The raw IMDb dataset is structured in such a way that most of its attributes andinformation is organized and stored separately in compressed plain text files. For instance, all of the roughly 600,000 movie ratings from the database are stored in the compressed text file ratings. List (e.g. ratings.list.gz), which includes textual informationabout the data as well as a table of film rank, the number of votes and film titles. Thus, some sort of cleaning, integration and preprocessing is likely to be required in order to make good use of the data for the purpose of data mining through supervisedmachine learning techniques. The data was collected using IMDB (java movie database) which contains the IMDB movie dataset of more than 30,00000 movies in the dataset. The dataset was transferred to MySQL, in form of tables.

3.2 Data Cleaning

Several SOL queries were run on the relational database to clean the data in order to reduce the data and select only the relevant attributes which would help in data analysís and prediction of movie success.

3.3 *Data Transfer*

The relational database table data was exported to excel files and stored in the CSV (comma separated values) format for further analysis.

3.4 Data Analysis and Prediction

The dataset is divided into training dataset and test dataset which contains the classes like Hit, Flop and Average and predicting variables like actor, actress, composer, genre, director producer and music director k-means clustering is used to analyze the training dataset to develop models which can be used for test dataset for analysis decision tree algorithm is used for predicting which factors.

4. Architecture Design

Figure 1 shows the architecture of a predictive modeling system. The data is collected for different parameters, and a database is formed. Here the data warehouses are central repositories of integrated data of one or more different sources. This database may contain several irrelevant information which we may not require in our predicting algorithm. Hence the database needs to be cleaned, and relevant analysis needs to be done. The database is refined and cleaned according to the requirement of the algorithm. The refined database is subjected to various tests. These different big data and data mining algorithms are carried out to study and learn more about the data. This gives data analyst a better grasp of the data they are working with. After carrying out different tests on the data, the best suitable algorithm which gives the most accurate result is chosen. In the proposed work , two different techniques are used to study dataset. It includes K-means clustering and Decision tree algorithm. The results of these algorithm gives us details about patterns and trends within our dataset as output. Patterns in data mining allows the users to analyse the data from different angles and dimensions. It helps us identify relationships within the data. Patterns also help in categorization and summarization of the given data. In fact, the term data mining refers to the large number of correlations and patterns found in the relational databases.

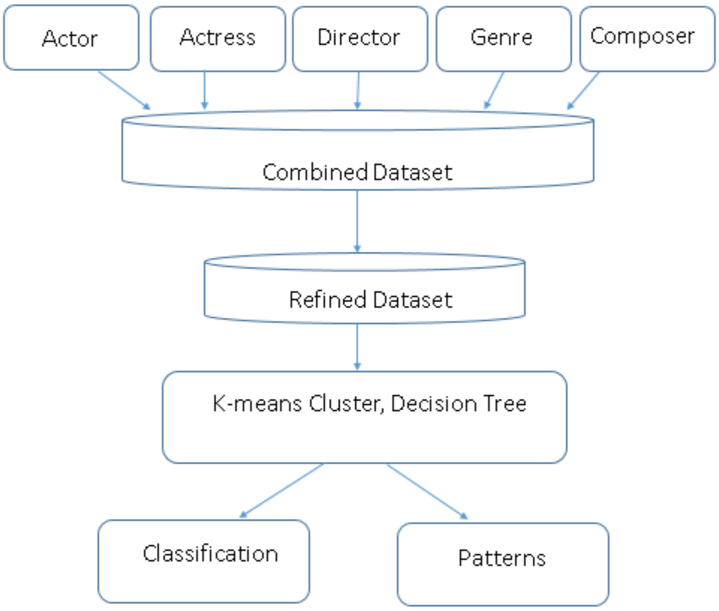

Figure 1 Architecture of Prediction model

5. Results

K-mean cluster is a used to classify the movies into two different categories. Clustering is a method, where Euclidean distance formula is used to make clusters from the data given to the algorithm. In the clustering method, clusters are formed, so that the inter cluster similarity is less, and intra cluster similarity is high.Table 1 shows the Actress Predictor showing cluster. Final objective is to classify movies into three categories, we decided to form three clusters. We wanted to see, if the actresses in a particular cluster,always gave the same class of movie. K means clustering aims to partition the observations into k clusters (we chose 3 here) in which observation belongs to a cluster with nearest mean serving as the prototype of the cluster. According to these algorithm, initially K centroids are chosen. Then the distance of each observation is calculated from each of those centroids using the distance formula. The observation is placed in the cluster where the centroid and the observation have the least distance. We applied this algorithm, between the variables hit movies and total movies of each different predictor. When clusters, were formed using R language the cluster mean was for each cluster was also obtained as the output.

Table 1 Table for the Actress Predictor showing cluster

ACTRESS	HITS	SUCCESS	CLUSTER	CLUSTER VERDICT
Juhi Chawla	1	0.047619048	3	FLOP
Kareena Kapoor	14	0.451612903	1	AVERAGE
Bipasha Basu	8	0.228571429	3	FLOP
Aishwarya Rai	6	0.24	3	FLOP
Rani Mukherji	10	0.357142857	1	AVERAGE
Priety Zinta	8	0.363636364	1	AVERAGE
Kajol	4	0.571428571	1	AVERAGE
Priyanka Chopra	13	0.419354839	1	AVERAGE
Vidya Balan	7	0.777777778	2	HIT
Katrina Kaif	12	0.631578947	1	AVERAGE
Kangana Ranaut	5	0.555555556	1	AVERAGE
Deepika Padukone	4	0.4	1	AVERAGE
Genelia D' Souza	1	0.5	1	AVERAGE
Anushka Sharma	3	1	2	HIT

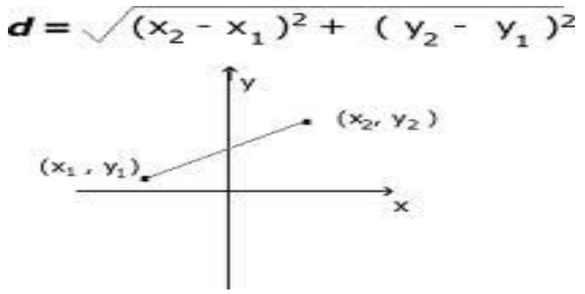

Figure 2 Distance formula for the K-means cluster algorithm

From the output we received, we tried classified each actress into different classes. However we noticed, an actress placed in the average category maybe a better actress than an actress in a hit category.

The cluster means obtained for the actress input are:

1. 1.00
2. 0.363
3. 0.688
4. The actress classified in a cluster with a cluster mean 1 was classified as a Hit category actress.

Similarly actresses in cluster with mean 0.688 were considered of average category and so on. However, there was not much significance of this classification. An actress in the hit category, gave flop movies as well. We used this classification technique for actors, directors, composers, and genres as well. In the K- means cluster for directors, composers, and genres we however observed that the classification made some significance. Composers, which were classified in the hit category, we observed were actually a part of most hit movies released after 2010. Same was the case for directors and genres as well. Hence we could somewhat deduce from the K means classification algorithm, that composers, directors, and genres had significance in predicting movies in Bollywood. While working with genres, we noticed that genres that were less common, had a success rate compared to genres like Romance and Comedy. Genres like Crime, thriller and horror which were less common did better at box office. After forming the clusters, we plotted the film crew against the cluster. The y-axis depicted the range of the cluster mean. x- Axis denoted the particular film crew.

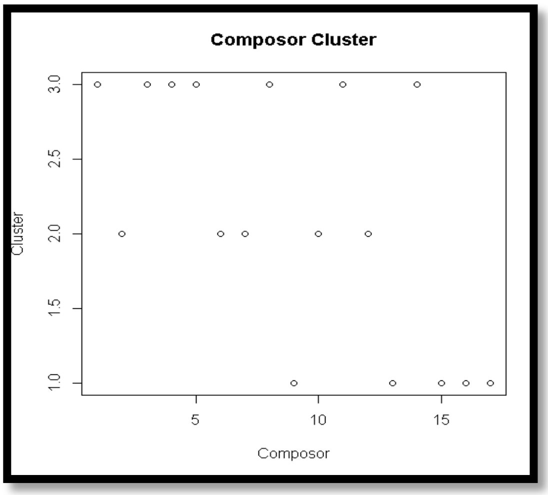

Figure 3 Plot depicting clusters for the plotted Composer predictor variable

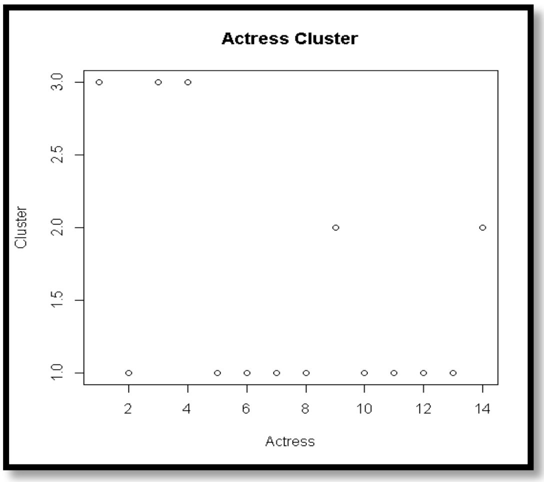

Figure 4 Cluster of Actresses according to different cluster mean

The experiment is carried out by Initial tuples of the training data set inclusive of class label

Table 2 Initial tuples of the training data set inclusive of class label

Actor	Actress	Composer	Directors	Genres	Producers	Verdict
0.66	0.24	0.3	0.35	0.25	0.3	Flop
0	0.3	0.4	0.45	0.4	0.4	Average
0.2	0.5	0.27	0.3	0.25	0.3	Flop
0.3	0.6	0.55	0.75	0.6	0.6	Hit
0.6	0.12	0.2	0.4	0.25	0.3	Flop
0.6	0.59	0.55	0.7	0.6	0.65	Hit
0.4	0.24	0.21	0.3	0.25	0.3	Flop
0.5	0.12	0.18	0.55	0.25	0.3	Flop
0.6	0.59	0.3	0.35	0.3	0.3	Flop
0.4	0.54	0.54	0.6	0.65	0.7	Hit

Table 2 consists of the few initial tuples of our training data set. Here the values for each tuple under Actress, Actor, composer, genre and director is their success rate. It is the ratio of their successful recent films in the last three years to the total number of films they were a part of the last few years. If the value is 0, it means that the particular film crew is a debutant. Then the RPART package was used to extend the features of R language and model a decision tree. RPART stands for Recursive and Partition tree.

From the decision tree, it is seen the most important node in the decision making algorithm is Genre. Then we can see the other significant determining factors are directors and composers. According to the decision tree, there are four decisions to classify the movies. After deriving a decision tree, we then had to prepare a table consisting of movies released 2011 onwards. This table prepared in excel was called the testing data set. The test data set's verdict which is the class label is determined by using the decision tree above.

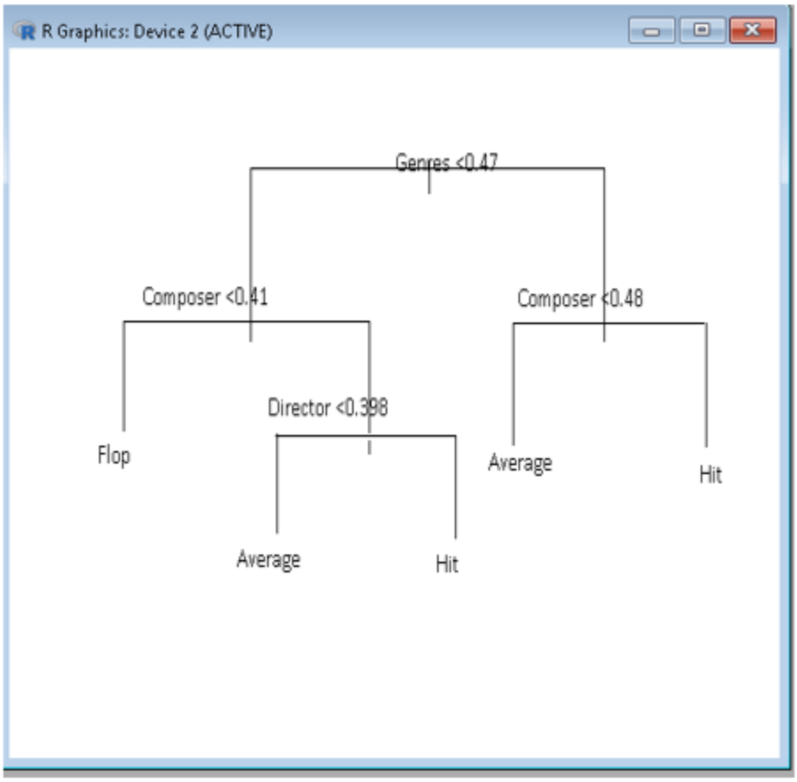

Figure 5 Screen Shot of Decision tree generated by data analysis in R

Table 3: - Initial tuples of the test data set inclusive of class label, and predicted class label

Title	actor	actress	Composers	Directors	Genres	Predicted Verdict	verdict
no one killed Jessica	0	0.44	0.37	0.29	0.39	Average	Average
dhobi ghat	0.24	0	0.2	0.56	0.42	Flop	Flop
patiala house	0.56	0.38	0.5	0.5	0.21	Average	Flop
7 khoon maaf	0.47	0.71	0.38	0.44	0.24	Flop	Flop
tanu weds manu	0.39	0.23	0.45	0.29	0.45	Average	Hit
Game	0.43	0.23	0.56	0.38	0.34	Average	Flop
Thankyou	0.61	0.1	0.2	0.56	0.42	Flop	Flop
dum maaro dum	0.43	0.44	0.24	0.56	0.43	Flop	Flop
Ready	0.71	0.23	0.56	0.67	0.49	Hit	Hit
double dhamaal	0.4	0.23	0.37	0.29	0.39	Average	Average
murder 2	0.31	0.41	0.56	0.67	0.49	Hit	Hit
ZNMD	0.63	0.42	0.68	0.45	0.49	Hit	Hit
Aarakshan	0.62	0.51	0.42	0.44	0.49	Average	Flop
Bodyguard	0.71	0.64	0.45	0.54	0.39	Hit	Hit

In the test data set, again we have calculated the success rate for each of the predictors. The success rate is the ratio of the total no of successful films out of the total films done by the film crew recently in the last three years. Here there is an additional column, called Predicted Verdict. The predicted verdict consists of the verdict predicted by the decision tree algorithm. The other column named verdict is the actual verdict of the movie. The actual verdict of the movie has been gathered from the official website of Box Office India. The same website gave us the class labels for the training data set as well. Hence the two columns can now be compared, to check for the accuracy of the system. The website is for use for two different types of users. The administrator can change, update and insert details in the database. The administrator has to continuously keep track of different films and different film crews. They have to keep making changes to the database with the release of each film, so that our prediction system gives better and accurate results. The names of different film crew is provided in a form to the user. The user fills the form with the details of a movie that hasn't been released yet. After the form is filled out with correct data about the film, to be analyzed the result is calculated according to the decision tree algorithm. The result is displayed on another page. Also in our project we have included a few graphs which act as visual aid to the user visiting the website. Through such graphs the person can get a better understanding of the statistics and logic implemented behind the prediction system. It will also help them remember the recent trends in the film industry.

6. Conclusion and Future Work

From the Experiment results, we found that it is difficult to apply data mining techniques to the data in the IMDb dataset. It requires proper cleaning and integration, and this consumed a large proportion of the time available for this analysis. In addition, much of the data is in textual rather than numerical format, making mining more difficult. The source data could not be integrated easily. By using natural language processing techniques the data can be integrated properly. For overcoming these problems, we performed some useful data mining technique on the IMDb data, and uncovered information that cannot be seen by browsing the regular web front-end to the database. More importantly, we believe that our research shows promise for further development in this area. Other interesting patterns can be identified by same technique, if additional dataset is available. A more accurate classifier is also well within the realm of possibility, and could even lead to an intelligent system capable of making suggestions for a movie in pre-production, such as a change to a particular director or actor which would be likely to increase the rating of the resulting film. Factors related to word-of-mouth (WOM), such as blogposts, play a key role in predicting the success of a movie within previous studies on predicting box office. WOM-based attributes might therefore also be useful as predictors within the domain of movie prediction in order to further improve prediction performance on such datasets. Further on it might also be relevant to Evaluate whether or not the prediction performance and results are generalizable over multiple datasets as well as over larger and smaller datasets in order to be able to draw moregeneral conclusions, as the current experiment setup only incorporated a single though well-established dataset. Likewise, it might also be of interest to evaluate whether or not the same is applicable to box office predictions, to further widen the scope of the study. Another possibility would naturally also be to include a wider range of algorithms and algorithm configurations, as the methods included in the current study were narrowed down to fit the given time frame.

References

1. Atta Badii, Ivo Keller, Mathieu Einig, Tobias Senst, Thomas Sikora, Volker Eiselein 2013 '*Prediction of movies box office performance using social media'*.
2. A. Kuhn, T. Senst, I. Keller, T. Sikora, and H. Theisel 2012 Comparison of Four Text Classifiers on Movie reviews pp. 387-392.
3. B.Abidi,N.Aragam,Y.Yi,andM.Abidi 2008 Feature Level Sentiment Analysis on Movie Reviews *ACM Computing Surveys*, **41**, 1 pp. 1–36.
4. B. Solmaz, B. E. Moore, and M. Shah 2012 CART approaches to mining incomplete data *IEEE Transaction on Pattern Analysis and Machine Intelligence* **34** 10 pp 2064-2070
5. Chandan Singh and Dipreet K Reddy 2014 Sentiment analysis of Indian movie review with various feature selection techniques *Survey Paper,Journal of Big Data.*

6. G. Diamantopoulos and M. Spann 2005 Performance analysis of CART and C5.0 using sampling techniques *Advances in Computer Applications*. **11** 3, pp. 233–243.
7. Honghai Liu, Shengyong Chen,and Naoyuki Kubota 2013 Data cleaning for data mining *ASurvey, in IEEE Computing for Sustainable Global Developmenton*.**9**,3, pp. 1222.
8. M. Patzold, R. Heras Evangelio, and T. Sikora 2010 Data cleaning: An abstraction based approach *in International Conference on Advanced Computation, Communication and Informatics* pp. 157-164.
9. R. Heras Evangelio and T. Sikora 2011 Recommender System Framework using Clustering and Collaborative Filtering *International Conference on Emerging Trends in Engineering and Technology* pp. 71-76.
10. Worapan Kusakunniran, Hongdong Li, and Jian Zhang 2009 A direct method to perform k – means clustering in R *DICTA*, pp. 250-255.
11. W.Wu, X.L.Chen,andJ.Yang 2005 Detection of association rules and patterns from datasets *IEEE Trans. Intelligent Computing Systems*, **6**, 4, pp. 378–390