Volume 1 No. 2, August 2011
International Journal of Science and Technology

http://www.ejournalofsciences.org

A Fine Grained Technique for Viral Marketing based on Social Network: A Machine Learning Approach

[1] Debajyoti Karmaker, [2] Hafizur Rahman, [3] Mohammad Saiedur Rahaman, [4] Md. Kamrul Bari
[1,2,3] Department of Computer Science. American International University-Bangladesh (AIUB).
Kemal Ataturk Avenue, Dhaka, Bangladesh.
[4] Department of Finance, American International University-Bangladesh (AIUB).
Kemal Ataturk Avenue, Dhaka, Bangladesh

ABSTRACT

Social networking sites are the platform to share familiar interest, and it is free of charge. But the e-commerce made it possible for the Social applications (specially the games) a potential business sector for different game development firms. Collectively the main focus of Facebook applications is to connect a user with a network and to do so the games are initially free of cost. Although the developing firms can make profit with it but the main earnings are from virtual currency (e-commerce) by providing extra benefits with the game. The problem is to identify and segmenting the potential buyer for a game and how to do marketing. This research presents a machine learning approach to generate some efficient rule based decision for viral marketing.

Keywords: *Decision tree, Decision rule, Post Pruning, Viral marketing, Machine learning, Entropy.*

1. INTRODUCTION

Facebook [16] is a social networking website intended to connect friends, friends of friends, family, and business associates. Facebook has over 500 million active users. Around 200 million of them log on to Facebook in any given day. An average of over 71 million pieces of content, which include links, images, videos, notes etc , are shared among them everyday [18]. Facebook allows a user to make new connections who share a common interest, expanding his/her personal network [13] [14]. Facebook is also popular for other social activities like versatile social applications like games (e.g. Farmville) where users have the freedom to play with their friends. it has become the social platform for anyone looking to promote just about anything online.

The main purpose of this paper is to design a classifier for segmenting and targeting the potential buyers for a Facebook application (game) with the help of previously analyzed data (same category game) to minimize the cost and time for marketing. In data mining Association Rule Mining (ARM) is one of the techniques used to extract hidden knowledge from datasets[1][2][3][4], which can be put to good use for business profit. This classifier identifies the best ways to promote a particular game on Facebook depending on 5 attributes (country, game type, age, sex, credit card type). The experimental results show that the proposed classifier achieves up to 78% accuracy.

As Facebook pays money for cost per click (CPC), and also for impression (CPM), so without direct purchasing of virtual goods, one can yet do a successful business through Facebook games. Yet the hot spot for profit remains the virtual items.

The organization of the paper includes the proposed scheme and the attribute selection criterion in chapter 2. Chapter 3 discusses on data collection process and limitations. Chapter 4 shows some analytical results using popular open source data mining tool WEKA [5][6] to examine the efficiency of our proposed model and presents some suggestions. Finally the paper concludes in chapter 5.

2. PROPOSED SCHEME AND THE ATTRIBUTE SELECTION CRITERION

For our proposed Decision tree initially we select 12 attributes and they are discussed bellow.

Country:

For this attribute, initially we did clustering. Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters [1][2][3].

The clustering is been done based on the attributes Number of Facebook users 1st July 2008, Number of Facebook users 1st July 2009, Number of Facebook users 1st July 2010,12 month growth percentage,24 month growth percentage and paid game users percentage[19][20]. So for clustering three key terms are kept in consideration for clustering and three key points are:

- Total no of users
- Growth rate per year
- Users percentage who purchase virtual goods

K-Means clustering is an exclusive clustering algorithm [7][8] which is used here to form the clusters. K-Means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. There are two issues in creating a K-Means clustering:

1) Determine the optimal number of clusters to create.
2) Determine the center of each cluster.

Considering the previously described three key points of country attributes we can divide the countries into three groups (Figure 2) to find out who amongst all the countries are the best targets, better targets and average targets.

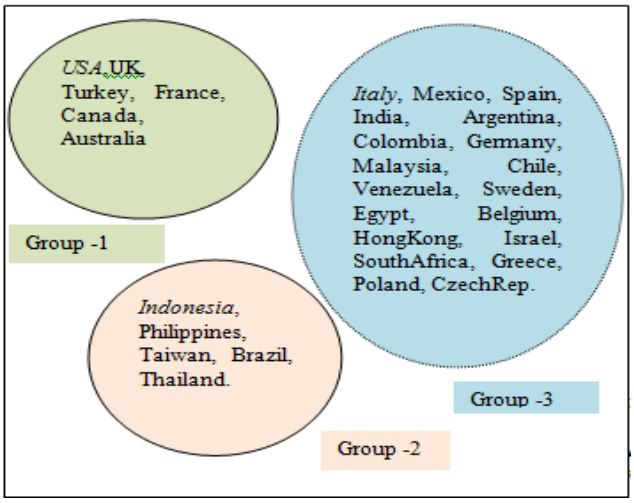

Figure 1: Country cluster

Best targets → *Group 1*
Better targets → *Group-2*
Average targets → *Group-3*

Where Group-1 represents the best group for target customers as they will to pay online for virtual currency, total no of users and the growth rate. Under all circumstances it would seem that we can only target these customers, but the fact is its very expensive to do promotional activities for the countries of Group-1 cluster and more importantly not all types of application is preferred by all countrymen. Three country groups represent three clusters so the value of k is set to three.

The algorithm works as follows:

I. Set the value of k =3.
II. Select 3 countries in an arbitrary fashion. Use these as the initial set of 3 centroids.
III. Assign each of the countries to the cluster for which it is nearest to the centroid.
IV. Recalculate the centroids of the k clusters.
V. Repeat steps III and IV until the centroids no longer move.

After running this algorithm on the dataset we come up with three country groups.

Group 1: *USA*, UK, Turkey, France, Canada, Australia.
Group2: *Indonesia*, Philippines, Taiwan, Brazil, Thailand.
Group3: *Italy*, Mexico, Spain, India, Argentina, Colombia, Germany, Malaysia, Chile, Venezuela, Sweden, Egypt, Belgium, Hong Kong, Israel, South Africa, Greece, Poland, Czech Rep.

Age Discretization

Discretization is used here to convert the continuous age data to form 6 groups and the groups in order to make it categorical so that we can apply these categorical values to our decision tree. The range of each group is given bellow

Age - 1	years	from	13	to	17
Age - 2	years	from	18	to	24
Age - 3	years	from	25	to	34
Age - 4	years	from	35	to	44
Age - 5	years	from	45	to	54
Age - 6	years	from	54	to	64

These groupings are done based on the respondent profile from an survey and users who want to use more application during the upcoming days [20].

Other Attributes

Other attributes are SEX, POWER UPS, WEARABLES, VIRTUAL GIFTS, which has only two possible values YES and NO. There is one more attribute, that is GAME CATEGORY which can have the values [ACTION, STATEGY, INDOOR, OUTDOOR]. Last one is the class that is the promoting way or promoting media for what we have designed the decision tree. This class contains the following values.

I. Facebook advertisement
II. Advertisement on other applications
III. Facebook live feed
IV. Invitation
V. Email
VI. Facebook notification

Facebook advertisement

Direct advertisement on Facebook is the most effective way for promotion. But cost is directly associated with it and it varies in different countries.

Advertisement on other applications

Banner advertisements through other sites or other applications may be a good platform for cross promotion, generally done through iframe.

Facebook live feed

Game progress information published on different user's wall as well as application user's wall is another type of promotional activity.

Invitation

In this regard invitation sent through game that user can receive on Invitation tab on their profile.

Email

Here emails are sent directly or by a specific application.

Facebook notification

Automatic and application oriented notifications on Social network are also used for promotional purpose.

The J48 Decision Tree

J48 is a version of an earlier algorithm developed by J. Ross Quinlan, the very popular C4.5[9][10][11]. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data[21][3][5].

Decision trees can represent data with diverse types. The simplest and most well-known is numerical data. discrete set of symbols might be used to represent those numerical data in categorical form. For example, a student with marks above 60% might be regarded as "First division", if marks is above 45% and bellow 60% can be called as "Second division" again if marks is above 33% and bellow 45% can be called as "Third division", else the mark can be labeled as "Fail" These values have no relationships or distance measures.

Decision tree algorithms function recursively. First, the root node must be selected. The root node must successfully split the data in order to form a efficient tree. Each split attempts to trim a set of instances until they all have the same classification. The best split is that one where the information gain is maximum.

Information comes from the concept Entropy. Often used to compute the amount of information for example, a database column of values

Entropy formula

$$H(P) = -\Sigma i{:}n\ p(si) * \log(p(si))$$

The entropy formula takes as input the probability distribution (denoted with a capital P) as the basis for computing an entropy score.

The J48 algorithm has several options for pruning. Here subtree replacement pruning is been used to overcome the problem of over fitting, i.e. the algorithm recursively runs until maximum accuracy is achieved. the tree may be generated with too much braches that leaves us with too many leafs. This may perform better on the tanning data set but might not perform well on a new data set as it may not be effective. Pruning always reduces the accuracy of a model on training data but will improve the performance. The overall idea is to gradually simplify a decision tree until it gains an equilibrium point for performance and accuracy.

Here the tree view of our decision tree is represented as a tree structure. The Country Group attribute is selected as the root element of the tree. Running this with 10 cross validation on the tanning set gives us a satisfactory result. The performance evolution and Confusion matrix (Table 1) is given bellow.

Table 1: Performance measures of the trained classifier

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.004	0.984	1	0.992	1	Live feed
1	0	1	1	1	1	Invitation
1	0.009	0.971	1	0.986	1	Advertise
1	0	1	1	1	1	Other game
0.957	0	1	0.957	0.978	1	Notification
0.944	0	1	0.944	0.971	1	Email
Weighted Avg	0.99	0.003	0.99	0.99	0.99	1

Figure 2 shows the total decision tree generated by using j48 algorithm which uses subtree replacement pruning. All the attributes are not present in the diagram as pruning is applied.

J48 pruned tree View

```
Group = gOne
|  age = one: othergame
|  age = two: notification
|  age = three: advertise
|  age = four: invitation
|  age = five: advertise
|  age = six: invitation
Group = gTwo
|  age = one
|  |  gametype = strategy: notification
|  |  gametype = indoor: advertise
|  |  gametype = outdoor: notification
|  |  gametype = fighting: invitation
|  age = two: advertise
|  age = three: advertise
|  age = four: livefeed
|  age = five: advertise
|  age = six: advertise
Group = gThree
|  VirtualGifts = NO
|  |  Weapons = NO: othergame
|  |  Weapons = YES: livefeed
|  VirtualGifts = YES: invitation
```

Figure 2: Country wise user's information

After creating the decision tree we took 300 instances as test set to test the performance of the tree and the findings are as follows

Evaluation

Correctly Classified Instances	297 (99%)
Incorrectly Classified Instances	3 (1%)
Kappa statistic	0.9877
Mean absolute error	0.0033
Root mean squared error	0.0408
Relative absolute error	1.2239 %
Root relative squared error	11.0653 %
Total Number of Instances	300

As initially the games are free, and users of different demographic (country, sex, age) group can join the game, it is difficult for the developers to identify the potential buyers for a game. But by analyzing a game(s) data, a game development firm can easily identify the potential customers for same category game.

3. DATA COLLECTION AND LIMITATIONS

As these types of analysis are required application user's personal data like age, sex, accounts information, it is difficult (sometimes not possible) to collect those information from Facebook because of Facebook's new privacy policy.

3.1 Methodology

For this analysis we have collected 5000 data from ibtGames [15] for their popular game T20 (Twenty Twenty cricket). Among the information we can successfully collect around 2300 user's gender and country information, its Facebook's policy that the application provider cannot store any user's information except user ID. Unfortunately for the change in privacy policy of Facebook we cannot collect the date of birth and other personal information like occupation. From the game installation information of ibtGames (BrainWaveU [17]) we have collected another 10000 data regarding users click on the game (T20) and how they know about the game for the first time. We have sorted the user's information country wise. The figure 4 describes the summery.

As from Figure 3 we found that the T20 user is max in Indonesia so we further analyzed the information for Indonesian users only.

3.2 Analysis

After country wise segmentation we got around 1100 T20 Indonesian user's information for further analysis. We can successfully collect user's name, city and gender information. First, we analyze the user's information city wise. The figure 4 describes the summery.

Second, we analyze the user's (T20 Indonesian user) information gender wise. The figure 5 describes the summery.

Facebook has commonly six type of application notification. These notifications are also used in application marketing purposes.

Third, we have analyzed the T20 user's information regarding how they got the information about the game for the first time and install the game using brainWaveU's [17] information. The figure 6 and 7 describes the summery.

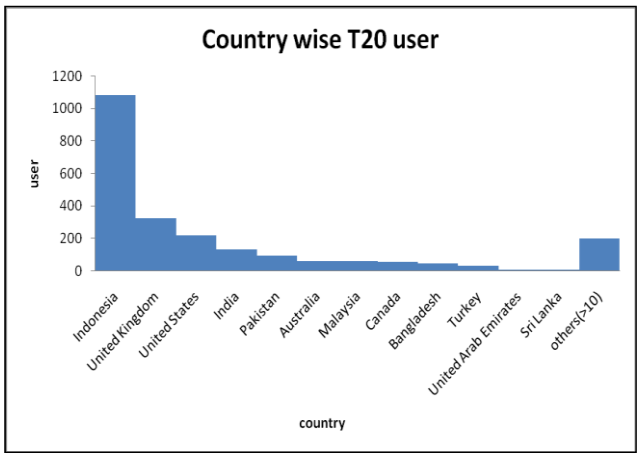

Figure 3: Country wise user's information

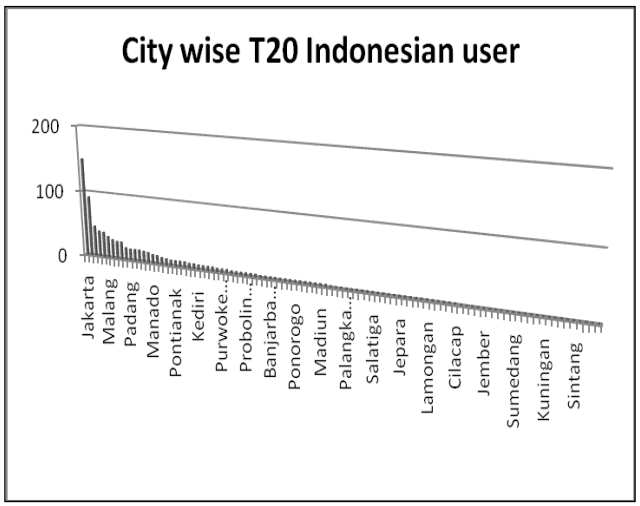

Figure 4: City wise T20 Indonesian user

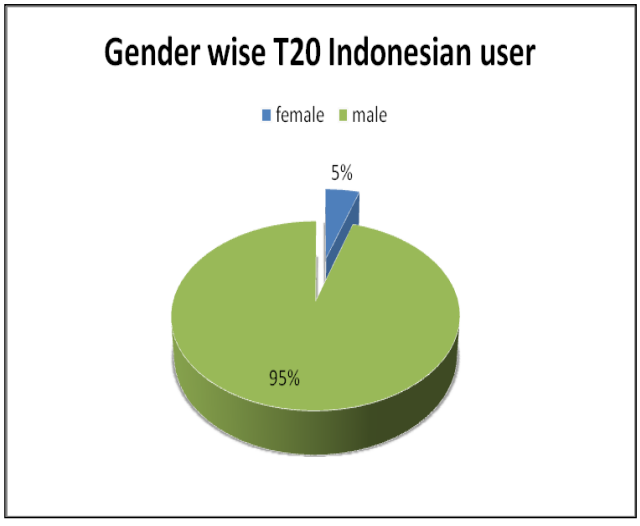

Figure 5: Gender wise T20 Indonesian user

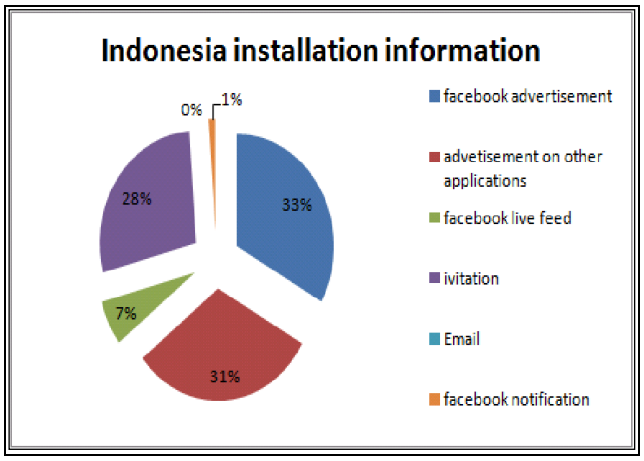

Figure 6: Indonesia installation information

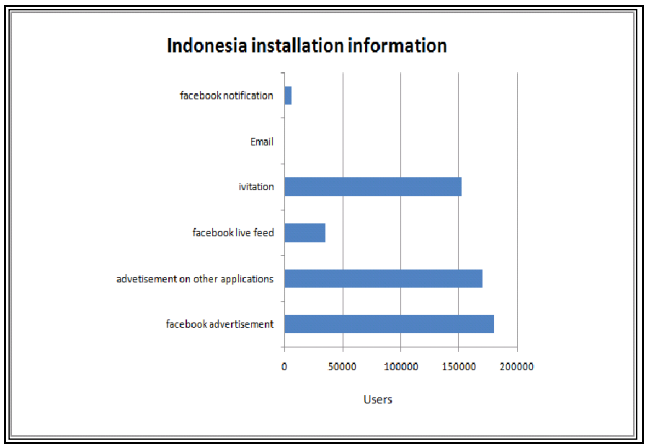

Figure 7: Indonesia installation information

4. RESULTS AND DISCUSSION

Findings

We found that the cities like Jakarta, Bandung, Surabaya, Semarang, Medan, Malang, Bekasi, Yogyakarta and Tangerang of Indonesia has the most number of users for T20 game application from figure 3. We also found form figure 4 that 95% of the users are male.

From the game installation information on figure 6 and 7, we found that the Facebook advertisement, Advertisement on other applications and Invitation has the most significance for Indonesian users. Email and Facebook notification has the least significance for them.

Country like Indonesia is under the country group 2 as shown in figure 1 and according to the figure 2 users of this game fall under the path Group (Group 2) → Age (Age 1) → Game Type (Outdoor) → Notification, which proves the correctness of our decision tree.

Also to notice that Group 1 and 2 users have different choices in correspondence with their ages. Group 2 → Age (Age 1) has the variant factor game type. Group

3 users does not differentiate with age factor, they are more influenced by those people who are around them.

Suggestions

From the findings we can target the users of Indonesia regarding their city and gender. As Facebook advertisement, Advertisement on other applications and Invitation has the most important way for Indonesian users to know about this game, the game developer can use these specific sectors to market T20 for Indonesian users.

When any game like T20 is installed for the first time by any user, the application providers can store the vital information to analyze the game market. But as per Facebook's privacy act, application providers cannot store any user's personal data on their database. If this can be possible, application providers can use Data mining techniques to predict the market before launching a new game on Facebook. They can also categorized the market based on different game types like action, adventure, sports etc and take different marketing strategy based on the findings on different perspectives.

5. CONCLUSIONS

Facebook applications are gaining popularity rapidly, specially the game applications. Different Companies earn massive amount of profit by developing Facebook applications and application add-ons. Companies market their games in a broad-spectrum, sometimes they customized their marketing strategy based on some assumed constraints. Facebook allows different advertisement practices based of different demographic values [12], but the other marketing tactics like customize notifications are developer dependent. Application developers can choose different tactics, sometimes customized game solutions, for different demographic constrains to market their games. If this kind of analysis is possible for the developers before launching any new game, they can easily predict the targeted demographic constrains and can develop unique and custom solutions to market their game applications.

REFERENCES

[1] Max bramer , "Principles of Data Mining" Undergraduate Topics in Computer Science ISSN 1863-7310.

[2] Han J, Kamber M, "Data Mining: Concepts and Techniques", Second edition, 2006, Morgan Kaufmann.

[3] Kantardzic, Mehmed, "Data Mining: Concepts, Models, Methods, and Algorithms", 1st edition, 2002, Wiley-IEEE Press.

[4] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, 1996.

[5] G. Holmes, A. Donkin and I.H. Witten, "Weka: A machine learning workbench", Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, 1994.

[6] Witten IH, Frank E, "Data Mining: Practical Machine Learning Tools and Techniques", Second edition, 2005, Morgan Kaufmann.

[7] MacQueen, J. B.. "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. MR0214227. Zbl 0214.46201. 1967.

[8] Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. (2009). "NP-hardness of Euclidean sum-of-squares clustering". Machine Learning 75: 245–249. doi:10.1007/s10994-009-5103-0.

[9] Quinlan, J. R,:C4.5: Programs for Machine Learning".1 edition, 1992, Morgan Kaufmann.

[10] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Proceeding of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007.

[11] J. R. Quinlan, "Improved use of continuous attributes in c4.5". Journal of Artificial Intelligence Research, volume 4, 1996.

[12] http://www.facebook.com/advertising,Facebook advertising, last visited at 8th May, 2010.

[13] http://hallman.nccommunities.org/glossary, what is Facebook, last visited at 8th May, 2010.

[14] http://www.wisegeek.com/what-is-facebook.htm, what is Facebook, last visited at 8th May, 2010.

[15] http://www.ibtgames.com, Infra Blue Technology Games, last visited at 8th May, 2010.

[16] http://www.facebook.com, face book advertising, last visited at 8th May, 2010.

[17] http://www.brainwaveu.com, game installation and accounts database of ibtGames, last visited at 8th May, 2010.

[18] http://sem-group.net/search-engine-optimization-blog/social-media/viral-marketing-on-facebook-7-points-you-just-cannot-neglect, last visited at 18th March, 2011.

[19] http://www.nickburcher.com/2010/07/facebook-usage-statistics-by-country.html, Facebook usages, last visited at 9th March, 2010.

[20] http://www.slideshare.net/michaelkim1/virtual-goods-market-courtesy-playspan, playspan july 2010, last visited at 2nd May, 2011.

[21] http://gautam.lis.illinois.edu/monkmiddleware/public/analytics/decisiontree.html, Decision Tree Induction, last visited at 3rd March, 2011.