Implementing User Clustering using Web Log Data

Saranya.S.V.[1] Sudhashree.A.[2] Dr. J. Frank Vijay[3]
[1, 2] Student [3]H.O.D.
[1,2,3]Department of Information Technology
[1,2,3] KCG College of Technology

Abstract—The sophistication of the web browsing depends on the likelihood of user reaching the intended web page at the earliest. Patterns of usage have to be tracked in order to conclude the user's intent. These usage patterns are analyzed with respect to URL's identified from user segmentation which is obtained from finding similar behavioral patterns. The technique employed here to achieve this phenomenon is the formation of user communities using pattern mining technique. Agglomerative clustering methods are used to typically discover the strongly linked communities. But these methods prove to be inefficient when implementing in large networks with several thousand nodes. Hence the method of triad formation, with graph mining can be applied to resolve new combination of edges which are not connected and establish links between them thereby increasing the identification of active users. The methodology is to form links between URL's that are most frequently used. The prerequisite for this is to detect communities based on the user's interests and increasing connectivity. The ultimate aim is to help the network providers serve their customers based on the user interest. The genetic programming method of churn prediction is applied to identify the possible customers who may leave the network with the help of which steps to retain the existing customers in a network can be made.

Key words: user segmentation, agglomerative clustering, graph mining, triad formation, churn prediction, genetic programming.

I. INTRODUCTION

Study of the behavior pattern of users in a network proves to be useful in various aspects. The telecom network providers benefit the most by recognizing the preference of the user and thereby serve them accordingly. Initially, users are segmented into user communities based on usage patterns. Web log data is obtained from web server log files that contain information about the request to the server, its response, the date and time of request and other such log details of client and server. Using a process mining technique which involves identifying user behavior with click patterns, the activities of a particular user can be found out. Based on these user activities, the user can be divided into segments called as communities. Communities are formed based on similar usage patterns and the formation of communities clearly helps us to track the various behavior exhibited by different classes of users. The behavior of a user is described in terms of navigation through various web pages, the frequency of usage etc.

II. LITERATURE ANALYSIS

From [1], the basic idea behind identifying user communities based on process mining techniques is derived and with finding the similar usage patterns the classification of the users in communities is made. Classifying users into communities help in the understanding of the behavior of users of different types and it also is a boon for the telecom companies to serve the target customers according to their needs.

In the paper [2], the various algorithms and techniques with which user communities can be formed are described. A conclusion is made after comparison that the CNM algorithm is the best existing algorithm to divide users into communities inside a network.

From [3], we get a graph theory based approach to further divide the communities into smaller sizes and closely-knit sub-units. With this graph based algorithms a better and detailed understanding of the working of the user communities in a network can be derived which in turn leads to better understanding of the customer behavior within a given telecom circle.

Triad formation [4], gives a whole new approach to improve the connectivity in a given network. Isolated edges and identified and generates new edges between nodes that are not connected but seem to have a common node between them. This is also a more dynamic approach as newly added nodes are identified and triads is formed even with the newly added nodes thereby expanding the network and increasing the possibility of potential edges.

The method of churn prediction [5] gives the telecom service providers and opportunity to identify the customers who might leave the network and steps to retain the existing customers can me made. In the current scenario, with lots of competition between many service providers it is hard to retain the customers and make them stick to a particular network and hence this method proves useful to maintain a network and retain its customers.

III. ALGORITHM ANALYSIS

The identification of formation of communities can be done using agglomerative clustering methods. They provide a way of linking closely linked communities. CNM algorithm is one of the best hierarchical agglomerative algorithm's which iteratively merges nodes into communities based on the gain in modularity.

In CNM algorithm first each node is taken as one community and adjacency matrix is drawn for that particular network. If A_{vw} is the adjacency matrix and $\delta\ (c_v, c_w)$ indicates the connectivity i.e., δ is equal to 1 if the vertices are connected, else 0,

$$m= \frac{1}{2}\sum_{vw} A_{vw}(c_v, c_w) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where, m is the number of edges in the graph and vertex=x belonging to the communityc_v.The fraction of edges making one community is,

$\frac{\sum_{v,w} A_{vw}\delta(c_v,c_w)}{\sum_{v,w} A_{vw}}$=m= $\frac{1}{2}\sum_{vw} A_{vw}(c_v, c_w)$…. (2)

This value is directly proportional to the probability of two edges merging in one community. But the value should be taken care of since, if equal to 1 this implies that all nodes belong to the same community. The main parameter which is taken care while splitting is the modularity. If the degree k_v of a vertex v is defined to be the number of edges incident upon it, the modularity (Q) is defined by,

Q= $\frac{1}{2m}\sum_{vw}[A_{vw} - \frac{k_v k_w}{2m}](c_v, c_w)$…….... (3)

CNM algorithm works only for undirected edges. The following equation shows it,

$(c_v, c_w) = \sum_i \delta(c_v, i)\delta(c_w, i)$ ……. (4)

CNM does not identify directed edges and so the algorithm goes into an infinite loop. The following modification was introduced:

If there exists an edge as A to B and B to A then swap B and A in the second case or else there would be an infinite loop and algorithm stops working. Once the nodes are swapped there will be only two copies of edges between A and B which resulted in direction of loop which solves the problem. But this method becomes highly inoperative and slow when applied to large networks with several nodes therefore a new method of graph mining is proposed which overcomes the disadvantages of the CNM algorithm.

A. ISOLATED COMMUNITIES

There may exist communities that may not be connected by anyways to other communities. They are called as isolated communities. Once communities in a network are identified the degree for each community is calculated. The degree is nothing but the count of the number of edges between one community to another. If the value of this degree is 0, it is called as isolated community. Members of an isolated community do not have any interactions outside their community.

B. FORMATION OF NEW EDGES

Graph G= (V, E) with V>0 and E>0 ….. (5)

1) Look for two nodes x, y that are non-adjacent and share a common neighbor.
2) If no such pair of nodes x, y exists, go to step 8, else go to step 3.
3) The edge weight between x and the common neighbour are labelled as w1.
4) The edge weight between y and the common neighbor is w2.
5) Check if w1 and w2 are more than threshold weight set by the user
6) If not got to step 8, else go to step 7.
7) Add edge (x, y) and assign a weight equivalent to average of w1 and w2.
8) Add element x, y to set E.
9) Go to step 1.

Consider the three nodes A, B and C as shown in figure 1. Here AC is connected and nodes B and C are connected. AC is given the weight w1 and BC is given the weight w2.

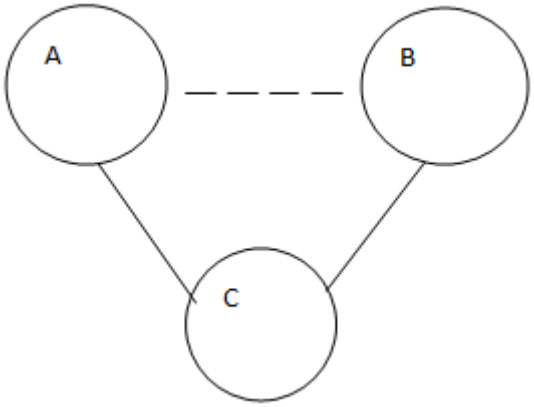

Fig. 1: formation of edges.

Now, with the threshold weight specified by the user w1 and w2 are compared and if both are higher than the threshold, the new edge AB is formed. The weight of the new edge is the average of w1 and w2. The formation of new edges results in better connectivity between the nodes of the community, thereby increasing the clustering and closeness between the nodes inside a community and also between them.

IV. CHURN PREDICTION

In the present telecom scenario the possibility that a customer leaves the network is very high. It is found that the cost of finding a new customer is ten times more than retaining an existing customer. A customer may want to leave the network due to various factors such as increased cost or reduced efficiency. With this data the customers who are likely to leave the network can be identified and steps to retain them can be implemented. In order to manage customer churn in a network it is important to build an effective and accurate customer churn prediction model. Some techniques used are:

1) Classification and regression trees (CART)
2) Logistic regression models (LRM)
3) Artificial neural networks (ANN)
4) K-means Cluster Algorithm

The most effective model compared to all the others mentioned above is the Genetic Programming (GP) model.

V. GENETIC PROGRAMMING

Identifying the exact reason as to why a customer would leave a particular network is practically not feasible. But the general reasons appear to be price, connectivity, coverage, satisfaction and so on. This process of genetic programming closely resembles the evolution in biology. The population in a network is evaluated using some fitness function and the candidates who are most likely to stay in the network, i.e., those with a best fitness score are moved to the next generation. In this way the GP technique is applied for every

generation as shown in figure 2 and the churn is each generation is found out. In a mobile telecom network the attributes with which the customers are evaluated can be as shown in table 1.

Field	Description
Age	Age of the customer.
Paid amount	Billing amount paid in a month
Net Local Calls	No. of local calls
NWD Call	No. of Nationwide calls
Local Mobile calls	No. of local mobile calls

Table. 1: customer evaluation attributes

The following diagram explains the GP approach:

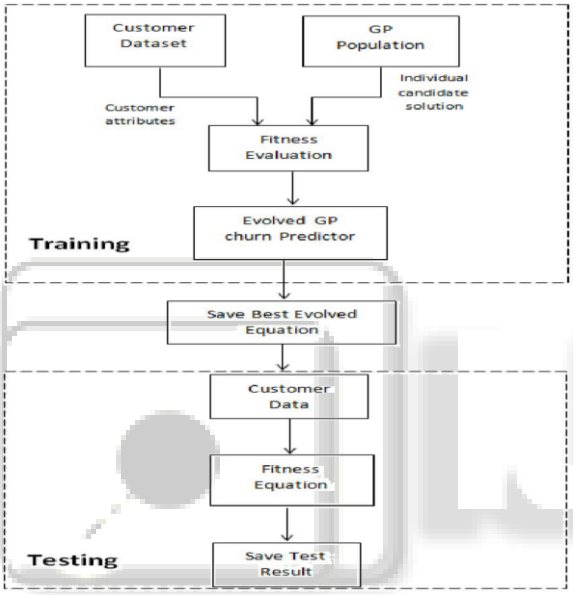

Fig. 2: the GP process

A. ATTRIBUTES USED IN GP SIMULATION

To develop initial population of customer churn prediction the following attributes are used in GP simulation: Age, Paid amount, net local calls, NWD Call, Local Mobile, NWD Mobile, and OS Calls etc.

B. FITNESS EVALUATION

In a generation, every individual candidate solution is evaluated and scored using the following algorithm based upon fitness of i i^{th} individual in GP population.

For i=1 to Ncust

{

If {fitnesscandidate> 0}

ChurnResult= 1;

Else

ChurnResult =0;

If {ChurnResult (XOR) ChurnID=1}

Fitnessi=1;

Else

Fitnessi=0;

Fitness=Fitness+Fitnessi;

}

Fitness=Fitness/ Ncust;

Where, Ncust is the total number of customers. Fitnesscandidate is the evaluation of different customer features on the candidate expression. ChurnID is the actual churn status of a record in the training data. Fitness is the fitness of ith record in the training data. Fitness value corresponding to an individual candidate solution is the mean value computed against the total number of customers in the training data set, and depicts the performance of an individual candidate solution. Likewise, fitness value of every individual in a population is calculated and the process is repeated generation by generation through genetic evolution.

C. COMPARISON WITH OTHER MODELS

Figure 3 demonstrates the performance comparison of the proposed technique with other state of the art techniques. From both training and testing it is cleared from the figure that the result of SVM is much better than K-mean cluster. On similar data, performance of GP based approach is much better than that of other two predictive models.

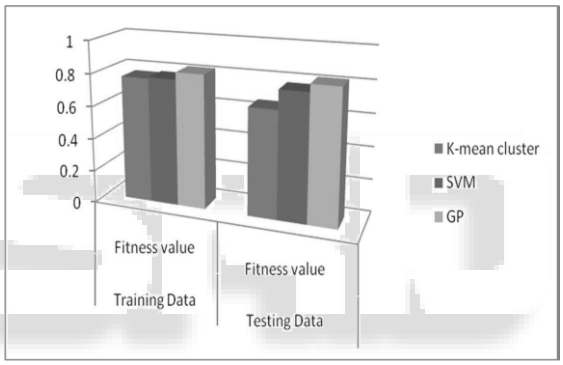

Fig. 3: comparison of GP fitness values with fitness values of other techniques.

1) Probability of error percentages for different predictive Models

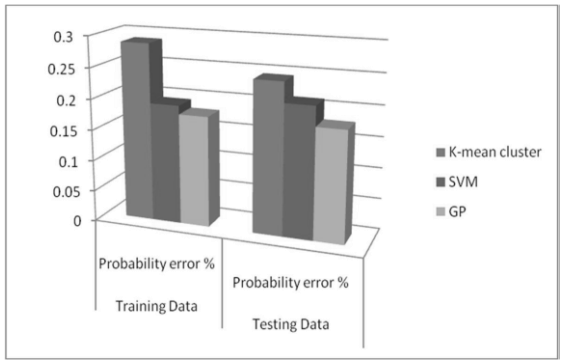

Fig. 4: error percentage comparisons for different models

Figure 4 shows the performance of the proposed technique with other techniques in terms of probability of error. Again we can see the induced error in the proposed technique is much less than SVM and K-mean cluster based approaches.

VI. RESULT

By the proposed method of identifying communities and triad formation the telecom network providers can understand the user preferences and can thereby serve their customers accordingly.

VII. CONCLUSION

The significance of triads also helps in identification of most frequently visited URL's. Generation of new edges will prove to be an advantage for end users in terms of functionality of customized search and also will provide service providers the aid to identify target audience for advertisements. With the means of identifying possible churn in a network it gives the network providers a possibility to satisfy the customer needs and retain the existing customers in the network thereby not losing their customer circle.

REFERENCES

[1] Katerina Slaninova, RadimDol'ak, Martin Miskus, "User segmentation based on finding communities with similar behaviour on the web site", 2010 IEEE/WIC/ACM International Conference on Web intelligent Agent Technology.
[2] M.E.J.Newman and M.Girvan, "finding and evaluating community structure in networks", phys. Rev. E 69,026113, 2004. doi:10.1103/PhysRevE.69.026113.
[3] M.Saravanan, G.Prasad, KarishmaSurana and D.Suganthi, "Labelling communities use Structural Properties", 2010 International conference on Advances in Social Network Analysis and Mining.
[4] N Nren Krishna, M Saravanan, "formation of triads in mobile telecom networks", DBKDA 2011: The Third International Conference on Advances in Databases, Knowledge and Data Applications.
[5] Imran Khan, Imran Usman, Tariq Usman, Ghani Ur Rehman, and Ateeq Ur Rehman, "Intelligent Churn prediction for Telecommunication Industry", International Journal of Innovation and Applied Studies.

