# Clustering Internet Usage Behaviours with SOM Neural Networks

U. Celenk, O. Ucan

*Abstract* — **According to different needs of users, there are different consumption habits. Consumption habits of people, which have the same age group or the same professions, are similar. A type of internet usage habits of people in this way is one of these habits. In recent years, developments in technology, GSM, and especially with 4G mobile internet usage have found applications in many areas of daily life. Enter to internet, wherever users need to, creates freedom. Messaging, media, finance and many different needs can be met through this connection. Users' occupation, age, gender, location, usage patterns according to different characteristics such as income level and the relevant properties are similar to each other according to the amount of internet usage (in Mb Download) connected to internet and internet usage frequency and duration of exposure can be clustered. SOM type of study, personal internet usage by artificial neural networks (data of the CDR) process and their profession, age, gender, location is to cluster usage patterns according to the values.**

***Index Terms***— **GSM Internet usage, Internet usage behaviours ,SOM neural Networks, Cluster analysis, Continuous Queries**

## I. INTRODUCTION

For a GSM operator, which has an increasing number of mobile subscribers and mobile traffic volume, understanding the subscribers' mobile traffic by using effective resources is important. Thus, by offering instantaneous promotions based on the using habits of subscribers, encouraging the mobile communication and charging this process, developing the systems require a great academic study. Consequently, observing the subscribers' old data traffic, instantaneous promotions for each user and developing high performance algorithms that define the unit cost of data usage are creating the aim of this study.

Analyzing mobile internet traffic, taking into account the using habits and the profiles, having an intelligent decision system for promotion systems, which is identifying the subscribers, will make a great profit for GSM operators. Thanks to this, operators will not only have an opportunity of offering instantaneous promotions based on their customer needs, but also increase the performance of charging systems and will acquire a system that have a lower energy consumption with lower IT infrastructure costs.

Today, the rapid technology evolution in the telecommunication sector requires that the current promotion solutions should be more flexible and focused on customer. According to strategy researches, it is emphasized that using new generation communication infrastructures, increase of diversity and complexity of products and services require more flexible and faster algorithms in promotion systems. Particularly, the necessity for the production of the tariffs that respond to immediate needs of customers in market and the start to offering new data packets with voice services is one of the most important obstacles that the promotion solutions will encounter in GSM sector.

## II. INFRASTRUCTURE OF MOBILE COMMUNICATION

In mobile networks; voice and data calls, primarily starts accessing of the mobile devices to BTS (Base Transceiver Station). Second demand is transmitted to BSC (Base Station Controller) and then transferred to MSC (Mobile Switching Center); it is provided to reach to GSM network. If it is not a voice-call, it is provided to customer to get an IP address from GGSN (Gateway SGSN-Serving GPRS Support Node) unit in main network. After this, directing the customer to Radius (Remote Authentication Dial-in User Service), it is questioned that the subscriber has a right of connection or not. After having the necessary authorization; login in with the name of PDP (Packet Data Protocol) with a protocol that works as a counter like CSG (Content Service Gateway), the process of internet access is started. Reaching out to a user's account by the related HLR (Home Location Register) or VLR (Visited Location Register) counters, a fixed quota is blocked. Meanwhile a call registration (CDR-Call Detail Record) is created. While the subscriber is creating data traffic by using internet, the counter starts to discount the used amount from the blocked amount. When the blocked amount, which is on the counter, is finished, having connection with HLR affirming the end of the blocked data, it is charged again and CDR comes to end. If

Manuscript received June 15, 2012; revised July 30, 2012. This work was supported by Republic of Turkey, Ministry of Science,Industry and Technology. San-Tez project with number 00874.stz.2011-1

Ulas Celenk is with the Istanbul University Electric and Electronic Department ,Istanbul,Turkey(phone:+90 535 644 73 18;e-mail:ucelenk@innova.com.tr).

Osman Ucan is with Istanbul Aydin University Electric and Electronic Department(e-mail:uosman@aydin.edu.tr).

the customer keeps at using the internet, the described processes in the previous steps are repeated again, so the new quota is blocked. These steps that are repeated during the usage of internet by the GSM customers are summarized in Figure 1.

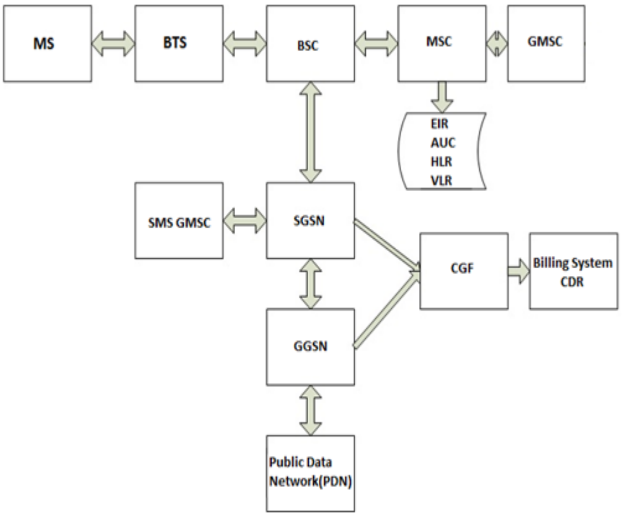

Fig. 1. Structure of Mobile Networks

During the invoicing internet access of customers for each created CDR, the charging process is done. In the related CDR data, there are too many individual information such as internet usage period, amount of the internet usage, date of internet usage and the locations and for statistical calculations CDR data can be used.

## III. CUSTOMERS DEPARTMENTALIZATION AND CONSTRUCTING CUSTOMER PROFILE

It is important to recognize your customer and to know what are their wants and needs to compete with other providers of mobile communications [1]. To implement this, segmenting customers and constructing customer profile are necessary. Constructing customer profile helps to extract the effective marketing strategy. This profile is based on the customer's attitude and subtracted from the calculation of the parameters. Constructing customer profile is a method of an external data of implementation to a prospective customer group. This method is based on current data, which can be used to investigate new customers and to recognize existing bad customers. The goal is to estimate a customer's attitude according to his/her data [2]. Constructing profile is done after the departmentalization of customers [3].

The departmentalization is a method of communication. This process defines attributes of the so-called partition or group of customers in data. It means to group community with similar qualifications and their distance. The departmentalization of customers is a preparatory step of classification of customers according to the division of every customer-defined customer groups. This is required to overcome today's dynamic consumer market. Using the departmentalization, the marketers become more effective in using the resources and catching opportunities.

- The quality and appropriateness of the information are required to create meaningful segments. If the company does not have sufficient customer data, the departmentalization of customer is not reliable or even useless.
- Too much data lead a complex and time-consuming analysis. Low-organized data (different formats, different welding systems, etc.) complicates to obtain considerable acquisition of information. Moreover, to complete effectively the results departmentalization can be very difficult for the company. Especially the use of a lot of variables in departmentalization can be amazing and departmentalization, which is not proper to management decisions, can occur. On the other hand, the effective variables cannot be detected. Most of these problems arise because of the missing customer details.
- Intuition: despite highly informative data, that's right for the analysis of data analysts to find the data required to produce consistently departmentalization hypothesis.
- Continuous follow-up: the departmentalization requires a continuous monitoring and updating to obtain a new customer data. In addition, the efficient departmentalization strategies have an effect on the affected customer attitudes; thus the classification of a new customer is to be revised. Moreover, the feedback, which is a direct e-commerce environment, should be updated almost every day in departmentalization.
- Further departmentalization: the department may be too small or insufficient to evaluate as a separate departmentalization. Editing departments can be achieved by a data reduction method depending on cluster algorithm category. This report will be discussed on several cluster algorithms compared to each other.

Constructing customer profile provides marketers to serve better existing customers and to keep them in the ground to communicate with them. This is made of a combination of personal and populous data of customers. The customer profile is used to identify new customers by using external sources taken from various sources such as census data. This information is used to find pre-established customer relation departments. This (the section on population and personal information belonging to the sum of each profile) enables a calculation. Thus, the estimated attitude can be helpful for each profile. Depending on the purpose, people choose which profile would be appropriate to the project. A simple customer profile file includes information such as age and gender. If a clear product is desired for profiles, the file contains information about the product and / or the amount spent.

## IV. CALL DETAIL RECORDS

Call detail records (CDR) in communication system are engaged automatically when a phone starts to use for monitoring and billing purposes. Information that is stored in search records can be edited to investigate the relationship between telephone users. In particular, with the approach mentioned in this report, the accumulated detail records can be identified mobile user communities effectively. The importance of understanding the attitudes of the search should be noted that communities and telephone companies are available to them. To manage appropriate communities in records, data transfer techniques and social network analysis should be used fully [4].

Subscribers connected to an account can be uploaded to a computer at a request time. If the provider is able to give a detailed bill to users, it can be seen in a long distance phone bills, for example records can be seen in every invoice.

Call detail records in telephone exchange include information about exchange during the whole of the calls. Call detail records is provided by Automatic Message Account (AMA) and is operated by the Operational Support System

Call Detail Record file can contain more than one type of call. For example, fixed-line voice and data hard can be put in the same file, but there is a separate evaluation purpose

Call account software or contact management software is usually used to recover and operate call detail record data.

A call detail record data consists of the total data defining telecommunication transactions such as:

- The subscriber's telephone number which calls (the caller)
- The telephone number which accepts the call (the receiver)
- Time when the call starts (date and time)
- Call duration
- The caller's calling fee
- The phone's identity who holds records
- Record sequence number
- Additional digits to be used to post the call
- Call results reporting the status of the receiver, for example, the number dialed is busy or the call is cut
- The direction of the beginning of the call
- The direction of end of the call
- Call type (voice calls, text messaging, etc.).
- Defective condition that may be encountered

## V. QUESTIONS ABOUT THE DATA FLOW IN CALL DETAIL RECORDS

Communication network management applications operate a fast, unpredictable and constantly flow of data including network performance metrics and package. Common DBMS, providing continuous online operation due to inadequacies of a sort of query, is open to an existing DBMS transaction management tool, the operation or preventing the use of closed questions simple, encrypted, continuous operation. Action management system allows network users to change establishment, relocation, and appropriate monitoring technique to support the management of the ISP's network. [5].

As a concrete example, consider an ISP that collects drafts of the package of the two link network (among others). The first one called as customer connection connects the customer's network of the ISP network. The second one called as basic connection connects two cutters in the ISP's network. In more simple terms, package connection includes five items listed in Figure 1. We use PTC and PTB to show drafts of the package collected from customers and basic connections.

TABLE I
RECORD STRUCTURE OF A PACKET HEADER

| Field Name | Description |
|---|---|
| Saddr | IP address of packet sender |
| Daddr | IP address of packet destination |
| Id | Identification number given by sender so that destination can uniquely identify each packet |
| Length | Length of packet |
| Time Stamp | Time when packet header was recorded |

Introduce the general architecture for processing continuous queries over data streams, illustrated in Fig. 2. For now let us consider a single continuous query Q with answer A operating over any number of incoming data streams. Multiple continuous queries can be handled within our architecture (as implied in the figure),. The query is over data streams only, although mixing streams and conventional relations poses no particular problems. When query Q is notified of a new tuple t in a relevant data stream, it can perform a number of actions, which are not mutually exclusive

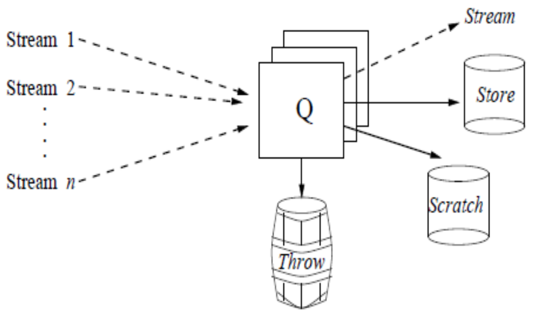

Fig. 2. Architecture for processing continuous queries over data streams.

- If a new tuple in A is known that it will be forever in A, then Q can send tuple A to a piece of the current shown in Fig. 2.
- If the new tuple A will not be in A even though it is decided to be in, then A is added to the part of the Storage shown in Fig. 2. In other words, Flow and Storage defines the answer A. The goal is to minimize the storage for the result, and then is to be ensure that they send tuples to the Stream instead of Storage when it is possible.
- The new current tuple t can cause the situation of deleted or updated of respond tuples. Respond tuples can be sent from Storage to Curent.
- T or the data taken from T should be saved so that in the future we are sure to calculate the outcome of the question. In this case, T (the data taken from T) is sent to Stratch in Fig. 2. Accordingly, we can send the data from Storage.

- In the case of transmission of T to the Waste part in Fig. 2, we have not needed T. It should not be forgotten that the Waste part has not storage. (If you're not interested in storing unnecessary data).
- As a result of the new current tuple t, we can receive data from a pre-recorded in Stratch (or in Storage) and instead of this, we can send to the Waste. If the goal is to minimize storage, we must be sure that unnecessary data is sent to the Waste instead of Stratch.

## VI DATA PREPARATION FOR GROUPING

The data come from the bill system defines attitudes of customers' expense and pay. Generally, mobile billing system data contains all kinds of service charges paid by the customer each month. Detailed call record data defines the usage attitude of the customer. They save data of each customer's call [6].

To prepare data for the grouping, particularly data on call details may take some time to prepare. This preparation is as follows [7]:

- Find and repair incompatible data formats, data codes, spelling, abbreviations and spelling errors.
- Delete unwanted data. It may include a lot of meaningless information in analysis such as data production switches or model numbers.
- Translate codes into text or significant numbers. It is needed to create new index in such cases the rate of the calling or receiving number, the rate of call duration and the rate of long distance calling, etc. Data may contain cryptic codes. These codes should be converted into a suitable text.
- Combine data by converting customer data from general to specific
- Find multi-use areas. Possible way of achieving this is to determine the field variables.

During this research is needed to prepare the following data:

- Check out abnormal, out of connection and uncertain values. Some of them may be true, but it can be almost impossible to explain.
- The damaged values replace the missing data values
- Add calculated areas as target or the input data
- Place constant data to regions
- Standardize the variables. There are two types of standardization. First, to standardize the values between [0.1]. The second type is to standardize the change to 1.
- Translate numeric data (example yes / no answers) into the metric measure.
- Translate text into number or numeric data.

New fields can be produced from the combination of frequency, average and minimum / maximum values. The aim of this approach is to create controllable variable number in correlation between variables. For this purpose, there is a connection between factor and the combined analysis used as techniques [8]. When there are large amounts of data, data reduction techniques (data accumulation, reducing the size and number, create concept hierarchy) are also useful to apply.

Reducing size is the selection of a suitable size among a group of features so that the diffuse of data result rate is as close as possible to all features of the original value distribution. To do this; additional works such as a large, random or experimental research, grouping, decision tree, or combination rules may require.

## VII. SOM NEURAL NETWORKS

Instead of classical statistical methods, artificial neural networks can be used in grouping studies. Artificial nerve Networks do not need the distribution assumptions for datum. Having a large quantity of elements and variables in a data set does not raise difficulties for neural Networks

The most used artificial neural network in grouping studies is SOM (Self-Organizing Maps) [9]. SOM Networks developed by Teuvo Kohonen in 1982. For this reason, they are also known as Kohonen SOM Networks. SOM networks can function at both K-Averaging and multiple dimensioned scaling methods which are in classical statistical. In other words it makes both grouping and mapping in the data label. Thus these Networks become very popular in recent years [10] .

SOM Networks are one layered Networks and they consist input and output neurons. Number of variables in the data set determines the number of input neurons. Each output neurons represents one group. A SOM network is seemed at Fig. 3. As aparting from other artificial neural networks, the positioning of neurons at the output layer is very important. This positioning can be linear, rectangular, hexagonal or cube shape. Mostly rectangular and hexagonal positioning are preferred. In practice, rectangular positioning is applied as quadratic. This positioning is important in terms of topologic neighbourhood. Reference vectors (code-book-vectors) show the connection between input neurons and each output neurons. It is conceivable to think these vectors as columns of coefficients matrixes. This topologic neighbourhood are used renewing the reference vectors while SOM neural Networks is trained.

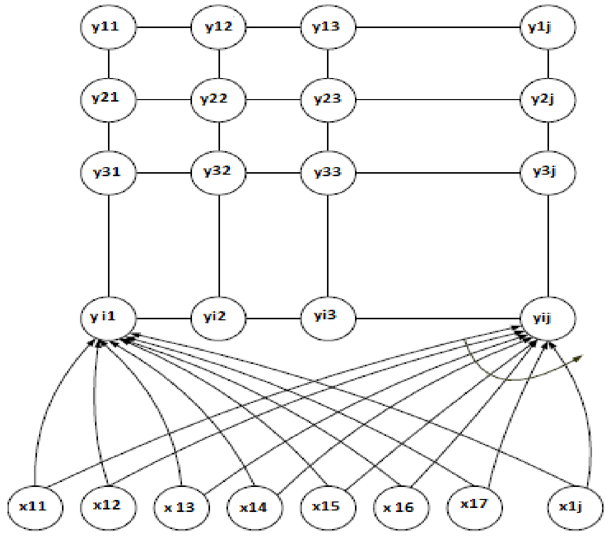

Fig. 3. Kohonen SOM Neural Network

## VIII. SOM LEARNING ALGORITHM

The used algorithm in Kohonen Networks is SOM (Self Organizing Maps) algorithm which also gives the name to these Networks. The learning algorithm used in these Networks is uncontrolled. As input vectors in the data set enter the network, the network arranges itself and form reference vectors. This algorithm is given at below [11].

The symbols that are used in this algorithm.

$w_{ij}$: the reference vector that belongs to output neuron which is at i. row j. Column

x: input vector

D (i,j) : square of Euclidean distance of x vector to the output neuron which at (i,j) coordinates.

i,j. : the coordinates of output neuron which is closest output neuron to the x vector.

α :learning coefficient

Algorithm

```
1).Assign initial value to $w_{ij}$ coefficients. Determine topologic neighbourhood parameters. Adjust learning coefficient parameters.
2).While finish condition is wrong , follow steps 3-9
3).For each x input vector follow steps 4-6
4).Calculate Euclidean distance D(i,j)= $\sum_{ij}\left(w_{ij}-x\right)^2$ for each i,j
5).Find value of i,j where D(i,j) is minimum
6).For all the output neurons in the defined neighbourhood of I,J $w_{ij}$ (new)= $w_{ij}$ (old)+ α (x-$w_{ij}$ (old))
7).Update learning coefficient
8).Decrease topologic neighbourhood parameter at specified times.
9).Control the finish condition
```

As it is understood from the algorithm above, firstly, an initial value is given to the reference vectors. Before starting to loop, a high value is assigned to learning coefficient (α) and neighbourhood variable (R). A value between 0 and 1 is assigned to α. It is preferred to have this value closer to 1. R variable starts with the value which is larger than height or width of positioning of output layer. One loop for algorithm is submission of all the rows in the data set to SOM network input. One of the row of data set is x vector. The square of Euclidean distance of x vector to each neuron at the output layer is being found. Each neuron at the output layer is represented by a reference vector ($w_{ij}$). Hence, this distance is the distance between x vector and wij. The smallest value is being found among the calculated distances. Whichever output neuron has this smallest distance is the Winner neuron. In other words, SOM Networks are ‘‘competitor’’ networks. The reference vectors of winner neuron and neighbourhood neurons are being calculated again. The linear neighbourhood of the winner neuron seemed in Fig. 4. and rectangular neighbourhood winner neuron seemed in Fig. 5. As it seems from these figures, there are more neighbourhood neurons around winner neuron in rectangular neighbourhood. $w_{ij}$(new)= $w_{ij}$(old)+ α(x-$w_{ij}$(old)) equation is used in this calculation. For this reason if a small values are given to references vectors as initial values, the value of α must be taken as closer to 1. In this way reference vectors have a chance to generate themselves. By this way, a loop will be completed when these process are completed for each row in the data set. Reference vectors keep changing as long as loops continue. A and R values decreases in particular periods of loop. There is no particular rule that determines that the values decreased in how many loops. There are different opinions for this subject. Mostly it is adequate to decrease it with a linear function. Looping finishes when the changing in the reference vectors finish.

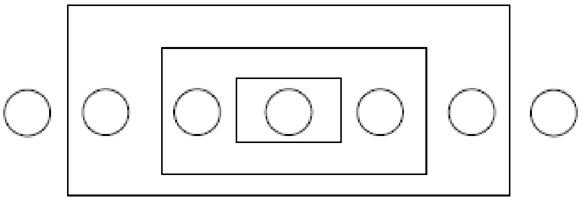

Fig. 4. Linear Neighbourhood of Winner Neuron (#) (in the order of inside to outside R=0, R=1, and R=2).

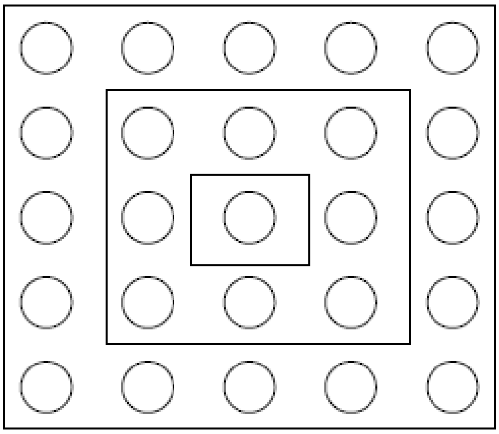

Fig. 5. Rectangular Neighbourhood of Winner Neuron (#) (in the order of inside to outside R=0, R=1, and R=2)

After completing the training of the network and forming the reference vectors, the elements that are in the data set are grouped together. All the rows in the data set are entered to network in a row. Entry vector is multiplied by reference vectors of output neurons. The element will be belonged to the group which the result is bigger. At the end of this process, elements are both grouped and placed in a two dimensional map. It is possible to see the elements that are close to and remote to each other from this map. If the positioning on the output layer 3 dimensional, the map will be 3 dimensional also. These maps can be colored and shadowed in different patterns according to features of the groups. Therefore, a more visual map can be gathered.

## IX.CONCLUSION

In this paper, it is aimed to group people with their internet usage by analyzing the SOM Clustering Neural Networks and the CDR records. Day by day competitive in telecommunications market is increasing. Operators should apply recognizable marketing strategy depending on customers' different attitudes to improve marketing results. The large data base in the form of call detail record can be deduced from the customer information using Self

Organized Maps (SOM). SOM is useful in not only for grouping, but also for reviving multi-dimensional data. Call detail records show the customer's attitude. The grouping analysis based on call detail records can provide more information than the grouping analysis of marketing management. Create a new index defining customer attitudes are very useful to identify to create the group of customers. This helps create groups that can be spotted in a better, marketing administrations can create more appropriate marketing strategies. It is possible to revive multi-dimensional data with SOM. This helps create groups that can be spotted in a better, marketing administrations can create more appropriate marketing strategies.

## REFERENCES

[1] McDonald, M. and Dunbar, I., Market segmentation. How to do it, how to profit from it. *Palgrave Publ.,* (1998).

[2] Verhoef, P., Spring, P., Hoekstra, J. and Lee, P., The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decis. Supp. Syst.,* vol. 34 (2002), pp. 471-481.

[3] S.M.H. Jansen , Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior *A Vodafone Case Study* , July 17, 2007

[4] Wei-Guang Teng , Ming-Chia Chou, Mining communities of acquainted mobile users on call detail records *SAC '07 Proceedings of the 2007 ACM symposium on Applied computing* Pages 957 - 958 New York, NY, USA

[5] Arasu, Arvind (2006) Continuous Queries over Data Streams. *PhD thesis, Stanford University*.

[6] Qining Lin, Mobile Customer Clustering Analysis Based on Call Detail Records , *Communications of the IIMA 2007* Volume 7 Issue 4 Pages 95-100

[7] Feldman, R. and Dagan, I., Knowledge discovery in textual databases (KDT). *In Proc. 1st Int. Conf. Knowledge Discovery and Data Mining,* (2005), pp. 112-117.

[8] Mattison, R., Data Warehousing and Data Mining for Telecommunications. *Artech House, (1997) Boston, London:*

[9] Teuvo KOHONEN, Self-Organizing Maps ,*Springer Series in Information Sciences,* 2001

[10] Hudaverdi BIRCAN,Metin ZONTUL,Ahmet Gurkan YUKSEK,A study of clustering exporting countries of Turkey using SOM Neural Networks, *Ataturk University Journal of Economics and Administrative Sciences,*Vol 20,No.2 ,2006, pp.219-238

[11] Laurene FAUSETT,Fundamentals of Neural Networks, *NJ:Prentice Hall,*1994