NETWORK PLANNING OF A VOIP-CAPABLE PBX
The use of data profiling techniques for an efficient network planning

Igor Ruiz-Agundez, Yoseba K. Penya and Pablo G. Bringas
DeustoTech, Deusto Institute of Technology, University of Deusto, Bilbao, Basque Country
{*igor.ira, , yoseba.penya, pablo.garcia.bringas*}*@deusto.es*

Keywords: Clustering algorithms; Network planning; VoIP services.

Abstract: Network planning presents a number of difficulties and risks but performed in an optimal way may turn beneficial. In this work we address a new way of achieving optimal network planning for Voice over IP (VoIP) services by using historical data to profile service usage. We show how to obtain a sound set of service use clusters based on calling behaviours by applying a Simple Expectation Maximisation (EM) algorithm. We successfully evaluate this methodology with real data and extract useful knowledge that can result in an improved network planning. Finally, we discuss the application of this method to other network services.

1 INTRODUCTION

Network planning aims at ensuring that a new network or service meets the needs of both the subscriber and the operator. These needs are sometimes in opposition, so that achieving a fair balance is not a trivial task. Traditionally, network administrators have used different methodologies to cope with this problem, including three steps: topological design, network synthesis, and network realisation. Each step must consider the dimensionality of user requirements to address problems such as peak-hour traffic and network resources consumption. In addition, the requirements of users change, and so decisions related to network planning must be approached iteratively.

In addition, every user within a company or an organisation presents unique needs. If there is a large number of users, it is hardly feasible to meet each of their individual requirements. Given this drawback, network administrators have traditionally used data profiles to group user requirements as well as possible. More specifically, a user is characterised in terms of one or more data profiles that describe her behaviour from different perspectives. This grouping process is traditionally performed by means of expert knowledge using a specialised method (Snasel et al., 2010).

Data profiling has been used to assign categories to data, develop quality metrics, assess risks, and visualise data, among other uses. Specifically, we use data profiling to extract knowledge that will facilitate decision-making in efficient network planning.

Among the services of the emerging convergence networks, Voice over Internet Protocol (VoIP) technology enables the transmission of audio data over an all-IP network. Still, as with more traditional technologies, VoIP also requires a network planning process. This process aims at guaranteeing that the subscribers, end-users, and operators of VoIP meet certain communication requirements.

In this study, we put forward the use of clustering algorithms to profile the behaviours of VoIP users. Since we do not know the number of clusters a particular algorithm may require, we use the *Simple Expectation Maximisation (EM)* algorithm because it provides different numbers of profiles.

Against this background, the contribution of this paper is three-fold. First, we introduce the use of the EM algorithm to profile users' *Call Detail Records (CDRs)*. Second, we evaluate this methodology for efficient network planning in the context of VoIP services within an organisational context. Finally, we discuss the proposed methodology and the possibility to apply it to other network services.

The remainder of this paper is organized as follows. Section 2 introduces the use EM algorithm. Section 3 illustrates empirical experiments and their results. Section 4 concludes the study and proposes avenues for future work.

2 Simple Expectation Maximisation (EM) algorithm

One of the main problems in data clustering algorithms lays on determining the number of clusters in a data set. This quantity is labelled as k, and its value is often ambiguous and depends on the application domain and the expert consulted. This optimal value k should strike a fair balance between the maximum compression of the data, which implies the use of a single cluster, and maximum accuracy, which is obtained by assigning a cluster to every data point (Kim, 2009).

There are several ways of determining the optimal number of clusters (Sugar and James, 2003). We can determine k manually by trying different values and analysing the results of the clustering algorithm until it is considered valid. This method, however, is normally used when an expert has an initial approximation of the existing number of clusters.

As we do not know the number of clusters into which our data set is divided, we employ the second method. Further, we use the EM algorithm, as it supports nominal, binary, empty nominal, and numeric attributes as well as allows us to manage missing data or unary attributes. This attribute types support that our data will be correctly processed (Keim and Hinneburg, 1999). Additionally, we decided to use this algorithm because, to our knowledge, it best fits our dataset as the results give a better knowledge representation, which is presented in Section 3.1. EM supports all of the attribute types we are working with, and it also identifies the number of clusters into which the dataset set must be divided. We also considered the proximity measure (i.e., how similar two data points are) and the clustering criterion (i.e., the cost function) of this algorithm for calculations.

3 EXPERIMENT AND RESULTS

3.1 VoIP data

We obtained our experiment data from the Asterisk PBX database records in a corporation. The data corresponds to all of the CDRs recorded in 2008, which comprise more than 700,000 entries.

These records are produced when two or more participants communicate over a partial or complete Internet-based voice connection. The call is normally initiated by one of the participants (i.e., the call initiator) and is received by one or more participants (i.e., call recipients). The call can be IP to IP, *Public Switched Telephone Network (PSTN)* to IP, IP to PSTN, mobile to IP or any other possible combination. In each case, the resulting CDR reflects the nature of the call and provides all the details on it.

For the sake of simplicity and due to our limited computational capabilities, we reduced the dataset to one month (October) with 48,849. To address data privacy concerns, all confidential information was made anonymous, guaranteeing the rights of users under compliance with the existing laws.

3.2 Interpretation of the results

The amount of clusters is defined by the results of the used clustering algorithm; the result was 10 different clusters. This amount of clusters is, in itself, highly representative, as it groups the CDRs and tells us how many different behaviours or requirements appear in the population of users. In other words, this number specifies the number of data profiles we are dealing with. With this information, we can already extract some useful conclusions.

If we are the service provider, we can modify the network infrastructure depending on the requirements of each profile. We can also predict network and infrastructure consumption. If we are the network administrator of the end-user's PBX corporation, we can negotiate *service-level agreements (SLA)* with the provider that best fit the requirements of each profile. The end-users should benefit from the quality improvements obtained by such measures taken by both the provider and the corporation's network administrator.

We can also perform a deeper analysis of the final results of the clustering. As previously mentioned, the output of these algorithms must be interpreted to extract added-value information for the current domain area. The first step in this interpretation consists of bringing the clustering algorithm results into a comprehensible format for analysis. We prefer visual data representation because it is a valuable method for intuitively analysing large amounts of data (Nocke et al., 2004). There are many clustering analysis visualisation methods, such as the rectangular view, the ThemeRiver technique and the Scatter Plot Matrix, among others (Chi, 2000). The second step is to interpret this representation with an expert in the domain area of the VoIP. This step should provide new information about the dataset and help us to infer conclusions regarding efficient network planning.

For representation, we chose the Scatter Plot Matrix because it is a visualisation method (Nocke et al., 2004) that has been proven to be reliable, and it can be applied to our dataset domain area. The dots in the Plot Matrix represent the centroid values of a certain

cluster, and the axis and the dot colours represent the attributes that are plotted. In this visualisation technique, we selected one attribute for $axis(x)$, one attribute for $axis(y)$ and one attribute for $class(colour)$. It is worth mentioning that attributes can be presented using more than one of these variables to obtain clearer information.

Among all the possible attribute visualisation combinations, we decided to focus on the hour that a call is established. We chose this attribute for four reasons. First, the hour indicates when calls are made, which helps us to detect peak demands so that better network planning can be performed. Second, some pricing schemes apply different price functions depending on when the service is used (e.g., morning or afternoon); with this information, we can negotiate better service level agreements with our providers (Chang and Petr, 2001). Third, the network administrator can see when a possible bottleneck may appear; this information can help improve the *quality of service (QoS)*. Fourth, both the provider and the end-user network administrator can balance their network and infrastructure loads depending on hourly demand.

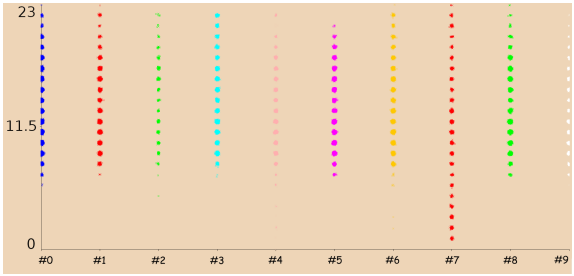

Figure 1: Clustering results. Axis(x) = Cluster. Axis(y) = Hour. Class(colour) = Cluster.

Figure 1 shows the clustering results. The cluster attribute appears both in terms of $axis(x)$ and $class(colour)$ to improve the readability of the graphic. The hour attribute corresponds to the moment a call starts. We can handle the call time in terms of hour for each data profile. All clusters have entries ranging from 07.00 to 22.00. This time range corresponds to the university's opening times. However, *cluster #7* presents entries all day, corresponding to special types of terminals, such as servers, modems and alarm systems. The calls related to the CDRs of this cluster may require special supervision to prevent call abuse. This information is also useful for modelling the normal behaviour of users and to anticipate the time of calls. We consider *cluster #7* to be exceptionally different from the rest.

Almost all of the data profiles show that users made their calls during working hours, and only one data profile had calls throughout the day. Table 1 shows the percentage of instances of the most relevant cluster in terms of instance number.

Table 1: Most relevant cluster instance percentages.

Cluster(s)	**Number of instances**	**Percentage**
Cluster #7	1846	3.78%
Remaining clusters	47003	96.22%
Total	48849	100%

This information could also be operated by the VoIP provider to improve the support systems. Information to prevent congestion may be inferred, marketing strategies shall be planned to offer special offers to the customers or infrastructure could be reorganised to improve efficiency.

In this case, we were thus able to propose an improvement for network planning. Currently, the corporation owns thirty channels in a primary trunk line with a time-based pricing scheme in which the calls are charged according to the time of call. The corporation should negotiate a call pricing scheme based on workday call times for 96.22% (or all but one) of their channels and a flat-rate pricing scheme for the remaining outgoing line. With this measure, the costs of PSTN calls should decrease for the end-user corporation. Table 2 presents best-case results following this network planning improvement. This maximum saving supposes that all calls of *cluster #7* are redirected through the flat-rate channel. As the cost per instance may vary, we simulated savings with $X = 10$ and $Y = 9$, obtaining a final savings of 35,074 monetary units, as shown in Table 2.

4 CONCLUSION AND FUTURE WORK

In this work, we have proposed a data profiling methodology for VoIP data for network planning. This methodology uses a clustering algorithm to process VoIP data records. The experiment shows that the proposed methodology is able to present knowledge in a highly comprehensible format. The possibility to accurately replicate this knowledge extraction methodology makes the proposed approach quite promising. We have used the EM algorithm to process our data records. This clustering algorithm determines the number of clusters into which the dataset is to be divided, and it splits the samples among them.

The results of the experiment demonstrate that it is possible to conduct data profiling using clustering algorithms. Indeed, the used clustering algorithm is one of the most widely applied algorithms. We could have

Table 2: Maximum possible savings through network planning improvements.

	Current situation	**With our network planning improvement**
Channels with a time-based pricing scheme	30	29
Channels with a flat-rate pricing scheme	-	1
Cost per instance for a time-based pricing scheme	X	X
Cost with a time-based pricing scheme per instance	-	Y
Cost	$\frac{48849}{30} \times 30X$	$\frac{47003}{29} \times 29X + 1846 \times Y$
Saving	$\frac{48849}{30} \times 30X - \frac{47003}{29} \times 29X + 1846 \times Y$	
Saving for $X = 10$ and $Y = 9$ monetary units	35074	

used other clustering algorithms and compared the results. The incremental clustering method is, however, the only one that with our data generates an explicit knowledge representation model that describes clustering in such a way that it can be easily visualised and understood (Witten and Frank, 2005).

We could have also incremented the granularity of the data, splitting the date attribute by year, month and day. Nevertheless, since the dataset in the experiment corresponded to one month, this criterion did not affect the results. We did not study whether or not the attributes are interdependent. Thus, for this experiment, all of the attributes were treated as covariant; additionally, Bayesian-network-based algorithms (Ruiz-Agundez et al., 2010b) may assist in the study of such dependencies.

We have proposed an improvement in network planning that could result in savings regarding telephone calls. The change consisted of adopting a flat-rate pricing scheme for one of the telephone channels (i.e., trunk lines) for a set of calls that would be more expensive with a time-based pricing scheme. We have contributed to improving the resource management and the service tariffing of the experiment corporation, further, this methodology could be generalised to any size corporation as the service usage data can always be clustered.

Finally, it is worth highlighting that the most exciting result of the proposed methodology is, in our opinion, the ability to produce highly representative results that could generate a high level of knowledge. The benefits and possible applications of such a methodology include: (i) user behaviour modelling, (ii) fraud detection through anomalous behaviour analysis (Ruiz-Agundez et al., 2010a), (iii) technical and economical implications (e.g. network planning), and (iv) pricing scheme modelling (Falkner et al., 2009).

Future work includes analysis of clustered data for use as training data for classification algorithms. These algorithms should allow us to predict the values of certain attributes. For example, we should be able to predict within a given probability a call's duration or the channel used to route the call given its source and destination.

REFERENCES

Chang, X. and Petr, D. (2001). A survey of pricing for integrated service networks. *Computer communications*, 24(18):1808–1818.

Chi, E. (2000). A taxonomy of visualization techniques using the data state reference model. In *Proc. of the IEEE Symposium on Information Visualization*, pages 69–75.

Falkner, M., Devetsikiotis, M., and Lambadaris, I. (2009). An overview of pricing concepts for broadband IP networks. *IEEE Communications Surveys*, 3(2):2–13.

Keim, D. and Hinneburg, A. (1999). Clustering techniques for large data sets from the past to the future. pages 141–181.

Kim, W. (2009). Survey topic: Parallel Clustering Algorithms.

Nocke, T., Schumann, H., and Bohm, U. (2004). Methods for the visualization of clustered climate data. *Computational Statistics*, 19(1):75–94.

Ruiz-Agundez, I., Penya, Y. K., and Bringas, P. G. (2010a). Fraud detection for voice over ip services on next-generation networks. In Samarati, P., editor, *Proceedings of the 4th Workshop in Information Security Theory and Practices (WISTP 2010)*, LNCS 6033, pages 199–212, Passau, Germany. IFIP International Federation for Information Processing 2010, Springer.

Ruiz-Agundez, I., Penya, Y. K., and Bringas, P. G. (2010b). Optimal bayesian network design for efficient intrusion detection. In *Proceedings of the 3rd International Conference on Human System Interaction (HSI 2010)*, pages 444–451, Rzeszow, Poland. IEEE.

Snasel, V., Abraham, A., Owais, S., Platos, J., and Kromer, P. (2010). User Profiles Modeling in Information Retrieval Systems. *Emergent Web Intelligence: Advanced Information Retrieval*, page 169.

Sugar, C. and James, G. (2003). Finding the Number of Clusters in a Dataset. *Journal of the American Statistical Association*, 98(463):750–763.

Witten, I. H. and Frank, E. (2005). *Data mining. Practical machine learning tools and techniques*. Diane Cerra.