Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

From IP port numbers to ADSL customer segmentation: knowledge aggregation and representation using Kohonen maps

F. Clérot & F. Fessant
France Télécom R&D, DTL/TIC/TNT, Lannion, France

Abstract

The very rapid adoption of new applications makes the segmentation of ADSL customers according to their network usage a critical step both for a better understanding of the market and for the prediction and dimensioning of the network. We show how Kohonen self-organizing maps can be used to build successive layers of abstraction starting from low-level traffic data to achieve an interpretable clustering of the customers and how the unique visualisation ability of the Kohonen maps makes the analysis quite natural and easy to interpret.

1 Motivation

Broadband access for home users and small or medium businesses and especially ADSL access is of vital importance for telecommunication companies since it allows them to leverage their copper infrastructure so as to offer new value-added broadband services to their customers.

The market for broadband access has three strong characteristics:

- there is a strong competition between the various actors,
- although the market is now very rapidly increasing, customer retention is important because of high acquisition costs,
- new applications or services may be picked up very fast by some segments of the customers and the behaviour of these applications or services may have a very strong impact on the quality of service delivered to all customers (and not only those using these new applications or services).

Two well-known examples of new applications or services with possibly very demanding requirements in terms of bandwidth are peer-to-peer file exchange systems and audio or video streaming.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

The above three characteristics explain the growing importance of an accurate understanding of the customer behaviour and of a better knowledge of the usage of broadband access. The very notion of "usage" is slowly shifting from a "bandwidth only" perspective to a much broader perspective which involves the discovery of the usage patterns in terms of applications or services. The knowledge of such patterns is expected to give a much better understanding of the market and to help anticipate on the adoption of new services or applications by some segments and allow the deployment of new resources before the effect hits all the customers.

"Bandwidth only" measurements are performed routinely at very large scale by telecommunication companies [1]. "Usage patterns" are most often inferred from polls and interviews which allow an in-depth understanding but suffer from the small size of the sampled population and cannot be easily extended to the whole population or correlated with measurements.

Below we report a new approach to the discovery of broadband customers usage patterns by directly mining network measurement data. This approach aims at giving an accurate insight in terms of applications or services usages patterns while being highly scalable and deployable.

After a short description of the data used in this study, we motivate and explain the data pre-processing and the main steps of the data-mining process. We then turn to the description of the data-mining technique and give a few illustrative results.

2 Network measurements and data description

The network measurements are performed on ADSL customer traffic by means of a proprietary network probe working at the SDH level between the Broadband Acces Server (BAS) and the Digital Subscriber Line Access Multiplexer (DSLAM). This on-line probe allows to read and store all the relevant fields of the ATM cells and of the IP/TCP headers. A detailed description of the probe architecture can be found in [2]. This is done transparently for the customer; no other use is made of such data than statistical analysis. For the study reported below, we gathered one month of data, on one site, for about one thousand customers. Once the probe in place, data collection is performed automatically.

In final form, the data give the volumes of data exchanged in the downstream direction on the 25 most active TCP ports (these ports were determined from a previous analysis of the traffics and include well-known ports such as ftp or http related ports and more "obscure" ports which later appeared to be used by various peer-to-peer file exchange systems) for every 6 minute window of each day and each customer (identified uniquely at the ATM level).

As such, these raw data suffer from two obvious drawbacks

- information on volumes exchanged at port level is too fine-grained to be easily understood in terms of applications or services usage;
- data volumes are very much likely to hide the usage patterns we are looking for as, for instance, a big e-mail user will likely consume much less bandwidth than a moderate peer-to-peer user.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

The first drawback is dealt with a two step data-mining analysis, as explained below, Section 3; the second drawback is dealt with data pre-processing, as explained below, Section 4.

3 From atomic activities to typical usage and to customer segmentation: a three step data-mining approach

The motivation of this study is a better understanding of the customers' usage patterns in terms of applications or services. Description of the customers' activity at the per day/ per port level is too fine-grained to give us a clear picture of what the usage patterns are. The problem is not only that different ports can be involved for a given application (ftp and ftp-data ports for the FTP application for instance) but also that what we are looking for is a set of combinations of applications (a pattern such as "big e-mail user and small peer-to-peer user").

There is a need to define what a proper "usage" is in terms of the "atomic usage data". Only when this notion of "usage" emerges from the data can we proceed to the segmentation of the customers in terms of such usages.

Our data-mining approach is therefore organized in three steps:

- in a first step, we first aggregate the per port data on a daily basis and end up with a set of per port volumes (one volume per port only) for each day and each customer : from now on, "usage" means "daily activity" (described by per port volumes). We then proceed to cluster the set of all the daily activities (irrespective of the customers). As a result, we end up with "typical daily activities" (the clusters) which allow us to understand how the customers are globally using their broadband access. These "typical daily activities" are interpreted as combination of high or low usages of different applications (usage patterns) and form the basis of all subsequent analysis and interpretations;
- in a second step, we turn to each individual customer described by their own set of "daily activities" and characterize him by a profile in terms of "typical daily activities" (each daily activity is attributed to its cluster);
- in a third step, we cluster the customers as described by their profiles and end up with "typical customers" (the clusters) described by profiles of "typical activities" (usage patterns).

This data-mining approach is reminiscent of Mobascher's approach to web usage mining where the "atomic data" (the pages seen by each customer with a time stamp) must first be aggregated at a relevant time-scale (the "visit") and classified so as to extract "typical visits". Then only, the behaviour of customers is analysed in terms of such typical visits [3]. Our clustering technique is however quite different and allows easier interpretations (see below).

4 Data pre-processing: the use of rank statistics

The distributions of the volumes on each port (all days and customers included) are known to be highly skewed with a lot of very small contributions and very heavy tails. Our data make no exception and, therefore, straightforward normalisations such as mean centring and normalising to unit variance will fail at

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

giving a reasonable aspect to the data. More complex modelling of the distributions is possible and many studies in traffic analysis are devoted to this topic; however, this may be quite involved and may prevent an easy automatic deployment of the analysis. Moreover, our objective is not to accurately represent each traffic but to be able to compare the level of usage of each traffic. Therefore, we simply rely on rank statistics: on a given port, each volume is replaced by its rank relative to the other volumes (all days and customers included) on the same port; the rank is then normalised between 0 and 1.

This normalisation is simple, robust and obviously delivers the kind of information we expect: a big (resp. small) customer on a given application should have many days with recoded values close to 1 (resp. 0) for the ports related to this application.

5 Clustering of daily activities

We choose to cluster our data with a Self Organizing Map (SOM). A SOM is a neural network made up of a set of nodes organized into a 2-dimensional grid. Each node has fixed coordinates in the map and adaptive coordinates (the weights) in the input space. The input space is relative to the variables setting up the observations. Two euclidian distances are defined, one in the original input space and one in the 2-dimensional space. The self-organizing process slightly moves the nodes coordinates in the data definition space -i.e. adjusts weights according to the data distribution. This weight adjustment is performed while taking into account the neighbouring relation between nodes in the map.

The SOM map has the well-known ability that the projection on the map preserves the proximities: close observations in the original input space are associated with close nodes in the map. For a complete description of the SOM properties and some applications see [4, 5].

We have developed a two level exploratory data analysis approach based on SOM. A first SOM is designed and trained with the observations. After learning has been completed, the map is segmented into a smaller set of clusters, each cluster being formed of similar nodes. To do this we run a hierarchical agglomerative clustering algorithm on top of the learned map [6]. This segmentation is made in order to simplify the quantitative analysis of the map.

In a second step, each input variable is described by its projection on the map of observations and these projections are transformed into vectors representative of the variables (the length of each vector being the number of nodes of the first map). A second SOM (the map of variables) is built with this second set of vectors. After learning, this map is also segmented in a smaller set of clusters with a hierarchical agglomerative clustering algorithm (the effect of this segmentation is to group together variables which have similar behaviour). Such approach allows the simultaneous visualization of clusters of observations and clusters of variables for exploratory analysis purposes. The two clusterings are consistent together: groups of observations have similar behaviour relative to groups of variables and vice versa. The scheme of the process is given fig. 1; more details and an application can be found in [7].

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

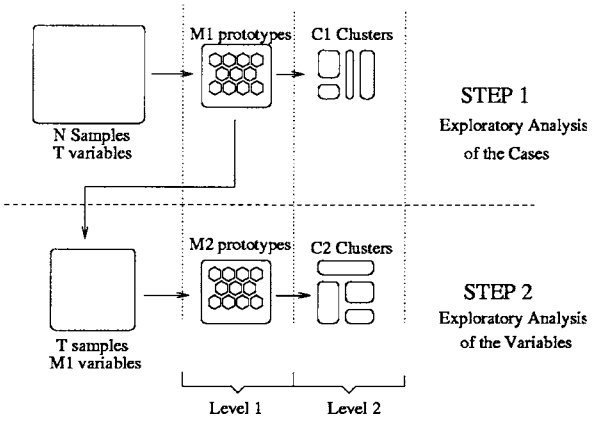


Figure 1: The two step two level approach.

We experiment with square maps with hexagonal neighbourhoods of size 12x12 for the clustering of observations and 5x5 for the clustering of variables.

Fig. 2 shows the 17 different groups of daily activities with similar behaviour revealed after the clustering of the first map; each of the clusters (the typical daily activities) being identified by a number (experiments on SOM have been done with the SOM Toolbox package for Matlab [8]). We can observe in the south-west corner of the map numerous little clusters.

The map of variables leads to the formation of 7 clusters of IP ports. Fig 3 shows the projections of the variables in one of them. Each map is representative of the corresponding variable on the map of observations. For each variable (each input dimension), the weights of its connections with the observations map nodes are represented regarding their arrangement in the map with a specific colour code which visualizes the spread of values of the component.

The three projections are very similar (i.e. with similar patterns in identical positions) indicating that corresponding variables are correlated and contribute in the same way to the description of the observations. Further investigation by experts of the domain revealed that, indeed, these three IP ports were commonly used by a peer-to-peer application (Direct Connect).

Fig 4. illustrates the characteristic behaviour of one typical daily activity of the map of observations (cluster 5). We plot the mean profile of the cluster, computed by the mean of all the observations that have been classified by the map nodes as belonging to the cluster. The variables of the profile are re-ordered and colour-coded according to their corresponding cluster of variable. We also give the deviation from the mean profile computed on all the observations.

The visual inspection of the figure shows that the typical daily activity associated with cluster 5 is mainly characterized by a high activity on classical applications (mail and http) and a high activity on FTP application.

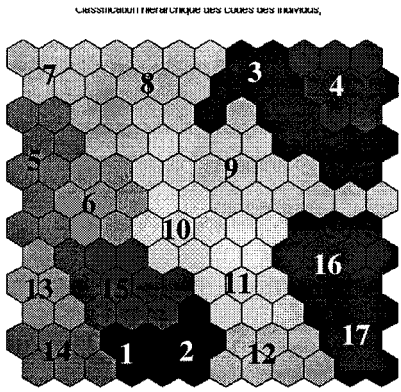


Figure 2: Hierarchical clustering of the SOM of daily activities.

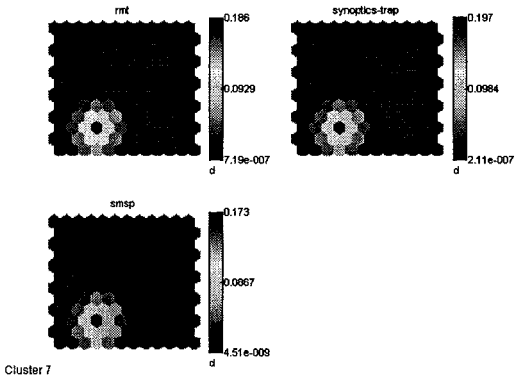


Figure 3: Projection of IP ports grouped in one cluster of variables.

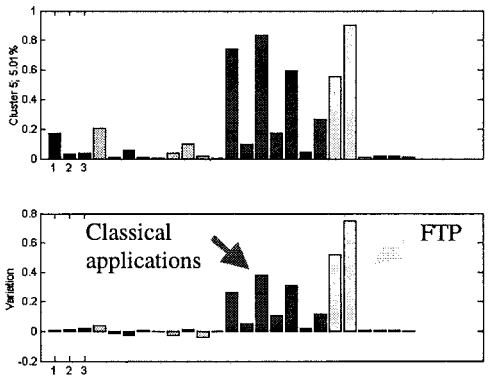


Figure 4: Profile of the typical daily activity 5 (up) and deviation from the mean profile (bottom).

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

All clusters can be described similarly, in terms of behaviour on applications (i.e. in terms of combination of low or high usage of the different applications). Discussions with an expert of the domain finally allowed to aggregate the 17 profiles in 6 main groups of profiles as follows:

1. clusters 3, 4 and 9 (north-east of the map) are characterised by a low global daily activity,
2. clusters 5, 6, 7, 8 and 10 (located at the north-west of the map) are characterised by a high daily activity on classical applications (mail, http) and ftp,
3. clusters 13, 14 and 15 (south west of the map) are characterised by a high daily activity on various Peer to Peer applications (except Kazaa),
4. clusters 1, 2, 11 and 12 (south of the map) are characterised by a high daily activity on Kazaa,
5. cluster 17 (south east of the map) is characterised by a high daily activity on http-alt (note that it is natural to find http-alt related activities in a specific cluster and not grouped with standard http related activities since http-alt ports are mainly used by http *proxies* while standard http ports are used by http *servers*),
6. cluster 16 (south east of the map) is characterised by a high daily activity on MsStreaming.

At this stage of the analysis we end up with a limited number of groups, the categories of typical daily activities, which characterize and summarize the activity of whole days on the different applications.

6 Description of customers on typical daily activities

We now turn to each individual customer whose activity has to be described on the basis of the clustering of daily activities detailed section 5. A customer is represented by his days as explained section 3. These days are projected on the clustering of daily activities obtained above; after this projection, each customer is represented by the proportion of days spent for each typical daily activity so as to define his "activity profile".

For example, suppose a customer doing 5 days of activity: 3 of these days are found, after projection on the map, to be low activity typical days, 1 is a day of classical activity and the last one is assigned to a typical Kazaa daily activity; the activity profile of this customer is therefore a 6-dimensional vector with the values [3/5, 1/5, 0, 1/5, 0, 0]. From now, each customer is represented by his activity profile composed of 6 descriptors.

7 Clustering of customers

A single SOM map is developed for the clustering of the customers defined by their activity profile (in the space of the typical daily activities). Fig. 5 shows the map obtained after its segmentation with a hierarchical agglomerative clustering algorithm; a number identifies each cluster. Finally, 17 clusters of customers have been created.

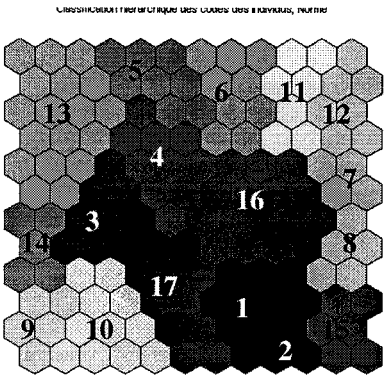


Figure 5: Hierarchical clustering of the SOM of customers.

This second clustering allows to link customers to usages. Fig. 6 details the characteristic behaviour of the cluster 15 (located in the south east corner of the map above, fig. 5). We plot the mean profile of the cluster, computed by the mean of all the customers that have been classified by the map nodes belonging to the cluster (6% of the population) and the mean profile of all the customers.

The mean customer profile distributes on low daily activities for 42%, on classical daily activity for 25%, on Peer to Peer but Kazaa daily activity for 15%, on Kazaa daily activity for 10%, on http-alt for 3% and on Ms-Streaming for 5%. The profiles of the clusters can be discussed according to their variation against the mean profile, in order to put forward their specific characteristics. For example, the typical customer represented by the profile of cluster 15 is characterized by a stronger contribution of classical daily activities and a lower contribution on low daily activity. The other components of the profile are close to the mean.

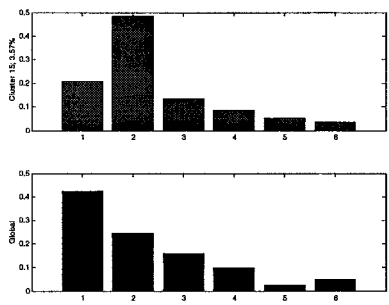


Figure 6: Profile of a typical customer (cluster 15, up) and mean profile (bottom).

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Table 1: Summary of the typical customer profiles in terms of typical daily activities.

Group index	Localisation on the map	List of clusters	Characteristics of the group
1	north-west, centre	3,4,5,6,13,16	low daily activities
2	south	1,2,10,17	Peer to Peer except Kazaa
3	south-west	9, 14	Ms-Streaming
4	north-east	7,11,12	Kazaa
5	south-east	8,15	classical daily activities

A detailed analysis of each cluster reveals 5 groups of typical customers profiles which are summarized table 1. An index identifies each group; we indicate also its localisation on the map given fig. 5, the list of the clusters belonging to it and its main characteristics in term of typical daily activities. It must be remembered that the "characteristic of the group" just sums-up what is specific to this group and does not represent the activity of the group: as can be seen on Figure 6, members of cluster 15, from the group labelled "classical daily activities" have more weight in their profile on the classical daily activities but also have weight on other activities.

The complete discussion of the behaviour of the different groups and of the clusters within those groups is however out of the scope of this paper. Let us just remark that the topological ordering inherent to the Kohonen map makes this discussion quite natural since clusters with similar behaviours will lie close on the map (fig. 5) and therefore it is possible to visualize how the behaviour evolves in a smooth manner from one place of the map to another.

The map is globally organized along two main axis: one axis goes from the north-west to the south-east, from low daily activities to intensive classical daily activities and the other goes from the south-west to the north-east, from Kazaa users to other peer-to-peer applications users.

8 Conclusion

This article has presented a complete scheme for the exploratory data analysis of ADSL customer behaviour from raw network data. This scheme relies on successive levels of abstraction obtained through the analysis and interpretation of Kohonen maps.

We are currently extending this work so as to investigate the variations of the usages from one site to another and the temporal behaviour of the clusters on a given site.

References

[1] Clément, H., Lautard, D. & Ribeyron, M., ADSL trafic : a forecasting model and the present reality in France. *WTC'2002*, Paris, France, September 2002.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

[2] François, J., Otarie: observation du traffic d'accès des réseaux IP en exploitation, *France Télécom R&D Technical Report FT.R&D/DAC-DT/2002-094/NGN* (in French), 2002.
[3] Mobascher, B., Dai, H., Luo, T. Nakagawa, M. & Sun, Y., Discovery of aggregate usage profiles for web personalization. *KDD'2000 Workshop on "Web mining for E-commerce, challenges and opportunities"*, Boston, 2000.
[4] Kohonen, T., *Self-Organizing Maps*, Springer-Verlag: Heidelberg, 2001.
[5] Oja, E. & Kaski, S., *Kohonen maps*, Elsevier, 1999.
[6] Vesanto, J. & Alhoniemi, E, Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11 (3), pp. 586-600, 2000.
[7] Lemaire, V. & Clérot, F., SOM-based clustering for on-line fraud behavior classification: a case study. *FSKD*, 2002.
[8] Vesanto, J., Himberg, J., Alhoniemi, J. & Parhankangas, K., SOM Toolbox for Matlab, *Report A57 Helsinki University of Technology, Neural Networks Research Centre*, Espoo, Finland , 2000.