Combining several SOM approaches in data mining: application to ADSL customer behaviours analysis

F. Fessant, V. Lemaire and F. Clérot

R&D France Telecom, 22307 Lannion, France
{francoise.fessant,vincent.lemaire,fabrice.clerot}@orange-ftgroup.com

Abstract. The very rapid adoption of new applications by some segments of the ADSL customers may have a strong impact on the quality of service delivered to all customers. This 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. Relying on a "bandwidth only" perspective to characterize network customer behaviour does not allow the discovery of usage patterns in terms of applications. In this paper, we shall describe how data mining techniques applied to network measurement data can help to extract some qualitative and quantitative knowledge.

1 Introduction

Broadband access for home users and small or medium business and especially ADSL (Asymmetric Digital Subscriber Line) 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 several 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 term of bandwidth are peer-to-peer file exchange systems and audio or video streaming.

The above characteristics explain the importance of an accurate understanding of the customer behaviour and a better knowledge of the usage of broadband access. The notion of "usage" is slowly shifting from a "bandwidth only" perspective to a much broader perspective which involves the discovery of 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 the adoption of new services or applications by some segments and allow the deployment of new resources before the new usage effects hit all the customers.

Usage patterns are most often inferred from polls and interviews which allow an in-depth understanding but are difficult to perform routinely, suffer from the small size of the sampled population and cannot easily be extended to the whole population or correlated with measurements (Anderson et al. (2002)). "Bandwidth only" measurements are performed routinely on a very large scale by telecommunication companies (Clement et al. (2002)) but do not allow much insight into the usage patterns since the volumes generated by different applications can span many orders of magnitude.

In this paper, we report another approach to the discovery of broadband customers' usage patterns by directly mining network measurement data. After a description of the data used in the study and their acquisition process, we explain the main steps of the data mining process and we illustrate the ability of our approach to give an accurate insight in terms of usages patterns of applications or services while being highly scalable and deployable. We focus on two aspects of customers' usages: usage of types of applications and customers' daily traffic; these analyses suppose to observe the data at several levels of detail.

2 Network measurements and data description

2.1 Probes measurements

The network measurements are performed on ADSL customer traffic by means of a proprietary network probe working at the SDH (Synchronous Digital Hierarchy) level between the Broadband Access 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 (Asynchronous Transfer Mode) cells and of the IP/TCP headers. From now, 9 probes equip the network; they observe about 18000 customers non-stop (a probe can observe about 2000 customers on a physical link). Once the probe is in place, data collection is performed automatically. A detailed description of the probe architecture can be found in (Francois (2002)).

2.2 Data description

For the study reported here, we gathered one month of data, on one site, for about two thousand customers. The data give the volumes of data exchanged in the upstream and downstream directions of twelve types of applications (web, peer-to-peer, ftp, news, mail, db, control, games, streaming, chat, others and unknown) sampled for each 6 minutes window for each customer. Most of the types of applications correspond to a group of well-known TCP ports, except the last two which relate to some well known but "obscure" ports (others) or dynamic ones (unknown). Since much of peer-to-peer traffic uses dynamic ports, peer-to-peer applications are recognized from a list of application names by scanning the payloads at the application level and not by relying on the well-known ports only. This is done transparently for the customers; no other use is made of such data than statistical analysis.

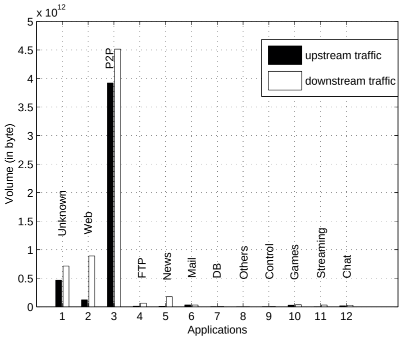

Fig. 1. Volume of the traffic on the applications

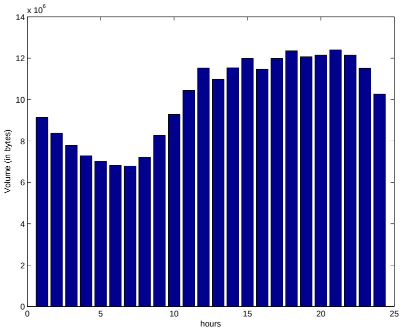

Fig. 2. Average hourly volume

Figure 1 plots the distribution of the total monthly traffic on the applications (all days and customers included) for one site in September 2003 (the volumes are given in bytes). About 90 percent of the traffic is due to peer-to-peer, web and unknown applications and all the monitored sites show a similar distribution. Figure 2 plots the average hourly volume for the same month and the same site, irrespective of the applications. We can observe that the night traffic remains significant.

3 Customer segmentation

3.1 Motivation

The motivation of this study is a better understanding of the customers' daily traffic on the applications. We try to answer the question: **who is doing what and when?**

To achieve this task we have developed a specific data mining process based on Kohonen maps. They are used to build successive layers of abstraction starting from low level traffic data to achieve an interpretable clustering of the customers.

For one month, we aggregate the data into a set of daily activity profiles given by the total hourly volume, for each day and each customer, on each application (we confined ourselves to the three most important applications in volume: peer-to-peer, web and unknown; an extract of the log file is presented Figure 3). In the following, "usage" means "daily activity" described by hourly volumes. The daily activity profiles are recoded in a log scale to be able to compare volumes with various orders of magnitude.

3.2 Data segmentation using self-organizing maps

We choose to cluster our data with a Self Organizing Map (SOM) which is an excellent tool for data survey because it has prominent visualization properties. A SOM is a set of nodes organized into a 2-dimensional[1] grid (the map). Each node has fixed coordinates in the map and adaptive coordinates (the weights) in the input space. The input space is spanned by the variables used to describe 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 location of the nodes 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 has the well-known ability that the projection on the map preserves the proximities: observations that are close to each other in the original multidimensional input space are associated with nodes that are close to each other on the map.

After learning has been completed, the map is segmented into clusters, each cluster being formed of nodes with similar behaviour, with a hierarchical agglomerative clustering algorithm. This segmentation simplifies the quantitative analysis of the map (Vesanto and Alhoniemi (2000), Lemaire and Clérot (2005)). For a complete description of the SOM properties and some applications, see (Kohonen (2001)) and (Oja and Kaski (1999)).

[1] All the SOMs in this article are square maps with hexagonal neighborhoods.

3.3 An approach in several steps for the segmentation of customers

We have developed a multi-level exploratory data analysis approach based on SOM. Our approach is organized in five steps (see Figure 6):

• In a first step, we analyze each application separately. We cluster the set of all the daily activity profiles (irrespective of the customers) by application. For example, if we are interested in a classification of web down daily traffic, we only select the relevant lines in the log file (Figure 3) and we cluster the set of all the daily activity profiles for the application. We obtained a map with a limited number of clusters (Figure 4): the typical days for the application. We proceed in the same way for all the other applications.

As a result we end up, for each application, with a set of "typical application days" profiles which allow us to understand how the customers are globally using their broadband access along the day, for this application. Such "typical application days" form the basis of all subsequent analysis and interpretations.

client	day	application	volume
client 1	day 1	unknown-up	volume-day-unknown-up-11
client 1	day 1	P2P-up	volume-day-P2P-up-11
client 1	day 2	unknown-up	volume-day-unknown-up-12
...	...	...	...
client 2	day 1	web-down	volume-day-web-down-21
client 2	day 3	unknown-up	volume-day-unknown-up-23
client 2	day 3	web-up	volume-day-web-up-23
client 2	day 3	web-down	volume-day-web-down-23
client 2	day 5	P2P-down	volume-day-P2P-down-25
...	...	...	...

Fig. 3. log file : each application volume (last column) is a curve similar to the one plotted Figure 2

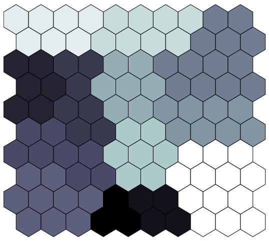

Fig. 4. Typical Web-down days

• In a second step we gather the results of previous segmentations to form a global daily activity profile: for one given day, the initial traffic profile for an application is replaced by a vector with as many dimensions as segments of typical days obtained previously for this application.

The profile is attributed to its cluster; all the components are set to zero except the one associated with the represented segment (Figure 5). This component is set to one. We do the same for the other applications. The binary profiles are then concatenated to form the global daily activity profile (the applications are correlated at this level for the day).

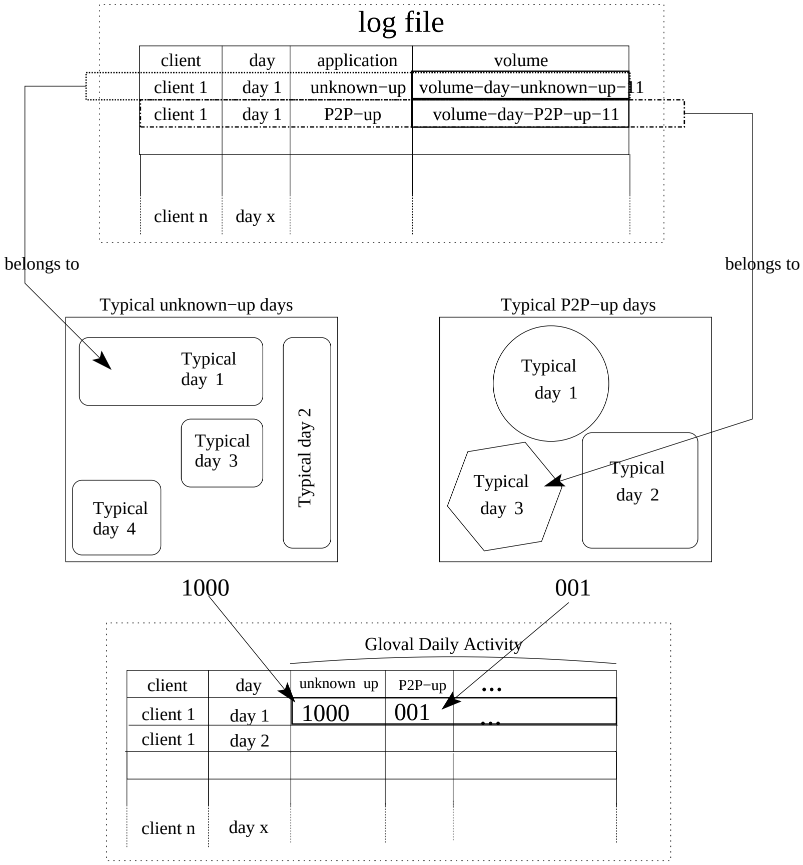

Fig. 5. Binary profile constitution

- In a third step, we cluster the set of all these daily activity profiles (irrespective of the customers). As a result we end up with a limited number of "typical day" profiles which summarize the daily activity profiles. They show how the three applications are simultaneously used in a day.

- In a fourth step, we turn to individual customers described by their own set of daily profiles. Each daily profile of a customer is attributed to its "typical day" cluster and we characterize this customer by a profile which gives the proportion of days spent in each "typical day" for the month.

- In a fifth step, we cluster the customers as described by the above activity profiles and end up with "typical customers". This last clustering allows to link customers to daily activity on applications.

The process (Figure 6) exploits the hierarchical structure of the data: a customer is defined by his days and a day is defined by its hourly traffic volume on the applications. At the end of each stage, an interpretation step allows to incrementally extract knowledge from the analysis results. The unique visualization ability of the self organizing map model makes the analysis quite natural and easy to interpret. More details about such kind of approach on another application can be found in (Clérot and Fessant (2003)).

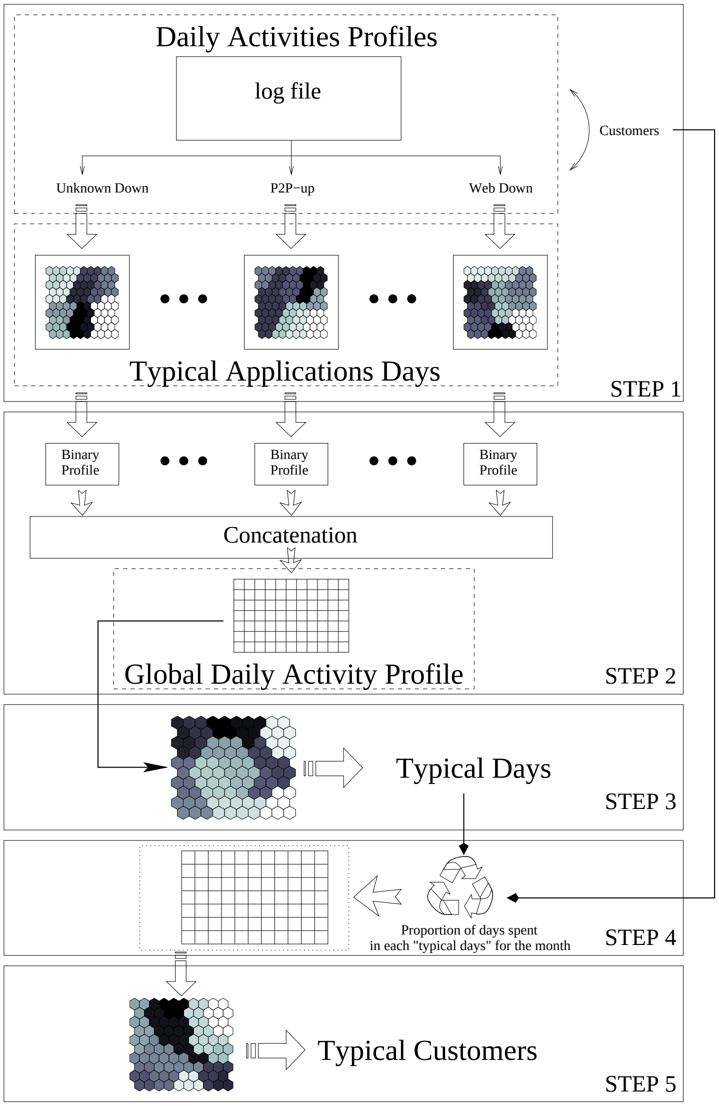

Fig. 6. The multi-level exploratory data analysis approach.

3.4 Clustering results

We experiment with the site of Fontenay in September 2003. All the segmentations are performed with dedicated SOMs (experiments have been done with the SOM Toolbox package for matlab (Vesanto et al. (2000)).

The first step leads to the formation of 9 to 13 clusters of "typical application days" profiles, depending on the application. Their behaviours can be summarized into inactive days, days with a mean or high activity on some limited time periods (early or late evening, noon for instance), and days with a very high activity on a long time segment (working hours, afternoon or night).

Figure 7 illustrates the result of the first step for one application: it shows the mean hourly volume profiles of the 13 clusters revealed after the clustering for the web down application (the mean profiles are computed by the mean of all the observations that have been classified in the cluster; the hourly volumes are plotted in natural statistics). The other applications can be described similarly.

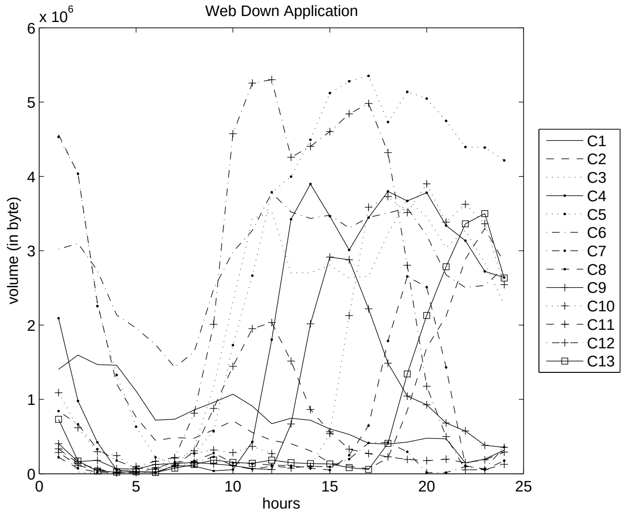

Fig. 7. Mean daily volumes of clusters for web down application

The second clustering leads to the formation of 14 clusters of "typical days". Their behaviours are different in terms of traffic time periods and intensity. The main characteristics are a similar activity in up and down traffic directions and a similar usage of the peer-to-peer and unknown applications in clusters. The usage of the web application can be quite different in intensity. Globally, the time periods of traffic are very similar for the three applications in a cluster. 10 percent of the days show a high daily activity on the three applications, 25 percent of the days are inactive days. If we project the other applications on the map days, we can observe some correlations between applications: days with a high web daily traffic are also days with high mail, ftp and streaming activities and the traffic time periods are similar. The chat and games applications can be correlated to peer-to-peer in the same way.

The last clustering leads to the formation of 12 clusters of customers which can be characterized by the preponderance of a limited number of typical days.

Figure 8 illustrates the characteristic behaviour of one "typical customer" (cluster 6) which groups 5 percent of the very active customers on all the applications (with a high activity all along the day, 7 days out of 10 and very little days with no activity). We plot the mean profile of the cluster (computed by the mean of all the customers classified in the cluster (up left, in black). We also give the mean profile computed on all the observations (bottom left, in grey), for comparison.

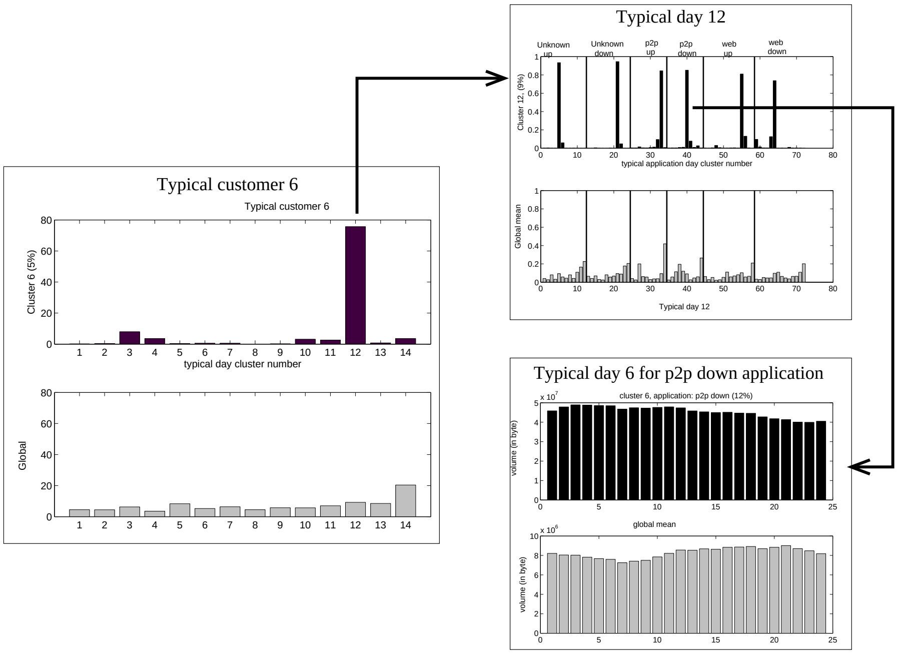

Fig. 8. Profile of one cluster of customers (up left) and mean profile (bottom) and profiles of associated typical days and typical application days

The profile can be discussed according to its variations against the mean profile in order to reveal its specific characteristics. The visual inspection of the left part of Figure 8 shows that the mean customer associated with the cluster is mainly active on "typical day 12" for 78 percent of the month. The contributions of the other "typical days" are low and are lower than the global mean. Typical day 12 corresponds to very active days. The mean profile of "typical day 12" is shown in the right top part of the figure in black. The day profile is formed by the aggregation of the individual application clustering results (a line delimits the set of descriptors for each application). We also give the mean profile computed on all the observations (bottom, in grey).

Typical day 12 is characterized by a preponderant typical application day on each application (from 70 percent to 90 percent for each). These typical application days correspond to high daily activities.

For example, we plot the mean profile of "typical day 6" for the peer-to-peer down application in the same figure (right bottom; in black the hourly profile of the typical day for the application and in grey the global average hourly profile; the volumes are given in bytes). These days show a very high activity all along the day and even at night for the application (12 percent of the days). Figure 8 schematizes and synthesizes the complete customer segmentation process.

Our step-by-step approach aims at striking a practical balance between the faithful representation of the data and the interpretative power of the resulting clustering. The segmentation results can be exploited at several levels according to the level of details expected. The customer level gives an overall view on the customer behaviours. The analysis also allows a detailed insight into the daily cycles of the customers in the segments. The approach is highly scalable and deployable and clustering technique used allows easy interpretations. All the other segments of customers can be discussed similarly in terms of daily profiles and hourly profiles on the applications.

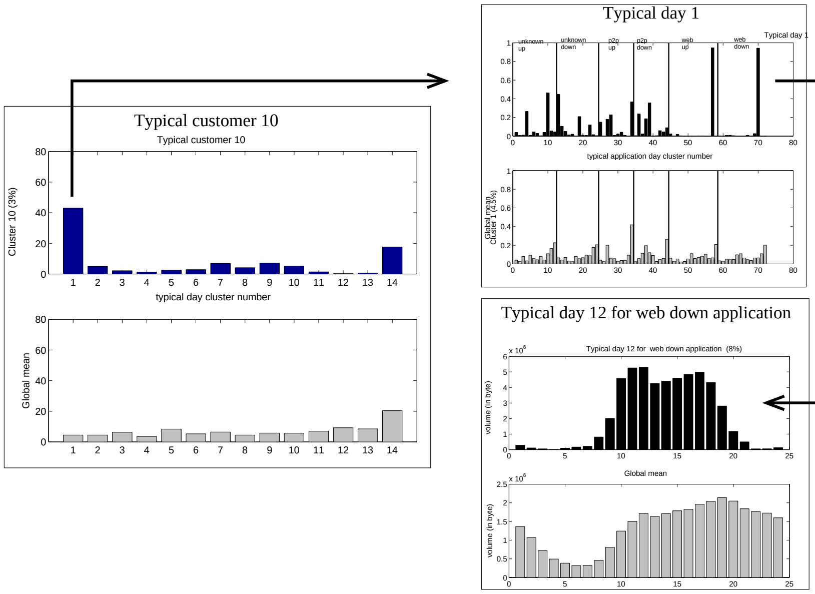

Fig. 9. Profile of another cluster of customers (top left) and mean profile (bottom) and profiles of associated typical days and typical application days

We have identified segments of customers with a high or very high activity all along the day on the three applications (24 percent of the customers), others segments of customers with very little activity (27 percent of the customers) and segments of customers with activity on some limited time periods on one or two applications, for example, a segment of customers with overall a low activity mainly restricted to working hours on web applications. This segment is detailed in Figure 9.

The mean customer associated with cluster 10 (3 percent of the customers) is mainly active on "typical day 1" for 42 percent of the month. The contributions on the other "typical days" are close to the global mean. Typical day 1 (4.5 percent of the days) is characterized by a preponderant typical application day on web application only (both in up and down directions); no specific typical day appears for the two other applications. The characteristic web days are working days with a high daily web activity on the segment 10h-19h.

Figure 10 depicts the organization of the 12 clusters on the map (each of the clusters is identified by a number and a colour). The topological ordering inherent to the SOM algorithm is such that clusters with close behaviours lie close on the map and 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 an axis going from the north east (cluster 12) to the south west (cluster 6), from low activity to high activity on all the applications, non-stop all over the day.

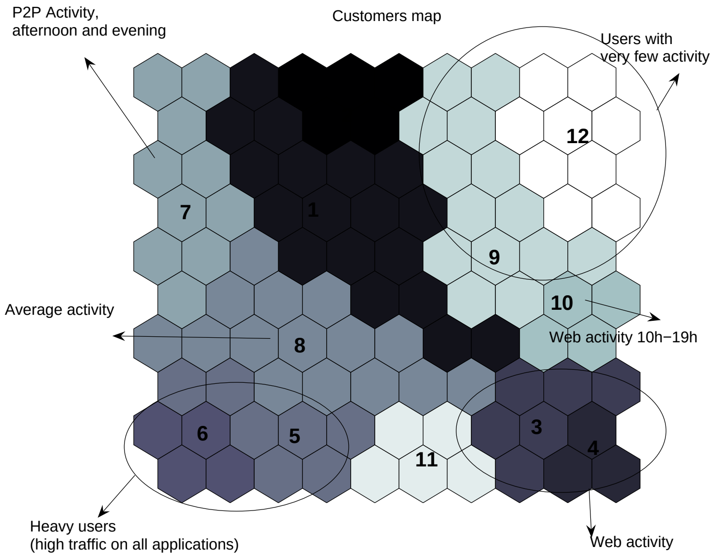

Fig. 10. Interpretation of the learned SOM and its 12 clusters of customers

4 Conclusion

In this paper, we have shown how the mining of network measurement data can reveal the usage patterns of ADSL customers. A specific scheme of exploratory data analysis has been presented to give lightings on the usages of applications and daily traffic profiles. Our data-mining approach, based on the analysis and the interpretation of Kohonen self-organizing maps, allows us to define accurate and easily interpretable profiles of the customers. These profiles exhibit very heterogeneous behaviours ranging from a large majority of customers with a low usage of the applications to a small minority with a very high usage.

The knowledge gathered about the customers is not only qualitative; we are also able to quantify the population associated to each profile, the volumes consumed on the applications or the daily cycle.

Our methodologies are continuously in development in order to improve our knowledge of customer's behaviours.

References

ANDERSON, B., GALE, C., JONES, M., and McWILLIAMS, A. (2002). Domesticating broadband-what consumers really do with flat-rate, always-on and fast Internet connections. *BT Technology Journal*, 20(1):103–114.

CLEMENT, H., LAUTARD, D., and RIBEYRON, M. (2002). ADSL traffic: a forecasting model and the present reality in France. In *WTC (World Telecommunications Congress)*, Paris, France.

CLEROT, C. and FESSANT, F. (2003). From IP port numbers to ADSL customer segmentation: knowledge aggregation and representation using Kohonen maps. In *DATAMINING IV*, Rio de Janeiro, Brazil.

FRANCOIS, J. (2002). 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).

KOHONEN, T. (2001). *Self-Organizing Maps.* Springer-Verlag, Heidelberg.

LEMAIRE, V. and CLEROT, F. (2005) The many faces of a Kohonen Map,. *Studies in computational Intelligence (SCI) 4, 1-13 (Classification and Clustering for Knowledge Discovery).* Springer.

OJA, E. and KASKI, S. (1999). *Kohonen maps.* Elsevier.

VESANTO, J. and ALHONIEMI, E. (2000). Clustering of the self organizing map. In *IEEE Transactions of Neural Networks.*

VESANTO, J., HIMBERG, J., ALHONIEMI, E., and PARHANKANGAS, J. (2000). Som toolbox for matlab 5. Technical Report Technical Report A57, Helsinki University of Technology, Neural Networks Research Centre.