

An SOM approach to analyse a web portal usage

Antonio Soriano-Asensi
Department of Applied Physics
University of Valencia, Spain
Email: Antonio.Soriano-Asensi@uv.es

José D. Martín-Guerrero
and Emilio Soria-Olivas
Department of Electronics Engineering,
University of Valencia, Spain
Email: {jdmg,soriae}@uv.es

Alberto Palomares
and Emili Balaguer-Ballester
R & D Department
Tissat, plc, Spain
Email: {apalomares,ebalaguer}@tissat.es

Abstract—In this work we are interested in the analysis of a citizen web portal usage, *Infoville XXI* (<http://www.infoville.es/>), in order to carry out an appropriate customization of such a portal. A straightforward way to customize the portal is based on analysing web-portal-user behaviours, thus being able to find similar users as well as differences among user behaviours. In particular, we will focus on useful representations of a set of users in order to better understand their behaviours. Due to the high number of services included in this portal, a representation of users in a two-dimensional plot becomes difficult to visualize. In order to overcome these limitations, we have used Self-Organizing Maps to analyse the way users interact with the web portal. The Self-Organizing Map has demonstrated to be a powerful algorithm that enables the projection of high dimensional data sets into two-dimensional maps. Several visualizations are tested to improve the analysis of the results provided by the Self-Organizing Map. These visualizations are used to obtain information about the characteristics of users, which can be used to improve the portal by means of personal customization.

I. INTRODUCTION

The Self-Organizing Map (SOM) is a neural network proposed by Kohonen [1], which maps a distribution of vectors of arbitrary dimension into a lower-dimensional space, typically one or two, while maintaining a high degree of topological ordering, or neighbourhood preservation. Therefore, it becomes a powerful tool in representing high-dimensional data.

In this work, SOM is used to analyse data from a citizen web portal, *Infoville XXI* (<http://www.infoville.es/>). Citizen web portals are an interactive gateway between citizens and the public administration. They involve citizens in the Information Society by offering a high number of services on the Internet, creating a new model for service delivery to the public as a result of the interaction between the basic services provided by the Government and private entities, which ends up at the citizen who made the request. The success and acceptance of these portals depend largely on their ability to attract the citizens and the public and private entities in the area. Finding out inter-user similarities and, in turn, creating groups of users with similar tastes helps in the customization of the portal. This is an easy way to make the site attractive to the majority of people.

An analysis of the interaction of users with the portal is required for a well-customized portal. However, in the case of *Infoville* as well as in other many web portals, this analysis becomes difficult due to the amount of accessed services. A

dimensionality reduction can be carried out by using the so-called descriptors, namely, thematic sections or labels which gather similar services without involving a relevant loss of information. However, a medium size portal can be formed by more than 20 descriptors, i.e., even using descriptors, the visualization of users becomes difficult due to the dimensionality of the vectorial space in which users are described (descriptor space). Therefore, a projection procedure, which ensures that whichever two similar vectors in the descriptor space must be represented close each other in the output space, is required to represent the portal. This task is done very efficiently by SOM-based algorithms [1]. Since SOM combines projection with vector quantization, the original data set is reduced after the training to a smaller, but still representative, data set to work with.

SOM-based algorithms have been widely used in data mining [2], [3]. A well-known application of the SOM algorithm in web mining is given by the *WEBSOM*. The *WEBSOM* is a method for organizing miscellaneous text documents onto meaningful maps for exploration and search [4], [5]. More recently, an SOM was used to analyse the information contained in a web page [6]. This is similar to our objective, although we are rather interested in users. The main purpose of our work is to make easier the work of the portal administrator by providing him with information related to users' behaviour. Several visualizations are taken into account in order to show the obtained results after the training with the SOM algorithm. Visualizations like component planes or data histograms [7] have demonstrated to be useful in order to get information about the shape and structure of a data set; in addition, a plot mixing both component planes and data histograms is used to analyse web users' behaviour. Other common visualizations, such as matrix distances [7] or smoothed data histograms [8] were also tested, but they did not provide extra information about the structure of the data set. Therefore, these latter results are not shown in this work.

The remainder of the paper is outlined as follows. The data set used in this study is presented in Sec. II. Sec. III describes the use of the Growing Grid algorithm to arrange user patterns in a two-dimensional mesh. The obtained results are shown in Sec. IV, ending up the paper with some conclusions and proposals for further work in Sec. V.

II. CITIZEN WEB PORTAL DATA: INFOVILLE XXI

We profiled user accesses to the region web portal Infoville XXI. This is an official web site supported by the Valencian Government, which provides citizens from Valencia, Spain, with more than 2,000 services, grouped into 22 descriptors, namely, public administration, agenda/events, children's area, town councils, street maps, channels (it consists of information on four specific matters: education, job-hunting, setting up business and housing), shopping, Infoville community (it enables the communication of people who access the portal by e-mail, fora, postcards, bulletin boards, ...), Infoville diary, education and training, finance, information for citizens, internal, register (internal and register are descriptors used for administration purposes), Lanetro (local information about where to eat, drink, dance, ...), SMS messages, entertainment, electronic newspapers, tourism in Valencia, national and international tourism, search and user utilities (personal agenda, site customization, personal web page, helping guide, ...). The hierarchical structure of the portal offers the information, first by descriptors, and then, by services. The most popular objects are highlighted, and they can be accessed by clicking on them from the main page. More than 50,000 homes are currently connected to Infoville XXI, recording more than 2 million accesses to date. Furthermore, the term *Infoville*, which was coined by the Valencian Government, has gone beyond the Spanish frontiers, reaching other European countries and even some non-European countries, such as Argentina or South-Africa.

We have used accesses from June 2002 to February 2003. The data recorded consists of user ID (a random number which does not offer any information about the identity of the user), session ID and service ID, together with the date and time corresponding to each access. A preprocessing procedure was carried out to eliminate data which did not provide useful information for our goals [9]:

- *Elimination of data corresponding to anomalous users.* Administrators of the portal tend to generate a high number of accesses for test purposes. These accesses are not useful for user profiling, and therefore, were removed from the data set. Moreover, some users only accessed the web portal once during the study period, and hence, could be considered as lost users. Finally, a low number of users appeared to be *outliers*, as they recorded an excessively high number of accesses; since they could bias the models, they were also removed from the data set.
- *Elimination of extremely accessed descriptors.* Those descriptors that recorded an extremely high or low number of accesses (more than three standard deviations from the mean value) were also removed, as they can bias the models obtained by SOM. As a result of this stage, only sixteen descriptors were taken into account for the visualization analyses.

For the sake of simplicity, only results with a reduced set of five descriptors are shown in this paper. The choice of these

five descriptors was carried out by analysing users' frequency of accesses as well as the relevance assigned to descriptors by people in charge of Infoville. A comparison between the whole set of sixteen descriptors and this reduced set showed the suitability of this choice to represent the web portal usage appropriately [9]. The reduced data set was formed by 1,676 patterns which consisted of the frequency of accessing the different descriptors by users within a certain session.

III. GROWING GRID ALGORITHM

The SOM algorithm introduced by Kohonen [1] consists of a set of neurons usually arranged in a one or two-dimensional grid. Although higher dimensional grids are also possible, they are hardly ever used because of their problematic visualization. Every neuron has a fixed position in the grid, and is represented by an n -dimensional weight vector $\mathbf{m} = [m_1, m_2, \dots, m_n]$, where n is the dimensionality of the input space.

A user pattern \mathbf{x} is randomly chosen from the data set on each training step. Then, the neuron whose weight vector is the most similar to the user pattern is searched, being this neuron the so-called Best Matching Unit (BMU). The weight vectors of the BMU and its neighbourhood are updated as follows:

$$\mathbf{m}^{t+1} = \mathbf{m}^t + \alpha(t)h(t)(\mathbf{x} - \mathbf{m}^t) \quad (1)$$

where t stands for the iteration number, $\alpha(t)$ is the learning rate, and $h(t)$ the neighbourhood kernel, whose centre is located at the BMU. The neighbourhood kernel determines which neurons around the BMU are updated, and how this update bears upon each neuron. During the training process, the weight vectors are updated, thus adapting the grid to the data.

A drawback of the SOM algorithm, which was indeed observed when analysing web users, stems from the size of the grid, which must be fixed at the beginning of the training process. Sometimes, the training has to be repeated since the right size of the SOM is not chosen at the beginning, and therefore, some trainings are required until the right size of the SOM is found out. In order to overcome this problem, the Growing Grid algorithm (GG) [10] is used in this work to analyse the behaviour of web users. This algorithm allows the SOM to grow by inserting a complete row or column of neurons, hence, it does not require to set the size of the output layer before the training process.

A criterion is required to decide where the new neurons are inserted, and to control the growing process. The original GG [10] introduces a resource variable which is increased every time a neuron is the BMU. The neuron whose resource variable is the highest is searched after the training in order to add neurons next to this neuron. This is similar to search the neuron which represents the highest number of user patterns at the end of the training process. In our case study, better results were obtained when considering the quantization error e_q , shown in Eq. (2), than considering the number of user patterns associated to neurons. This is because there were

several users who showed the same pattern, and therefore, all of them were well represented by one neuron. The criterion which considers the number of patterns per neuron would insist in adding neurons next to well represented regions with a high number of identical patterns.

$$e_q = \sum_{k \in N} \sum_{i \in k} \frac{\|\mathbf{m}_k - \mathbf{x}_i\|}{e_0} \quad (2)$$

where N is the number of neurons which make up the SOM, k denotes a particular neuron, and i is an index which refers to the patterns associated to neuron k . The initial quantization error e_0 (standard deviation of the data set) is introduced to reduce the effect of the data set distribution on the stopping criterion. Thus, e_q provides information about the improvement of the data set representation due to the addition of neurons.

The GG algorithm can be summarized as the following iterative procedure:

- 1) Search of the neuron with the highest quantization error after each SOM training (“e” in Fig. 1).
- 2) The dissimilarities between the “e” neuron and its neighbours are calculated by considering the Euclidean distance. The most dissimilar neighbour is marked as “d” in Fig. 1.
- 3) A new row or column is inserted between “e” and “d”, as shown in Fig. 1. The dark-shaded neurons in Fig. 1 correspond to added neurons.
- 4) The weight vectors of these added neurons, \mathbf{m}_{new} , are initialized by using the mean value of the weight vectors of “e” (\mathbf{m}_e) and “d” (\mathbf{m}_d).

$$\mathbf{m}_{\text{new}} = \frac{\mathbf{m}_e + \mathbf{m}_d}{2} \quad (3)$$

- 5) A new training with the SOM algorithm is performed.
- 6) Return to 1 until any of the following criteria is fulfilled:
 - e_q is fewer than a predefined value.
 - The improvement of e_q is lower than a certain threshold.

The training with the SOM algorithm is faster than with GG because the GG repeats the training with the SOM until the grid size is appropriate to represent the data set. However, the SOM requires to know the correct size of the map in advance, while the GG algorithm is capable of finding the best size of the output layer to represent the data set by itself. Apart from this capability, the characteristics of the SOM and the GG are basically the same.

An improvement of GG algorithm is given by the growing hierarchical SOM (GHSOM) [11]. In addition to providing similar results to GG, it also reduces considerably the time required to train the map because of its hierarchical procedure to split data. However, since the visualization of hierarchical results become rather difficult to interpret, results with this algorithm are not shown in this paper.

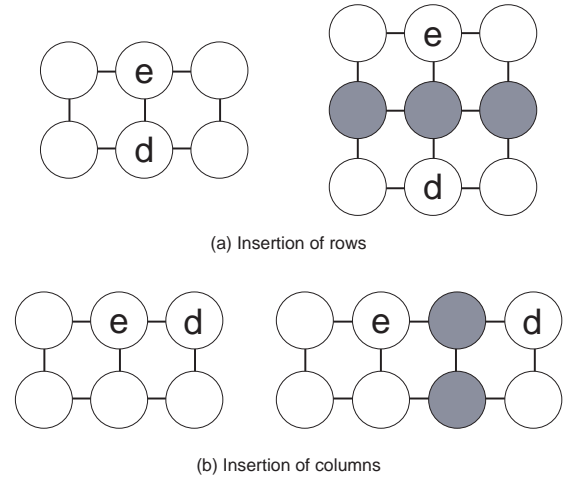


Fig. 1. Growing Grid algorithm: insertion of neurons.

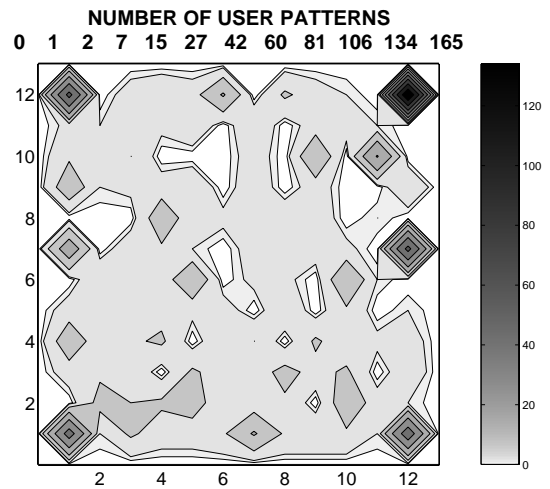


Fig. 2. Distribution of users in the output layer.

IV. RESULTS

The information obtained after the training with the GG algorithm was the number of users and the features represented by each neuron of the output layer. The number of users were represented in data histograms while the features of user patterns were plotted in component planes. Both the number and features of users were plotted in two-dimensional maps, where these characteristics were represented considering the position of the neurons in the SOM. The representation of user features in a regular grid allows a straightforward comparison among different visualizations.

The distribution of users in the output layer is presented in a data histogram (Fig. 2) [7], in which the number of user patterns associated to each neuron is shown. From this representation it is possible to find out which parts of the map represent more users. A friendly visualization of this information was obtained by plotting the contourlines of the data histogram, as shown in Fig. 2. The number of users associated to each neuron was plotted in a greyscale representation. The

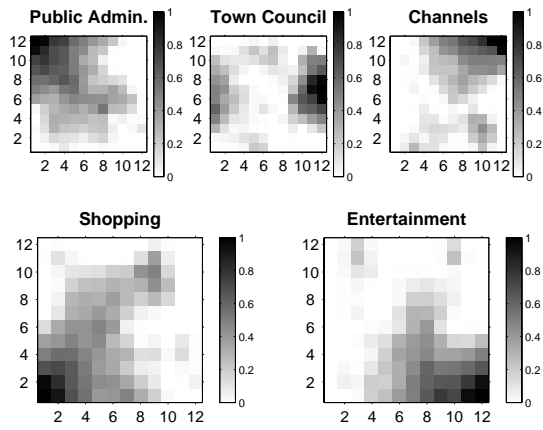


Fig. 3. Weight vectors of each neuron for the different descriptors. The darkest neurons represent users with the highest likelihood of accessing the descriptor.

more users the neuron had associated the darker its colour.

The five darkest regions in Fig. 2 were placed in the corners, and in the right part of the map at middle height. Fig. 3 shows that these regions represented users who only accessed one descriptor. Taking into account the number of users involved in these darkest regions, the conclusion is that almost half of users were interested in services included in only one descriptor.

The component planes [7] of Fig. 3 show information about how user features were mapped into each region of the output layer. In particular, Fig. 3 shows the weight vectors associated to each neuron for the different descriptors in different subplots. By using an RGB-plot instead of a greyscale representation, it is possible to group triplets of descriptors and represent them in the same figure, although sometimes the interpretation of the results becomes rather difficult.

Since user patterns were coded according to frequency of access to a descriptor, and weight vectors represented user patterns, Fig. 3 actually shows the likelihood of users to access a descriptor. This likelihood, which can range between zero and one, is plotted in a greyscale representation, in which the higher likelihood the darker colour. In other words, a likelihood equal to zero is represented by white colour whereas a likelihood equal to one is represented by black colour.

Fig. 3 shows that users interested in public administrations were located in the upper-left corner of the map; those interested in town council information were placed in the right side of the map, at middle height; users who mainly accessed the descriptor 'Channels' were placed in the upper-right corner; and finally, users who accessed shopping and entertainment were located in the lower-left and lower-right corners of the output layer, respectively.

The information shown in Fig. 2 and Fig. 3 was merged into a unique plot (Fig. 4) in order to obtain a more comprehensive representation of the results. It is similar to component planes shown in Fig. 3, but in this case the size of cells depends on the number of user patterns represented by each neuron. The

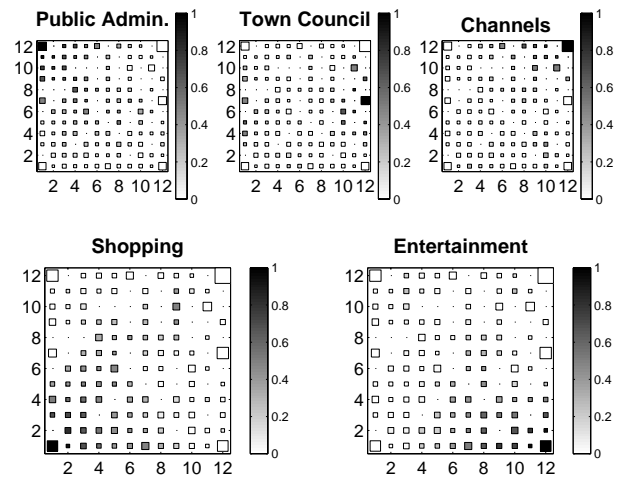


Fig. 4. Weight vectors and number of users represented by neurons.

relationship between cell sizes and the number of user patterns was not linear. Since the number of user patterns associated to each neuron was very different, the visualization was enhanced by introducing a non linear relationship between the number of associated users and the cell size.

The reason to introduce size as a relevant variable was to obtain qualitative information of the number of users associated to neurons. The size of cells plotted in Fig. 4 is proportional to the cube root of the number of users represented by the neuron. It is easier to obtain information about the spreading of user patterns from Fig. 4 than from Figs. 2 and 3. Fig. 4 shows both the kind of users represented by each neuron and a qualitative approach about the number of users associated to the neuron. As the relationship between cell sizes and the number of patterns was not linear, the exact number of user patterns represented by a neuron was obtained from Fig. 2 instead from Fig. 4.

Fig. 2 shows that there were a few neurons which represented more than 20 users, while most of neurons had associated less than 2 users. Moreover, there were many users who only accessed services from a unique descriptor. All of them were appropriately represented by one neuron because all of them had the same pattern. Since SOM training can be interpreted as if neurons are pulled by each data pattern, many patterns repeated in the data set introduce an extra stress in the SOM that makes difficult the correct mapping of regions with fewer neurons. In other words, the mapping of regions between neurons which had associated a lot of user patterns might not be good enough.

In order to achieve a better mapping of these regions, patterns of users who only accessed one descriptor were removed from the data set. The resulting reduced set contained a 60 % of the user patterns from the original data set. The training process with the GG was repeated with this new data set. As shown in Fig. 5, the distribution of users in the output layer was smoother than in the complete data set. The number of user patterns associated to each neuron ranged from 0 to 28,

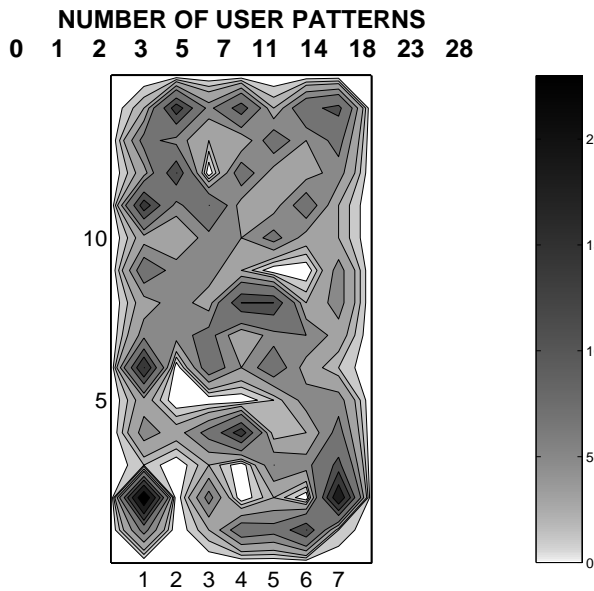


Fig. 5. Distribution of the reduced set of users in the output layer.

but most of neurons had associated a number of patterns that ranged from 4 to 8. These results indicate that user interests were shared among several descriptors in many different ways, i.e., there were a wide range of user preferences. Therefore, a relatively high number of neurons to represent the data set was required. Although the majority of neurons had associated from 4 to 8 user patterns, there were some patterns which were frequently repeated. This fact led to regions where the number of users was larger than in the rest of the map. For instance, there were two regions of neurons which represented a high number of user patterns close to the lower-left and lower-right corners of the map (Fig. 5). This means that there was a high number of users with a similar behaviour mapped into these regions. As it is shown in Fig. 6, these regions corresponded to users who accessed both 'Public Administration' and 'Town Councils' in the lower-right region, and users who accessed 'Town Councils' and 'Channels' in the lower-left region.

As every user who was interested in only one descriptor was removed, there were no cells represented in black in Fig. 6. The likelihood of accessing each descriptor was spread over the map, but there were different emphasized regions associated to each descriptor. These regions were more clearly appreciated in Fig. 3 than in Fig. 6 because of the stress introduced by the patterns of users interested in only one descriptor.

By merging the information about the number of user patterns associated to each neuron and its weight vectors (Fig. 7), the interpretation of results became easier. In this case, as the number of patterns associated to every neuron was similar, a linear relationship was chosen between cell sizes and the number of patterns associated to neurons.

In order to reference any neuron, the following notation was used: its position was indicated as (row, column). Rows were numbered from the bottom to the top of the map, and columns

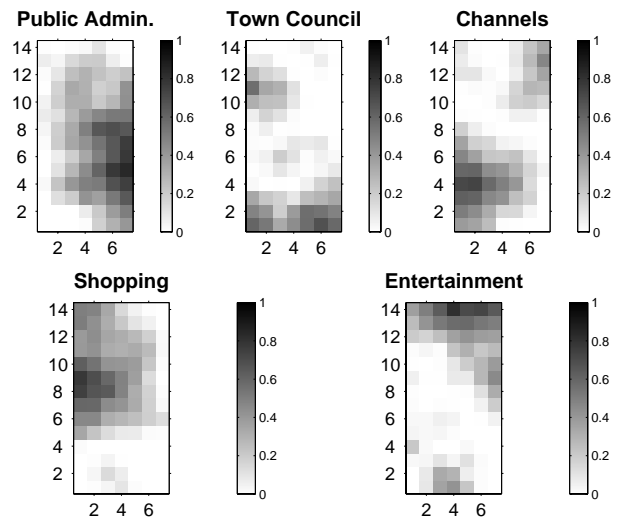


Fig. 6. Weight vectors of each neuron for the reduced data set.

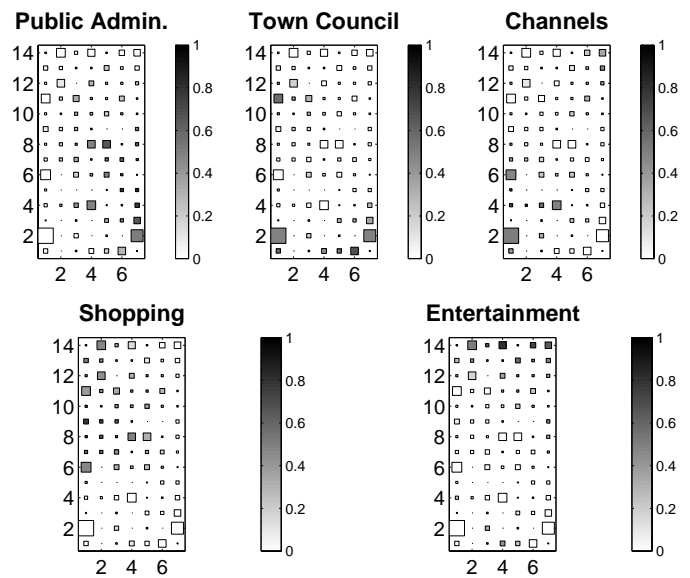


Fig. 7. Weight vectors and number of users represented by neurons for the reduced data set.

were numbered from the left to the right side of the map. The neuron with the highest number of associated patterns was neuron (2,1), which was located in the lower-left region of the map. This neuron represented those users who were interested in descriptors 'Channels' and 'Town Councils'. In the lower-right region of the map, there were three neurons which represented a group of users interested in descriptors 'Town Councils' and 'Public Administration'. These neurons were the following: (3,7), (2,7) and (1,6). Users represented by (3,7) were more interested in 'Public Administration' than in 'Town Councils', while users represented in (1,6) were more interested in 'Town Councils'. All users represented by these three neurons could be merged into a group representing users interested in 'Town Councils' and 'Public Administration'.

Although each of these three neurons represented less users than neuron (2,1), the whole group represented by these three neurons was bigger than the aforementioned group of users interested in 'Channels' and 'Town Councils'.

The aforementioned groups of users were the most important groups shown in Fig. 7, but there were other smaller groups of users to take into account:

- (6,1) represented users interested in 'Channels' and 'Shopping'.
- (8,4) and (8,5) represented users interested in 'Shopping' and 'Public Administration'.
- (4,4) and (4,3) represented users interested in 'Channels' and 'Public Administration'.
- (14,2), (14,3) and (14,4) represented users interested in 'Shopping' and 'Entertainment'.

Summarizing, many users were interested in only one descriptor. Almost the rest of users were interested in services included in two descriptors. Moreover, there was a group formed by a large amount of users, represented by neurons (11,1), (12,2) and (13,1), who were interested in 'Shopping', 'Town Councils' and 'Public Administration'.

V. CONCLUSIONS AND FURTHER WORK

The capability of SOM to arrange high-dimensional data in a two dimensional grid has been successfully exploited in this work to analyse the behaviour of users from the web portal Infoville XXI. A low-computational-burden tool has been implemented to analyse the shape and structure of a set of web users. Both SOM and GG algorithms have been applied to obtain a wide range of visualizations to explore a data set of users from the citizen web portal Infoville XXI.

Since SOM is not capable of adapting the size of the map to the requirements of each data set, the GG algorithm has been used to represent user behaviours. It has provided better results than SOM in exchange for the higher computational burden involved by GG. Several visualizations, like component planes or data histograms have been used to show the training results and also to analyse the characteristics of the data set, namely, the different user behaviours that appear in this citizen web portal. Therefore, these visualizations have helped portal administrators in order to identify groups of similar users.

It has been found that 40 % of users were only interested in services included in one descriptor. These user patterns might correspond to users who had accessed to the portal previously, and therefore, they knew where to find the desired information. Moreover, smaller groups of users, who were interested in services from more than one descriptor, have also been found. The most heterogeneous group of users is formed by users interested in services included in three different descriptors ('Shopping', 'Town Councils' and 'Public Administration').

Users interested in 'Public Administration' have been mapped close to people who accessed 'Town Councils' in both data sets taken into account (complete data set and reduced data set). Furthermore, people who were mainly interested in 'Shopping' and 'Entertainment' have been placed right on the other side of the map. Although descriptors were mapped into

different regions of the map for the two data sets which were taken into account, the neighbourhood relationships among patterns were the same. Therefore, web portal users could be divided into two large groups: users interested in information from public organizations, and users accessing the web portal for leisure items.

Our ongoing research is devoted to include the information extracted from this study in order to improve the recommender system of Infoville XXI. At present, this recommender system is based on Naïve-Bayes methodology [12]; it also has an SOM-based clustering tool, but recommendations do not use this clustering yet (so far, it is only used for consultive purposes). Two main ways of improvement based on this study can be taken into account:

- Development of a collaborative recommender based on user profiling.
- Use of GG algorithm for user profiling instead of the classical SOM algorithm, which is the algorithm currently implemented in Infoville.

Further work will be focused on analysing how users access services rather than descriptors. This analysis can provide interesting information about users interested in different services within the same descriptor.

REFERENCES

- [1] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Berlin: Springer-Verlag, 1989.
- [2] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, 2000.
- [3] A. Rauber, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarities," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR), Paris, France., 2002*.
- [4] S. Kaski, "Data exploration using self-organizing maps," *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82*, March 1997.
- [5] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen, "WEBSOM—self-organizing maps of document collections," in *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6. Espoo, Finland: Helsinki University of Technology, Neural Networks Research Centre, 1997*, pp. 310–315.
- [6] K. Smith and A. Ng, "Web page clustering using a self-organizing map of user navigation patterns," *Decision Support Systems*, vol. 35, pp. 245–256, 2003.
- [7] J. Vesanto, "SOM-based data visualization methods," *Intelligent-Data-Analysis*, vol. 3, pp. 111–26, 1999.
- [8] E. Pampalk, A. Rauber, and D. Merkl, "Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN'02)*. Madrid, Spain: Springer, August 27-30 2002.
- [9] J. Martín, "Determinación de tendencias en un portal web utilizando técnicas no supervisadas. aplicación a sistemas de recomendaciones basados en filtrado colaborativo (in Spanish)." PhD Thesis, University of Valencia, Spain, November 2004.
- [10] B. Fritzke, "Growing grid- a self-organizing network with constant neighborhood range and adaptation strength," *Neural Processing Letters*, vol. 2, no. 5, pp. 9–13, 1995.
- [11] A. Rauber, D. Merkl, and M. Dittenbach, "The growing hierarchical self-organizing map: Exploratory analysis of high dimensional data," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1331–1341, November 2002.
- [12] E. Balaguer and A. Palomares, "AI recommendation engine of Tissat, S.A." Tissat, plc." Internal Technical Report, 2003.