SELF-ORGANIZING MAPS
An Approach Applied to the Electronic Government

Everton Luiz de Almeida Gago Júnior, Gean Davis Breda, Eduardo Zanoni Marques and Leonardo de Souza Mendes
School of Electrical Engineering, University of Campinas, Campinas, SP, Brazil

Keywords: Self-organizing Maps, Data Mining, e-Gov.

Abstract: With the facilitations and results offered by automated management systems, more and more municipalities seek to eliminate physical documents by digitally storing their information. One of the direct consequences of that is the generation of a large volume of data. This paper proposes a model to support decision making based on self-organized maps. Applied to electronic government tools, this model can help identify unknown data patterns guiding the decision making process. For the accomplishment of the case study, available information from the city of Campinas, São Paulo, Brazil has been provided.

1 INTRODUCTION

Information and Communication Technology (ICT) is recognized by its potential of interactivity with users and service providers. The ICT have been broadly used in the public sector aiming to help administrators manage the resources and make feasible the monitoring of results on the implementation of public policies in the society (Mourady and Elragal, 2011). The integration and application of ICT technologies to the area of public administrators is usually known as electronic government (e-Gov) (Klischewski, 2003).

The large amount of data produced by ICT technologies might be an issue for public administrators whose intend to make decisions using information from management systems, since these data may be incomplete or also presented in an incomprehensible way. A last aggravating is the utilization of different databases for each agency in the public sector. Most of the times, these bases are built up on distinctive platforms, threatening information exchange. Hence, public organizations demand software solutions which can help the identification of possible flaws and business opportunities from smart analyses of their operational data (Yan and Guo, 2010).

Many procedures have been done in order to provide better management of public resources enabling efficient monitoring of the results on the implementation of public policies in the society. Oliveira, 2009; Braga, 2010; Mourady and Elragal, 2011 demonstrate that platforms to support public planning and decision support tools may contribute to the tributary, fiscal and economic development through the exploratory analysis of their data.

The data exploratory analysis may also be used by the public sector to self-evaluate its performance leading to better application of technological, human and financial resources, optimizing processes and speeding up the pace of administrative documents and protocols, as seen in the work of Kum (2009). The authors propose a system for knowledge discovery on self-evaluation of results by the public sector.

The studies previously mentioned use controlled vocabulary, thesauri and have limited environments able to operate only on a pre-established number of variables. The handling of these tools raises the operational cost, once it needs knowledge engineers to create ontology and analyze patterns which will be set up as prior conditions for exploratory analysis tools.

This kind of solution hampers the achievement of new knowledge from the operational data of public institutions.

1.1 Objetives

This paper proposes a Generic Model for Representation of Samples and Extraction of

Luiz de Almeida Gago Júnior E., Davis Breda G., Zanoni Marques E. and de Souza Mendes L..
SELF-ORGANIZING MAPS - An Approach Applied to the Electronic Government.
DOI: 10.5220/0003904104610470
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 461-470
ISBN: 978-989-8565-08-2

Knowledge (GMRSEK) which enables the identification of unknown patterns by mining the operational data of the public institutions. To identify unknown patterns in large volume of data, we shall use a non-supervised classification technique called self-organizing maps.

Self-organizing maps are neural networks of competitive learning. On this kind of network the processing units, called neurons, compete with one another for the right of representing an input datum. The neuron whose distance is shorter, regarding to the input datum, wins the competitive process. The winner neuron and its neighbors are adapted towards the input datum; however, these contiguous neurons are adapted with less intensity (Braga, 2010).

By using this data mining technique, it is expected to get data gatherings which show similar information between them, so that, it is possible to find classes of logs and possible patterns existing in the data. The patterns and gatherings found after exploratory analysis of the data will be treated as knowledge. The knowledge got through the exploratory analysis must be stored in the GMRSEK which provides an organized structure, enabling the generation of reports and the use of this information by electronic government systems.

2 ELECTRONIC GOVERNMENT

Many countries throughout the world stimulate reforms in the public institutions due to growing expectations of citizens, regarding to their governors. The success of public management is measured based on the benefits they assure to society. Private organizations, communities and citizens demand efficiency and accountability in public resources management, as well as, ensure the delivery of better services and results.

In this new scenario, the countries seek to revitalize their public administrations by innovating their structures and procedures, and qualifying their human resources. In this context, the utilization of Information and Communication Technologies (ICT) has a fundamental role in managing and creating an environment propitious to social and economic growth, leading to the achievement of these goals (Mourady and Elragal, 2011).

2.1 Structure of e-Gov

To establish and regulate the standards of integration and exchange of services between government, companies and citizens, it is important to define the e-Gov structure. This structure makes easy to understand the implementation process of the electronic government and the implications of this process (Ebrahim and Irani, 2005).

The generic structure of e-Gov proposed by Ebrahim and Irani (2005) is divided into four layers, as we can see in figure 1:

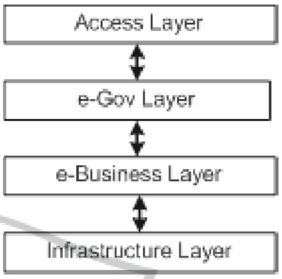

Figure 1: Structure of e-Gov (Ebrahim and Irani, 2005).

The access layer provides the means for distribution of services, products and information provided by e-Gov. These means consist of on-line access channels, such as portals that can be accessed via computer or mobile devices (Ebrahim and Irani 2005).

The e-Gov layer can be seen as a repository, where all the services offered by the government are allocated. The purpose of this layer is to establish a single entry point for users, enabling the search and utilization of services.

The e-Business layer is where the IT data an services of different agencies and public departments can be integrated. In this layer, the common data and services between different agencies and public institutions should be shared through a distributed interface, allowing the various public departments to access information from a single point (Yan and Guo, 2010).

On the other hand, the infrastructure layer concentrates the hardware solutions, which enable to provide information and services via on-line access channels.

We can list as elements of infrastructure layer the application servers, routers and other pieces of equipment which clears the way for distribution of services via Internet, Intranets and Extranets (Ebrahim and Irani, 2005).

2.2 Classification of e-Gov

e-Gov systems have broad applications and hold users of distinctive needs and profiles. In order to group the services offered to each class of users, it comes the necessity of classifying the electronic government systems. The e-Gov systems are classified as follows: *Citizen to e-Gov, Business to e-Gov, Government to e-Gov, Internal Efficiency Applications, and Effectiveness and Global Infrastructure* (Yan and Guo, 2010).

The Citizen to Government system class concentrates the services offered to the citizens. In general these services are communication channels which permit the citizen to ask the public institutions for the execution of a given task, for instance, cleaning and mowing a park, or simply issuing a form copy of a document, such as *IPTU* (tax for urban territorial property), or a debt clearance certificate (Yan and Guo, 2010).

The Business to Government system class holds part of the services offered to the companies, such as, printing forms, copying taxation documents, thus, facilitating communication between government and business. Among the services offered to entrepreneurs, it is usual, in this class of electronic government, the occurrence of electronic bids, where competitive propositions are made for getting the right of taking over a public enterprise or a service bound to be outsourced (Marques, 2010).

The Government to Government class deals with the services which must be shared between the various agencies and departments of the government itself. In general, governmental agencies and departments do not adopt a single solution for software and data storage; on the contrary, pieces of information are kept in separate and distinctive environments. In this scenario, the sharing of information and services is a challenge for the public institutions, which demand software and hardware solutions that can lead to the solution of this problem (Yan and Guo, 2010).

The class of systems called Internal Efficiency and Effectiveness deals with applications that aim to improve quality and efficiency of internal processes in governmental agencies and departments. To exemplify these applications, one can mention the work of Kum (2009), which proposes a system of knowledge discovery for self-evaluation on the results achieved from public departments and agencies, allowing to monitor the implementation results of public policies in the society.

The class of global infrastructure comprises matters concerned to interoperability of e-Gov applications, providing quality and assurance of services. The solutions employed in this class of systems put together hardware and software resources. As an example of global structure, one can mention the work of Mendes (2009), which establishes communication networks, enabling the integrations of governmental agencies and departments through the distribution of services via on-line service channels (Mendes, 2009).

3 BUSINESS INTELLIGENCE

Business Intelligence (BI) comprises a set of techniques which permit to identify behavior trends from a frame of events. These trends can help the process of decision making in business. With the fast-evolving computing sector and the enhancement of data storage mechanisms, organizations turned to store all pieces of information coming from their daily activities, such as sending protocols and documents, recording activities performed by clients, like ordering, purchasing, and so on (Mourady and Elragal, 2011).

The organizations begin to see these data as source of information that could guide their evolution and development just by utilizing the information concealed in the large volume of data stored during long periods of gathering. The growing competition between organizations and the demand for better services by clients prompted the development of more efficient techniques which permit to analyze large volumes of data in an intelligent way. The BI has emerged as a popular expression to cover these needs and is classified as systems for supporting the decision (Kum, 2009)

The large amount of data and the complexity of its relations make difficult the understanding and extraction of useful information for decision-making. Thus, there is the need of storing these data in simplified environments where the degree of relationship among the data is lower, leading to better performance in queries and cross-checking information. To meet these requirements it comes the Data Warehouse's concept (DW), which are multidimensional data bases with a lower level of standardization compared to transactional databases. In data warehouse, queries can be done more quickly and the data do not suffer from having constant modification (Kum, 2009).

Figure 2 displays the BI environment and technologies involved in it:

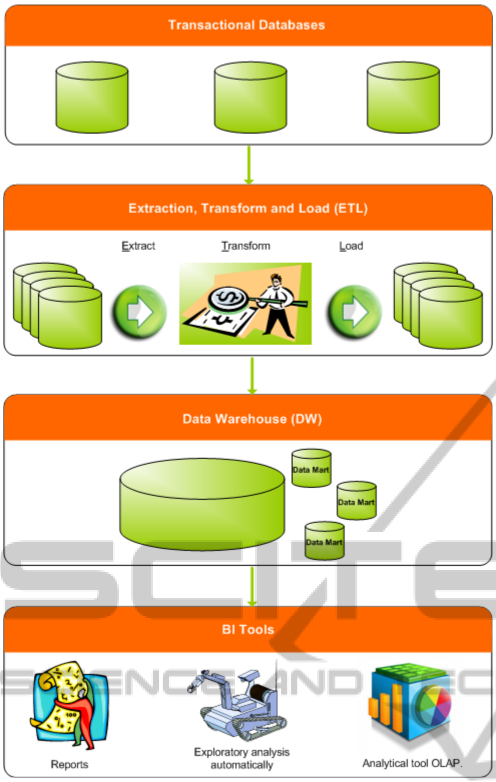

Figure 2: BI environment.

As we can see in figure 2, the DW is fed by data coming from transactional databases. The insertion of data in the DW is done by tools called Extraction, Transformation and Load – ETL. The data in the DW are in a suitable format for exploration through supporting tools for decision, without duplicities and integrated as for terminologies and formats (Mourady and Elragal, 2011).

3.1 Data Mining

Data mining is the analysis of large volumes of data in order to recognize new patterns and trends coming from information of an organization. Generally, these data are cached in transactional databases or in DW, and data mining uses techniques of pattern recognition which searches for existing similarities among the data under analysis. These patterns are characterized based on recurrent events, for instance, several people get the same disease in a given time of the year. If this event occurs again in the following years, it can be considered a pattern. Data mining can identify this kind of behavior by eliminating those less cyclical facts (Kum, 2009).

Data mining enables knowledge discovery, *i.e.*, it gets unknown information among the data. When there is no previous knowledge about the data to be mined, it is used, in general, techniques of non-supervised exploratory analysis. The self-organizing maps are examples of these techniques, not requiring any previous knowledge about the data, *i.e.,* they operate on large amount of non-classified data, of unknown types, classes and groups (Kum, 2009).

The self-organizing maps are competitive neural networks which are organized into two layers: the input-layer and the output-layer. Each neuron of the input-layer is connected to all the neurons of the output-layer through the vectors of weights (Haykin, 1999). The completion of these neural networks supposes the presence of a set of data, taken randomly and in a repetitive way in which every neuron has a weight vector associated with each input of the total of inputs. There is competition among all neurons to win the right of representing the data displayed in the network. The neuron whose vector is closer to the input datum wins the competition and gets the name of Best Matching Unit (BMU), (Haykin, 1999). The BMU neuron alters its vector of weights in order to get even closer to the displayed datum, increasing the likelihood of winning again on the occasion of appearance of the same datum. In order to identify groups, the neighbor's neurons of the winner neuron will also have their weight driven to the same input, with less intensity, though. (Haykin, 1999).

4 PROPOSED MODEL

This section presents the proposed model for application of self-organizing maps in order to identify patterns in databases of public institutions. the Generic Model for Representation of Samples and Extraction of Knowledge (GMRSEK) provides mechanisms for storing data which enables automated exploratory analysis of these data through self-organizing maps. The need for a generic environment to carry out data exploratory analysis is due to the different software solutions adopted by Brazilian municipalities. The GMRSEK is capable of storing an undetermined number of samples made up of dimensions and values that, after being submitted to the self-organizing map, results in a set of new pieces of information which, hereafter, we call knowledge.

The expected results from data mining are concentrations of data, whose meaning can be represented by the GMRSEK through a hierarchical structure. Figure 3 presents the Extraction Process of Knowledge:

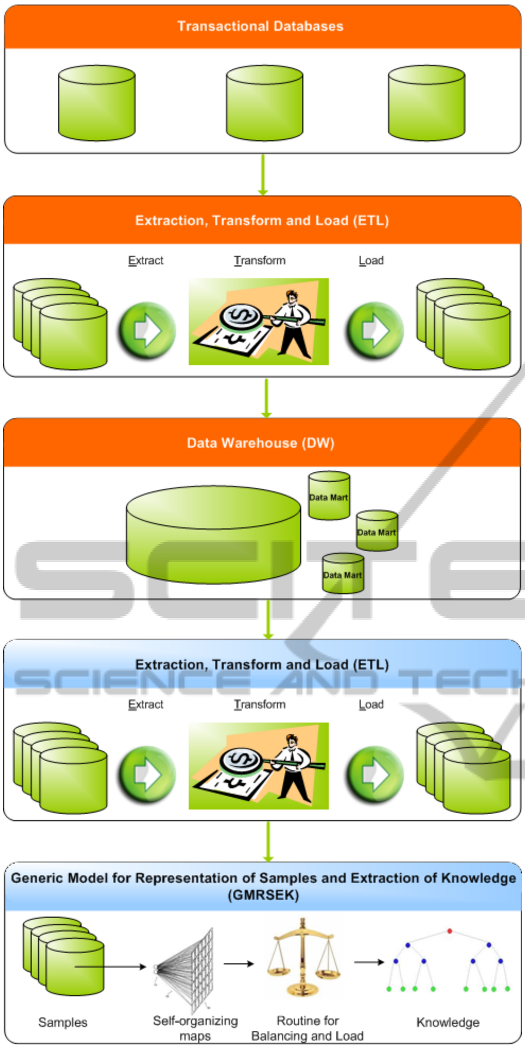
Transactional Databases
Extraction, Transform and Load (ETL)
Extract
Transform
Load
Data Warehouse (DW)
Data Mart
Data Mart
Data Mart
Extraction, Transform and Load (ETL)
Extract
Transform
Load
Generic Model for Representation of Samples and Extraction of Knowledge (GMRSEK)
Samples
Self-organizing maps
Routine for Balancing and Load
Knowledge

Figure 3: Extraction Process of knowledge.

As we can see in figure 3, data originated from transactional databases go through a process of transformation and are stored in a Multidimensional Conceptual Model (MCM). In the MCM, the data must be in an appropriate format for the application of exploratory analysis, without duplicities and integrated, as for the different terminologies existing in the transactional databases. Although data stored in the MCM are in a suitable format for exploratory analysis, public organizations, in general, have distinctive environments, turning the data mining process difficult, due to the need of integration and utilization of specific routines for data mining. In this scenario, the GMRSEK is capable of storing an undetermined number of samples coming from the MCM, and, further, submit them to a self-organizing map.

4.1 Multidimensional Model

The MCM to be used in the extraction process of knowledge as showed in figure 3 is the model proposed by Marques (2010). Marques proposes a Multidimensional Conceptual Model (MCM) of business intelligence application in electronic government. In his work, he describes the application of the MCM for analyzing operational data in the Social Assistance area. Marques' MCM comprises the integration of different tools and open sources technologies which regards from data gathering and transformation to availability of tools for end users to analyze and deal with pieces of information stored in the MCM, according to a model of intuitive use. Marques uses a structure divided into three layers: ETL layer, Storage and Availability of Data Visions, and End Users Applications Layer.

The ETL layer is responsible for the process of extraction, transformation and load of data in the repositories of operational data to the database of MCM. In this structure the ETL process is divided into sub-layers: Motor ETL and Middleware.

To implement the ETL Engine sub-layer, it was adopted a tool called Talend Open Studio, which is specialized in integration and migration of data. To choose this tool, Marques has taken into account the available documentation and the facility of providing exports routines in *.jar* extensions (Marques, 2010). Diversely, in the Middleware sub-layer it was adopted the JDBC driver. The changes applied to the data include the removal of duplicated records, integration of terminologies and values, such as, monetary values, dates and profile data, like gender, types of disabilities, race, color, and so on. Besides the mentioned changes, data originated from transactional databases undergo a structural adequacy, accommodating the pieces of information in an issue-oriented structure whose main focus is the social care carried out to citizens.

The Storage and availability of data layer is responsible for the controlling of stored data in the MCM, resulting from the ETL process performed by the previously described layer. This layer is divided into the sub-layers Physical data in the BI databases, whose function is to provide mechanisms for storing data, and Logical Layer of BI data, accountable for generating representations of data to upper layers. For storing data in the sub-layer Physical data in BI databases, Marques uses the Data Management System MySQL, for its support to the various types of indexes and its rapidness on data loading. As for the sub-layer Logical Layer of BI data, it was adopted the OLAP Mondrian server, which allows the execution of multidimensional queries on a relational database. Along with the server, the Mondrian Schema Workbench tool is released to help the multidimensional mapping of relational data, facilitating the completion of the mapping files in XML format (Marques, 2010).

The End Users Applications Layer has as its objective to provide solutions that allow users to intuitively analyze available data in BI environment through pre-defined visions. The OpenI tool has been sorted out for this purpose. The OpenI tool allows users to check BI data through a WEB application where the results are presented in form of multidimensional tables and graphs (Marques, 2010).

4.2 Generic Model for Representation of Sample and Extraction of Knowledge

The Generic Model for Representation of Samples and Extraction of Knowledge (GMRSEK) offers a centralized environment capable of storing a large volume of data, made up of an undetermined number of dimensions and values. The GMRSEK consists of a set of entities in charge of storing data which will be utilized by data mining process, and summarizes knowledge obtained through this process in a hierarchical structure, thus, providing a single and agile access for search. Figure 4 presents the GMRSEK:

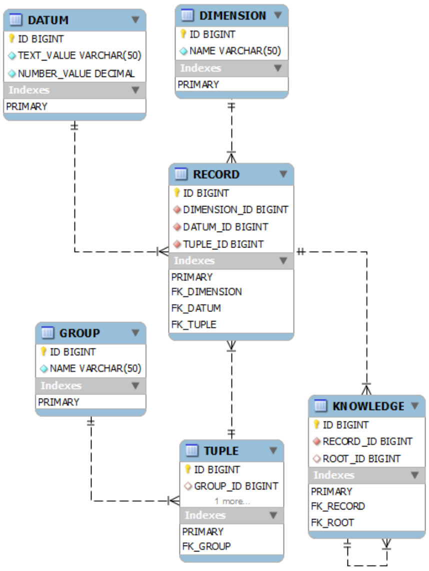

Figure 4: Generic Model for Representation of Samples and Extraction of Knowledge.

In figure 4 we can see the existence of six entities: Dimension, Record, Datum, Group, Tuple and Knowledge. The entity Dimension stores columns of a sample, identified in a single way, while the entity Datum stores different values that each dimension can take. It stores, along with the descriptive value, a numeric constant which will be used by the self-organizing map while running data mining. This numeric constant will be utilized to calculate the similarity between topological regions of the self-organizing maps and the input data.

The entity Record relates to a value of the entity Datum, where the dimension and the datum belong to the same input. Each input has a record in the entity Tuple, which relates to the entity Record, as well. The entities Dimension, Datum, Tuple and Record are Entities of Sample Representation (ESR), in charge of storing all the data to be submitted to the data mining process through the self-organizing map.

The ESRs will be fulfilled with data stored in the MCM proposed by Marques (2010). These data will be extracted from the MCM through a specific conversion routine and recorded in the GMRSEK. The extraction, Transformation and Load process (ETL) must be done through a specific routine which reads the data of the multidimensional model, changing these data to be stored in the entities of sample representation of the GMRSEK.

The option for using the MCM proposed by Marques (2010) took place due to being the data converted into a suitable format for the mining process, with a possible reduction of the number of variables, which summarize and integrate the data to be submitted to the extraction of knowledge. The dimensions of MCM proposed by Marques (2010) used by the routine of load are: gender, race, social program, location, education and disability.

The mining of data will be accomplished through a self-organizing map, due to its capability of non-supervised classification and identification of groups based on the similarities of data. The option for this technique lies on the fact of not previously knowing the data to be mined, so that it is not possible specify a set of training which comprises all the possible classes of objects existing in the data. The data stored in the ESRs must be submitted to the self-organizing map, triggering the non-supervised exploratory analysis process.

The parameterization of the neural network and choice of the self-organizing map topology comprise parameters like: initial radius of the neighborhood function, number of events for the learning process, initial value for the adaptation pace (learning rate) and the number of neurons existing in the self-organizing maps. The choice of these parameters is an empirical process whose goal is to get a point of convergence with the least possible number of neurons, thus, minimizing the processing time. The convergence point is reached when the configuration of the self-organizing map does not undergo significant changes from an event to another. This occurs because the vectors of synaptic weights reached the minimum locations of the function to be represented (Haykin, 1999).

The choice of the number of neurons is also an empirical process, so that few neurons may not represent all the groups existing in the data. On the other hand, an excessive number of neurons can be computationally costly. So, the appropriate number of neurons is the one that represents all existing groups in the data with the lowest number of units in the self-organizing map. The interpretation of results from the self-organizing map can be presented through a graphical representation by the Unified Distance Matrix (U-Matrix) and through analytical representation by assessing the relation between the records stored in the entity Knowledge, of the GMRSEK.

It follows the routine for training routine self-organizing map.

```
$w_{i,j}$ = weight.start();
r = network.getSize();
α = null;
δ = 0.9;
k = 0;
som ($V_1\{v_1, v_2 \dots v_n\}, V_2\{v_1, v_2 \dots v_n\}, V_n\{v_1, v_2 \dots v_n\}$){
   while((α=null||α<>0)&(k < 10000){
     $d_{i,j}(t) = min\sqrt{\sum |v_i - w_j|}$;
     $\alpha(t) = \exp - (|n_i - d_{i,j}|^2/2.r^2(t))$;
     $w_j(t+1) = w_j(t) + \delta(t).\alpha(t).[v_i(t) - w_j(t)]$;
       neighbors.reduce();
       knowledge.reduce();
       α = som(t) - som(t + 1);
       k = k + 1;
   }
}
```

As it can be seen in the routine previously described, the synaptic weights $w_{i,j}$, between the input layer and the network neurons are initialized at random. The neighborhood radius of neurons, represented by the variable *r*, is initially as large as the network but it is reduced in all learning iterations. The variable α stands for the difference of the map in the time status *t – 1* and when this difference is equals zero we say there was data convergence. The conditions for stopping neural network come through either data convergence or through a number k, which limits the iterations in case of no convergence.

The learning rate must be initiated by having a fixed value; in the example, the learning rate δ starts in 0.9, but must be gradually reduced as the network learning goes on. The variable $d_{i,j}(t)$ stands for the winner neuron that is the closest one to the input provided to the network. In the sequence, the neighborhood function is calculated, which affects the degree of adaptation of the neuron and its neighbors. After concluding the data mining by the self-organizing map and identification of the groups by the entity Group, the hidden information concerned to the data is already in the GMRSEK.

Although the knowledge is stored, reaching these pieces of information may be a costly task under the computational point of view, once the set of stored data in the GMRSEK may be big. There is, then, the need of a structure which leans the information, making knowledge available in an agile and unique access channel. This channel allows other applications of electronic government which makes use of knowledge achieved through the data mining process for decision-making, therefore, enhancing quality of reports and information provided to users.

As it can be seen in figure 4, the entity Knowledge has self-relationship, featuring a hierarchical structure, in such a way that enables interdependent relationship among the entities Data, Dimensions and Tuples. Hierarchical structures are known by their representative capability and access agility, however, the performance during the access to these structures is closely related to data balancing represented by them. The data must be distributed, so that, the information tree does not grow indiscriminately in just one side. If this occurs, the access performance will be like a list and not like a tree.

In the sequence, it is presented the Routine for Balancing and Load which summarizes the knowledge achieved by data mining in the hierarchical structure comprised by the entity Knowledge of the GMRSEK:

```
$g_i$ = groups.getAll();
dimensions.order();
for each $i$ of $g$ do {
  $t_k$ = $g_i$.getTuples();
  for each $k$ of $t$ do {
     $r_m$ = $t_k$.getRecords();
     integer j = 0;
     for each $m$ of $r$ do {
       if ($m < 1$) then {
         knowledge.save();
       }
       else {
         knowledge.save($r$, $r_{m-1}$);
       }
     }
  }
}
```

As it can be seen in the above routine, to load all the data in the entity Knowledge to the entities Dimension, Record, Datum, Group and Tuple, they have to be fulfilled in. The loading process of these entities starts with getting all the groups found after mining data, along with the organization of the set of samples. The dimensions of the set of samples must be organized in accordance with the number of variations of their sides, in such a way that the dimension with the lowest number of variations must be presented first. We also notice that in the first iteration of each record *r*, the entity Knowledge refers to the record *r*. In the other iterations, the entity Knowledge refers to the record *r* and to the record *r[t - 1]*, where *t – 1* stands for the previous iteration record.

5 CASE STUDY

This sections presents the results achieved from a real case study applied to data gathered from services rendered to beneficaries of social programs of *Prefeitura Municipal de Campinas*, SP, Brazil.

5.1 Operacional Data Source

The Brazilian Federal Government holds the control on the service addressed to the beneficiaries of social programs by using a data gathering tool in order to characterize the status of families called Family Development Index (FDI) (MDS, 2011). Although this data gathering tool is established by the Federal Government, public institutions look for complementary solutions which can bring bigger efficiency to the management of operational data, thus, allowing visualization of managerial reports and graphs regarding to social services. The SIGM is a good example of these solutions. It is a software focused on the need of municipal management, providing mechanism for dealing with all services, records of citizens, process management and other relevant data for municipal administration. This system is developed on the structure of multiple layers, by using the EJB technology for distributing the business objects, and managing relational database system for data storage (Marques, (2010). For managing operational data, Marques, (2010) has adopted the SIGM module for Social Management by loading in its MCM all bits of information from the SIGM transactional database. Throughout the ETL process, the data underwent a format change in order to fit the MCM. This change leads to the redistribution of information in the dimensions of multidimensional model and the elimination of duplicated records, with no integration of values, as the data come from a unique source, that is, from the SIGM transactional base.

In the MCM by Marques (2010) there are around 21,000 social care records, of which 1,621 were loaded to the ESRs of GMRSEK. Only the most consistent data records were sorted out, taking in consideration the logs of the following pieces of information: Gender, Race, Disability, Education, Attends School, Type of social benefit and metropolitan region. These dimensions were selected because they can portrait individuals and are capable of characterizing them without interfering in their privacy.

5.2 Loading Data in the GMRSEK

While loading data from the MCM to the GMRSEK, some dimensions had their values grouped into broader classes aiming to provide closer data similarity. For this reason, some different disabilities were not considered, making this dimension to be a Boolean one, that is, only saying whether the person is disabled or not. Nevertheless, the item Education had its various levels grouped into four categories: None, Low, Fair, and High. The kind of social care, likewise, had a reduction of variables, packing the different benefits into five types. They are: Income transference, Housing benefit, Social-educative benefit, Child and Teen Care and Youth-addressed Programs. The reduction for these variables was necessary to bring bigger similarity in the data set, once the similar social programs benefit citizens with the same features.

5.3 Data Mining

During the exploratory analysis carried out through the self-organizing map, it was possible to notice that the free parameters of the neural network and the choice of the number of neurons have directly influenced the convergence of results, sometimes, even the results themselves. Initially, the self-organizing map had been defined with many neurons, having 841 processing units, that is, a grid of 29 x 29 neurons. This is a generous estimative, taking in consideration the number of available data. It took the exploratory analysis 15 hours to reach the convergence point, and, in the end, it was possible to identify the existence of five groups of data. By decreasing the number of neurons to 64 units, the convergence time has dropped to approximately 90 minutes; however, only four groups have come to evidence, being the fifth one embodied to the others. In the trial of establishing an intermediate value, a grid of 100 processing units was then defined. In this last configuration, it was possible to obtain the same five groups resulting from the first execution with fewer processing units and to shorten the convergence time to around 12 hours.

For achieving convergence of results, the neural network had to go through several learning events. One could notice that when modifying the pace of adaptation of the neural network, the number of necessary events to convergence had been different. Taking the 100-neuron grid, with the learning pace starting at 0.9, it was necessary about 2300 events to have the occurrence of convergence of results. Nevertheless, by initiating at 0.3, it was necessary around 1000 events for having convergence of results. A third trial was carried out with the same 100-neuron grid, but with the adaptation pace at 0.6. This way the convergence point was reached with about 700 events. Figure 5 shows the U-Matrix which illustrate the 100 neurons of the self-organizing map and the groups found:

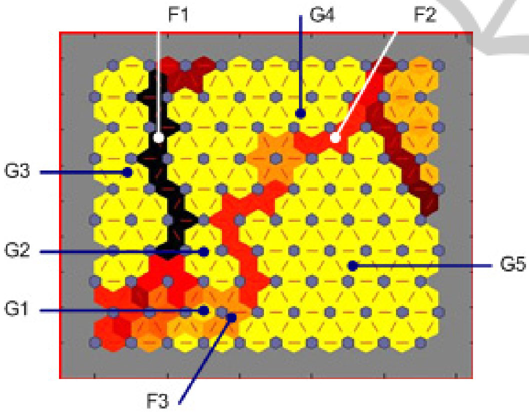

Figure 5: U-Matrix.

Figure 5 is a graphical representation of the groups found by the data mining process through the U-Matrix. In it, we can notice the division of the input data into five different groups. Between one group and another there is a delimitation done by frontiers of bigger or smaller intensity. The darkest frontiers *F1* and *F2* are those representing smaller similarity among the groups. This way the clearest frontier *F3* represents bigger similarity among the groups. It is possible to notice the existence of a small group *G1* made up of just one hexagon, surrounded by clearer frontiers, at the bottom left of the image. This group has little difference from *G5* and *G2* groups. It is important to highlight that there is smaller similarity between *G5* and *G2* groups, bordered by the darkest frontier *F2* between them. Other two groups *G3* and *G4* can be seen on the top of figure 5. The frontier *F1* between them shows that there is little similarity between the two groups.

5.4 Balancing and Load

After the end of the automatic exploratory analysis carried out by the self-organized map, the five groups found had already been identified in the entity Group of the GMRSEK. Although the groups were associated with the tuples, that is, to the inputs which generated them, the big volume of data made difficult the understanding of results in an analytical way. The results, in their analytical form, were better understood after having summarized the knowledge, representing the results through a hierarchical structure provided by the entity Knowledge of the GMRSEK. This was possible thanks to Balancing and Load Routine. The data summarized by the balancing and load routine show that people who claim for programs addressed to the young public are generally of female gender.

The young female, deficiency holders, have a high level of education and are no longer attending school and live in the east zone of town. Nevertheless, the young female, non-deficiency holders, have medium education level and are still attending school, usually downtown, as shown in figure 6:

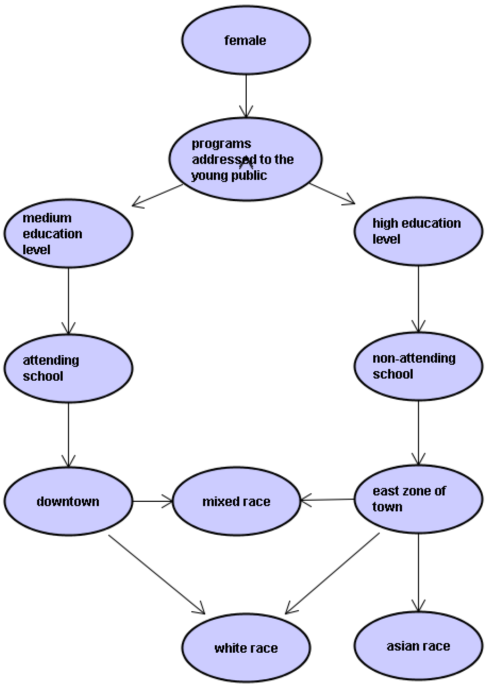

Figure 6: Descriptive representation of groups.

This piece of information may be useful for the municipal institutions, as it allows the actions to be taken in order to improve services to these beneficaries. The case study has shown that, in the east zone of town, social attendance points must provide accessibility to disabled people beyond forwarding the most qualified beneficiaries to the job market, assuring that other people can be helped in this region.

6 CONCLUSIONS

The U-Matrix was efficient to graphically represent the groups found, bearing a big reduction of dimensionality without having information loss. Although data may be represented by the U-Matrix, this representation only shows the topological display of groups, not allowing the end user to analytically visualize the results achieved.

The Generic Model for Representation of Samples and Extraction of Knowledge (GMRSEK) has shown itself as flexible and capable of operating on a variable number of analysis dimensions with multiple values associated with them. Through the entity Knowledge, the result achieved with the data mining process can be represented under a hierarchical form, allowing the decision-maker to have a good perception of existing patterns about the data.

The representation of groups by a hierarchical structure permits users to visualize the data groups in an analytical way, enabling the understanding of the patterns. The utilization of a hierarchical structure for the representation of knowledge has ended up having a good performance during the queries carried out; however, the balancing and load process of data in this structure is costly under the computational point of view. The data mining, as well as, the balancing and load routine for the representation of the knowledge are periodically performed, being more often the utilization of this model for asking queries and issuing reports.

The MCM proposed by Marques (2010) has brought more quality to the knowledge discovery process, once the data had already gone through treatments, such as duplicity removal, uniformity and grouping of values, beyond bearing totals in their table of facts, thus, enhancing the integrity of samples in the GMRSEK.

Future work may use the knowledge stored in GMRSEK to classify new records, without a new mining of data types. A supervised learning mechanism can be training with the knowledge stored in Knowledge organization, and from then classify new data registered by ICT technologies.

REFERENCES

Braga, C. V., 2010. Rede Neural e Regressão Linear: Comparativo entre Técnicas Aplicadas a um Caso Prático na Receita Federal. Dissertação de Mestrado. Faculdade de Economia e Finanças IBMEC.

Ebrahim, Z., Irani, Z., 2005. E-government adoption: structure and barriers, Business Process Management Journal, v. 11 n. 5.

Haykin, S. A Comprehensive Fundation, 1999. McMaster University Hamilton, Ontario, Canada, Pearson Education, 1999.

Klischewski, R., 2003. Semantic Web for e-Government, Springer Berlin.

Kum, H., Duncan, D. F., 2009. Stewart, C. J.: Supporting self-evaluation in local government via Knowledge Discovery and Data Mining. Government Information Quarterly.

Marques, E. Z., Miani, R. S., Gago Junior, E. L. A., Mendes, L. S. Development of a Business Intelligence Environment for e-Gov Using Open Source Technologies. In: Data Warehousing and Knowledge Discovery, 2010, Bilbao. Lecture Notes in Computer Science. Berlim: Springer, 2010. v. 6263. p. 203-214.

Mendes, L. S., Bottoli, M. L., Breda, G. D., 2009. Digital cities and open MANs: A new communications paradigm, LATINCOM'09. IEEE Latin-American Conference on Communications.

Ministério do Desenvolvimento Social http://www.mds.gov.br/programabolsafamilia/noticias/aplicativo-do-indice-de-desenvolvimento-da-familia-ja-esta-disponivel/, 2011.

Mourady, A., Elragal, A., 2011. Business Intelligence in Support of eGov Healthcare Decisions. European, Mediterraneam & Middle Eastern Conference on Information System, Athens, Greece.

Oliveira, T. P. S., 2009. Sistemas Baseados em Conhecimento e Ferramentas Colaborativas para Gestão Pública: Uma proposta ao Planejamento Público Local.

Yan, P., Guo, J., 2010. Researching and Designing the Structure of E-government Based on SOA, Proceedings of the 2010 International Conference on E-Business and E-Government.