Explaining Diverse Application Domains Analyzed from Data Mining Perspective

Alberto Ochoa, Lourdes Margain, Rubén Jaramillo,
Javier González, Daniel Azpeitia, Claudia Gómez,
Jöns Sánchez, Julio Ponce, Sayuri Quezada,
Francisco Ornelas, Arturo Elías, Edgar Conde, Víctor Cruz,
Petra Salazar, Emmanuel García and Miguel Maldonado

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/49947

1. Introduction

This chapter proposal explains the importance of adequate diverses application domains in different aspects in a wide variety of activities of our daily life. We focus our analysis to different activities that use social richness data, analyzing societies to improve diverse situational activities based on a decision support systems under uncertaainty. To this end, we performed surveys to gathering information about salient aspects of modernization and combined them using social data mining techniques to profile a number of behavioural patterns and choices that describe social networking behaviours in these societies.

We will define the terms “Data Mining” and “Decision Support System” as well as their contrast and roles in modern societies. Then we will describe innovative models that captures salient variables of modernization, and how these variables give raise to intervening aspects that end up shaping behavioural patterns in social aspects. We will describe the data mining methodologies we used to extract these variables in each one of these diverse application domains including the analysis of diverse surveys conducted in diverse societies, and provide a comparative analysis of the results in light of the proposed innovative social model.

On the rest proposed chapter, we will describe how our model can be extended to provide a means for identifying potential social public politics. More particularly, we make allusion to behavioural pattern recognition mechanisms that would identify the importance of use techniques from Data Mining. We will close with concluding remarks and extended discussions of our approach and will provide general guidelines for future work in the area

INTECH
open science | open minds

of application of Data Mining in diverse application domains, including further analysis on how those public politics organize and operate in social rings, and how they use technology to that end. While our main focus will be on pure social networks such as Facebook. Our literature review will include cases of implementation of correct public politics, and some issues, challenges, opportunities, and trends about this diverses social problems.

The proposal of this chapter is to explain the importance of use Social Data Mining in a wide variety of activities in our daily life, many of these activities, which are online and involve many social networkings using in many ways using Media Richness. Social Data Mining techniques will be useful for answering diverse queries after gathering general information about this given topic. This kind of behaviors will be characterized by take a real implementation of a correct solution, each one of these taking diverse models or multi agents systems for adequate this behavior to obtain information to take decisions that try to improve aspects very important of their lifes organized in different application and fields of knowledge.

First, in section 1 of this Social Data Mining techniques will be useful for answering diverse questions after gathering general information about the given topic. This type of behaviors will be characterized by a real application of a correct solution, each one of these taking diverse models or multi agents systems. This is for adequate this behavior to have information and make decisions that try to improve aspects very important of their lifes organized in different application and fields of knowledge.

First, in section 1 of this chapter explain the concept of Social Data Mining and as this behavior affect in different way to people in differnet aspects in societies´ people –Viral Marketing to determine boughts on inmobilarie sector–. In other sections we explain the way to generate a correct analysis In correspondent sections we explain the way to make a correct analysis of diferent activities of daily life as in Electrical Industry (section 2), Classification of Images and its analysis which explain the effects in their analysis including Medical advances (section 3), a comparative analysis using people profile according to describe a possible social benefits in diverse applications domains (section 4). In section 5 we explain the results obtained in e-commerce data mining and emergent kind of techniques which resolve and propose specific kind of marketing according at life style of consumers, and in the section 6 are try to describe the use of Data Mining to Mobile Ad Hoc Networks Security which will be used to determine the possible changes on our modern society. In section 7 we described another specfic applications domains as: organizational models, organizational climate, zoo applications to classify more vulnerable species or identify the adequate kind of avatars on a roll multigame players and finally our conclusions about the future of Data Mining in diverses uses to different activities of our daily life.

2. Data mining and their use on viral marketing

The use of traditional media like radio, television and newspaper, has been replaced by new digital media like social networks Facebook and Twitter. According to (Salaverría, 2009) the increase of broadband users, both on mobile devices, home and workplace, has raised the replacement of traditional media by digital media. Likewise, mentions that these new tools allow the user to interact with the issuer, thanks to several factors that facilitate interaction such as, frequency of updates, including multimedia such as videos and photographs, among others.

On the other hand, (Orihuela, 2002) mentions that existing Internet interactivity has been subverting the paradigms within communication processes in mass media. As (Salaverría, 2009), mentions the ability of interactivity, customization and upgrade, as central in replacing traditional media to digital (Figure 1).

Figure 1. Interaction between users of social networks.

In their study, (Orihuela, 2002), concludes that the public announcement raised in the new digital media is sufficient justification to redefine the requirements in the media, the procedures and content of information, all within trends changing as a result of network usage.

Due to the above, the social networks like Facebook, are an important tool in the marketing strategy of companies. Its low cost (sometimes zero) helps not only in communicating the customer value, but also improves the customer-consumer relationships. According to (Orihuela, 2002), social networks like Facebook sometimes take characteristics of traditional media, however, incorporate a higher level of interaction.

2.1. Corporate use of social networks

After analyzing the above, we can say that the use of social networks helps greatly reducing advertising costs and implementation of new marketing strategies. But even if there are different tools to monitor and observe the behavior of users, there is little research evidence that reveals different patterns of consumption, transmission of messages or lack of them, and observes the behavior of these consumers on trademarks and their experiences with them within the social networks like Facebook.

According to (Salaverría, 2009) the online advertising industry grew by 800 percent from 2004 to 2009 demonstrating a steady development in which social networks and contextual advertising play an important role in the marketing or advertising on social networks, without But there is no scientific evidence on the behavior of users in such networks and the dissemination of messages received and sent within these networks and what encourages you to do or not.

(Sandoval et al., 2010) mentions that social networks have changed the human relations approach and have potentiating its most important feature: Easy to find and develop relationships with other members with similar interests. Similarly mention that social networking services have proliferated targeting people in specific regions or some similar interests as, ethnic, religious, sexual and political (Figure 5). Thus, the fact of having a community segment showing a potential interest in a particular company or product, is useful when performing a specific marketing strategy. In addition to marketing strategies, companies can use such networks in the recruitment, internal communication and interaction with consumers.

Figure 2. Nested groups of similar interests.

2.2. Research objectives

Having analyzed the use of social networks in business, the importance of the restaurant industry in Mexico and specifically the problem of insecurity in Juarez, perform the following research questions:

- What specific objectives seek restaurant sector companies to use social networks?
- What digital social network use most frequently?
- What percentage of these companies has replaced the traditional media advertising advertising on social networks?
- What marketing strategies used in online social networks?
- What correlation is there between; use of social networks and increased sales?
- When beginning their presence within social networks?
- How many users is made up your network?
- How often publish information within social networks?
- What correlation exists between the periodicity of the publications and the time spent in the network, with the number of users in the network?

2.3. Methodology

The conclusive results of this research were obtained through an exploratory study of the use of social networks in companies in the restaurant industry in Juarez and factorial designs were performed to find some correlations between different variables.

First we made a query of the restaurant industry to recognize his presence in Mexico and in the locality. This was done through the National Chamber of the Restaurant Industry and Seasoned Foods (CANIRAC) and National Chamber of Commerce (CANACO) found in the locality.

From the list of registered companies in the industry by these cameras restaurateur, was searched to select those that were present within the digital social networks, regardless of upgrade or number of users connected to their groups. Were interviewed and application of survey of 20 companies with the largest number of users within your network, to meet their business openly in online social networks, specifically Facebook. Took place through a careful study of social networks to find that participation in that network have to know the frequency and topics of their publications, as well as general information of relevance to publish within their Facebook page. He knew the date they started their activities in the network.

Once the information was held after his capture to be analyzed in statistical software to find relevant values.

3. Competitive learning for self organizing maps used in classification of partial discharge

Competitive learning is an efficient tool for Self Organizing Maps, widely applied in variety of signal processing problems such as classification, data compression, in anothers. With the huge volumes of data being generated from the different systems everyday, what makes a system intelligent is its ability to analyze the data for efficient decision-making based on known or new cluster discovery. **The partial discharge (PD) is a common phenomenon which occurs in insulation of high voltage, this definition is given in [1]. In general, the partial discharges are in consequence of local stress in the insulation or on the surface of the insulation**. We evaluate the performance of algorithms in which competitive learning is applied of partial discharge dataset, quantization error, topological error and time in seconds per training epoch. The result from classification of PD shows that *Winner-takes-all* **(WTA)** has better performance than *Frequency Sensitive Competitive Learning* **(FSCL)** and *Rival Penalized Competitive Learning* **(RPCL).** The first approach in a diagnosis is selecting the different features to classify measured PD activities into underlying insulation defects or source that generate PD's (Figure 3).

The phase resolved analysis investigates the PD pattern in relation to the variable frequency AC cycle (Cheng et al., 2008). The voltage phase angle is divided into small equal windows. The analysis aims to calculate the integrated parameters for each phase window and to plot them against the phase position (ϕ).

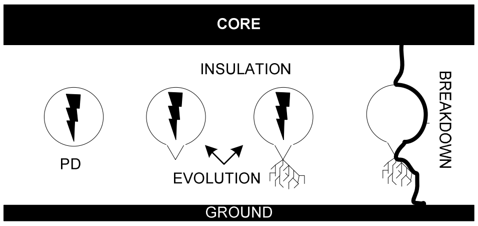


Figure 3. Example of damage in polymeric power cable from the PD in a cavity to breakdown.

- (q_m −ϕ) : the peak discharge magnitude for each phase window plotted against ϕ, where q_m is peak discharge magnitude.

3.1. Self organizing map

The Self Organizing Map developed by Kohonen, is the most popular neural network models (Kohonen, T., 2006 & Rubio-Sánchez, M., 2004). The SOM is a neural network model that implements a characteristics non-linear mapping from the high-dimensional space of input signal onto a typically 2-dimensional grid of neurons. The SOM is a two-layer neural network that consists of an input layer in a line and an output layer constructed of neurons in a two-dimensional grid.

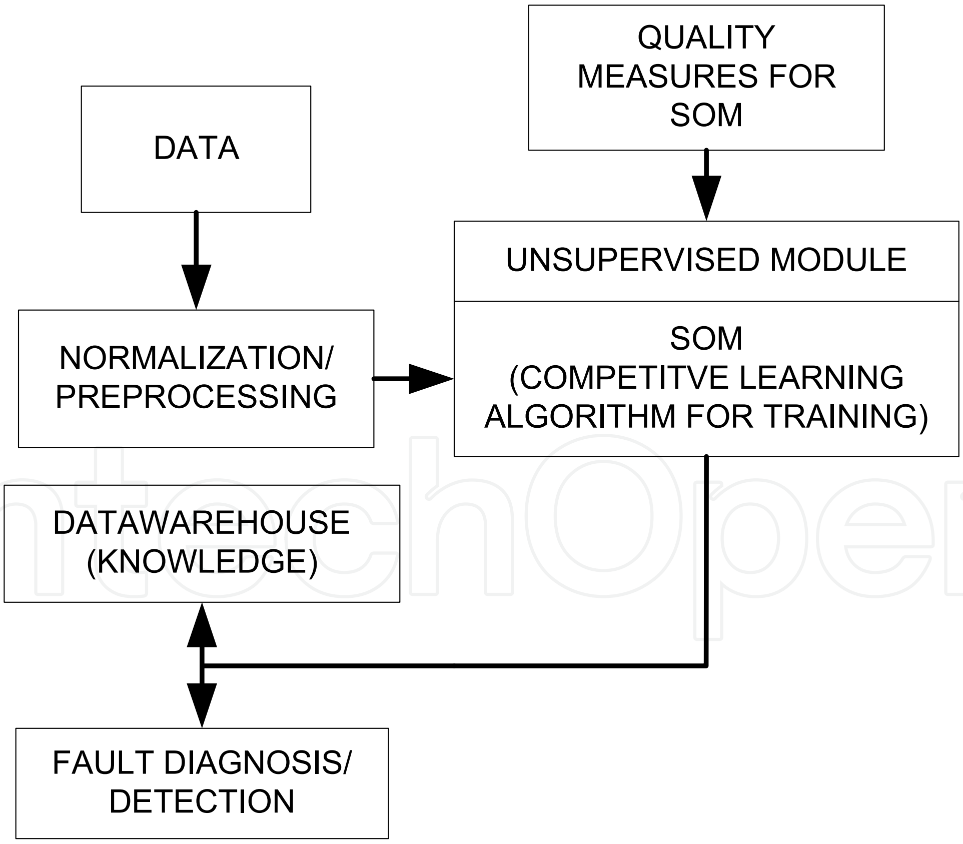

QUALITY MEASURES FOR SOM
DATA
UNSUPERVISED MODULE
NORMALIZATION/ PREPROCESSING
SOM (COMPETITVE LEARNING ALGORITHM FOR TRAINING)
DATAWAREHOUSE (KNOWLEDGE)
FAULT DIAGNOSIS/ DETECTION

Figure 4. The component interaction between SOM.

PD measurements for power cables are generated and recorded through laboratory tests. Corona was produced with a point to hemisphere configuration: needle at high voltage and hemispherical cup at ground. Surface discharge XLPE cable with no stress relief termination applied to the two ends. High voltage was applied to the cable inner conductor and the cable sheath was grounded, this produces discharges along the outer insulation surface at the cable ends. Internal discharge was used a power cable with a fault due to electrical treeing. Were considered the pattern characteristic of univariate phase-resolved distributions as inputs, the magnitude of PD is the most important input as it shows the level of danger, for this reason the input in the SOM the raw data is the peak discharge magnitude for each phase window plotted against ($qm - \phi$). **Figure 2** shows the conceptual diagram training. In the cases analyzed, the original dataset is 1 million of items, was used a neurons array of 10×10 cells to extract features. As it is well known, in fact, a too small number of neurons per class could be not sufficient to represent the variability of the samples to be classified, while a too large number in general makes the net too much specialized on the samples belonging to the training set and consequently reduces its generalization capability. Moreover a too large number of neuron per class implies a long training time and a possible underutilization of some of the neural units. PD patterns recognition and classification require an understanding of the traits commonly associated with the different source and relationship between observed PD activity and responsible defect sources. This paper shows the performance of SOM using different competitive learning algorithms to classify measured PD activities into underlying insulation defects or source that generate PD's, its showed that WTA is the better algorithm with less error and training time, but its overall performance are not always satisfactory, being alternative in accord at the performance FSCL or RPCL algorithms.

4. Classification of images using Naive Bayes and J48

This research work's approach is related to artificial vision due to extraction from information contained in images (human faces) by using methods to obtain RGB coloration and statistic values. Extraction takes place by performing several tests of image splitting into different sizes, later classifying sets of instances with data mining techniques, and analyzing classification results to determine which of the algorithms is the best for this particular case.

Knowledge database contains 100 images built from 20 people and 5 pictures each. Mean and standard deviation were employed as statistic values, which are also used as attributes of the instances classified by Naïve Bayes and WEKA J48. It is important to mention that no pixel is disregarded to obtain instances, both of the pixel groups the ones inside face and outside of it are considered. Impact of splitting images into parts.

Table 1 shows that splitting images into both 16 and 64 obtain the same amount of correctly classified instances except with NaiveBayes classifier under cross validation where splitting into 16 obtains 3% better of correctly classified instances, partial conclusion from this table is splitting images into both 16 and 64 is better than splitting them into 4 and no splitting them.

Table 2 shows both Naïve Bayes and J48 under use training test option obtain 100 % of correctly classified instances from splitting images into 4 parts which reveals splitting images helps for classification process.

		Parts			
Classifier	**Test options**	1	4	16	64
NaiveBayes	use training set	94%	100%	100%	100%
J48	use training set	97%	100%	100%	100%
NaiveBayes	cross validation folds 10	85%	90%	97%	94%
J48	cross validation folds 10	71%	86%	85%	85%
NaiveBayes	percentage split 66%	76.4706%	70.5882%	88.2353%	88.2353%
J48	percentage split 66%	64.7059%	55.8824%	64.7059%	64.7059%

Table 1. Results of splitting images into parts including all attributes.

		Parts			
Classifier	**Test options**	1	4	16	64
NaiveBayes	use training set	90%	100%	100%	100%
J48	use training set	92%	100%	100%	100%
NaiveBayes	cross validation folds 10	72%	93%	97%	96%
J48	cross validation folds 10	69%	83%	84%	89%
NaiveBayes	percentage split 66%	58.8235%	88.2353%	94.1176%	94.1176%
J48	percentage split 66%	61.7647%	70.5882%	73.5294%	76.4706%

Table 2. Results of splitting images into parts including best 30% attributes.

4.1. Impact of attribute selection

Table 3 shows that for any test option, both classifiers obtain greater amount of correctly classified instances considering all of the attributes which are 6.

Table 4 shows that both classifiers obtain 100% of correctly classified instances under use training set test option. Under cross validation Naïve Bayes classifies 3% better selecting 8 attributes and J48 classifies 3% considering all of the attributes. Finally under percentage split both classifiers perform better selecting 8 attributes.

		6 attributes	**2 attributes**
Classifier	Test options		
NaiveBayes	use training set	94%	90%
J48	use training set	97%	92%
NaiveBayes	cross validation folds 10	85%	72%
J48	cross validation folds 10	71%	69%
NaiveBayes	percentage split 66%	76.4706%	58.8235%
J48	percentage split 66%	64.7059%	61.7647%

Table 3. Results of attribute selection without splitting images.

		24 attributes	**8 attributes**
Classifier	Test options		
NaiveBayes	use training set	100%	100%
J48	use training set	100%	100%
NaiveBayes	cross validation folds 10	90%	93%
J48	cross validation folds 10	86%	83%
NaiveBayes	percentage split 66%	70.5882%	88.2353%
J48	percentage split 66%	55.8824%	70.5882%

Table 4. Results of attribute selection splitting images into 4 parts.

Table 5 also shows a 100% of correctly classified instances for both classifiers under use training set test option. Under cross validation , Naïve Bayes classifies equal amount of correctly classified instances selecting 29 attributes as selecting all of them, similar situation occurred with J48 with 1% greater for selecting all of the attributes. Finally, under percentage split both of the classifiers perform better selecting 29 attributes.

		96 attributes	**29 attributes**
Classifier	Test options		
NaiveBayes	use training set	100%	100%
J48	use training set	100%	100%
NaiveBayes	cross validation folds 10	97%	97%
J48	cross validation folds 10	85%	84%
NaiveBayes	percentage split 66%	88.2353%	94.1176%
J48	percentage split 66%	64.7059%	73.5294%

Table 5. Results of attribute selection splitting images into 16 parts.

Table 6 shows once again a 100 % of correctly classified instances under use training set test option for both cases of attribute selection. Under cross validation and percentage split both of the classifiers perform better selecting 116 attributes.

		384 attributes	**116 attributes**
Classifier	Test options		
NaiveBayes	use training set	100%	100%
J48	use training set	100%	100%
NaiveBayes	cross validation folds 10	94%	96%
J48	cross validation folds 10	85%	89%
NaiveBayes	percentage split 66%	88.2353%	94.1176%
J48	percentage split 66%	64.7059%	76.4706%

Table 6. Results of attribute selection splitting images into 64 parts.

4.2. Analysis of classifiers effectiveness based on test options

Table 7 shows that J48 performs better than Naïve Bayes without splitting images but not in a significant way. Considering any other splitting image scheme or attribute selection show a 100 % of correctly classified instances.

		Classifiers	
Attributes	Parts	NaiveBayes	J48
All of them	1	94%	97%
30%	1	90%	92%
All of them	4	100%	100%
30%	4	100%	100%
All of them	16	100%	100%
30%	16	100%	100%
All of them	64	100%	100%
30%	64	100%	100%

Table 7. Results of classifiers effectiveness under use training set.

Table 8 shows Naive Bayes performs better than J48 for any splitting image scheme and attribute selection.

		Classifiers	
Attributes	Parts	NaiveBayes	J48
All of them	1	85%	71%
30%	1	72%	69%
All of them	4	90%	86%
30%	4	93%	83%
All of them	16	97%	85%
30%	16	97%	84%
All of them	64	94%	85%
30%	64	96%	89%

Table 8. Results of classifiers effectiveness under cross validation.

Table 9 shows Naïve Bayes performs better than J48 except for selecting best 30% attributes without splitting images. Experiments of splitting images into parts allow concluding that splitting images into 16 parts is enough for satisfactory classification. Statement from previous paragraph can be asserted due to results in Table 1 show splitting images into 64 parts obtains equal amount of correctly classified instances as performing such split into 16 parts, Table 1 even shows a reduction of 3% in correctly classified instances for NaiveBayes classifier under cross validation. Next stage of experiment consisted on selecting best 30%

attributes, which reveals Naïve Bayes generates greater amount of correctly classified instances from splitting images into 16 parts, J48 obtains 5% better in 64 parts under cross validation and 2.9412% in 64 parts under percentage split. Due to improvement for 64 parts is not significant, it is concluded splitting into 16 parts is enough. Experiments of attribute selection allow concluding that selecting best 30% is enough. This can be validated from both table 1 and table 2 which show that splitting images into 64, 16, and 4 parts selecting best 30% obtains greater amount of correctly classified instances than considering all attributes. J48 throws 1% better for splitting into 16 parts and 3% better into 64 parts with all attributes, but this is disregarded due to it is not significant. Experiments analyzing effectiveness of classifiers allow to conclude Naïve Bayes performs better due to it obtains greater amount of correctly classified instances under most splitting images case and test option except for use training set test option and no splitting images.

		Classifiers	
Attributes	Parts	NaiveBayes	J48
All of them	1	76.4706%	64.7059%
30%	1	58.8235%	61.6747%
All of them	4	70.8552%	55.8824%
30%	4	88.2353%	70.5882%
All of them	16	88.2353%	64.7059%
30%	16	94.1176%	73.5294%
All of them	64	88.2353%	64.7059%
30%	64	94.1176%	76.4706%

Table 9. Results of classifiers effectiveness under percentage split.

4.3. Medical visualization in data mining

A field that is becoming a rich area for the application of data mining is that of medical imaging. The tremendous advance in imaging technologies such as X-rays, computed tomography, magnetic resonance, ultrasound and positron emission tomography has led to the generation of vast amounts of data (Figure 5). Scientists are interested, of course, in learning from this data, and data mining techniques are increasingly being applied in these analyses.

There are interesting techniques for finding and describing structural patterns in data as a tool for helping to explain that data and make predictions from it. The data will take the form of a set of examples from the patients. The output takes the form of predictions about new examples. Many learning techniques look for structural descriptions of what is learned, descriptions that can become fairly complex and are typically expressed as sets of rules. Because they can be understood by people, these descriptions serve to explain what has been learned and explain the basis for new predictions. People frequently use data mining to gain knowledge, not just predictions. Databases are rich with hidden information that can be used for intelligent decision making. Classification and predictions are two of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large.

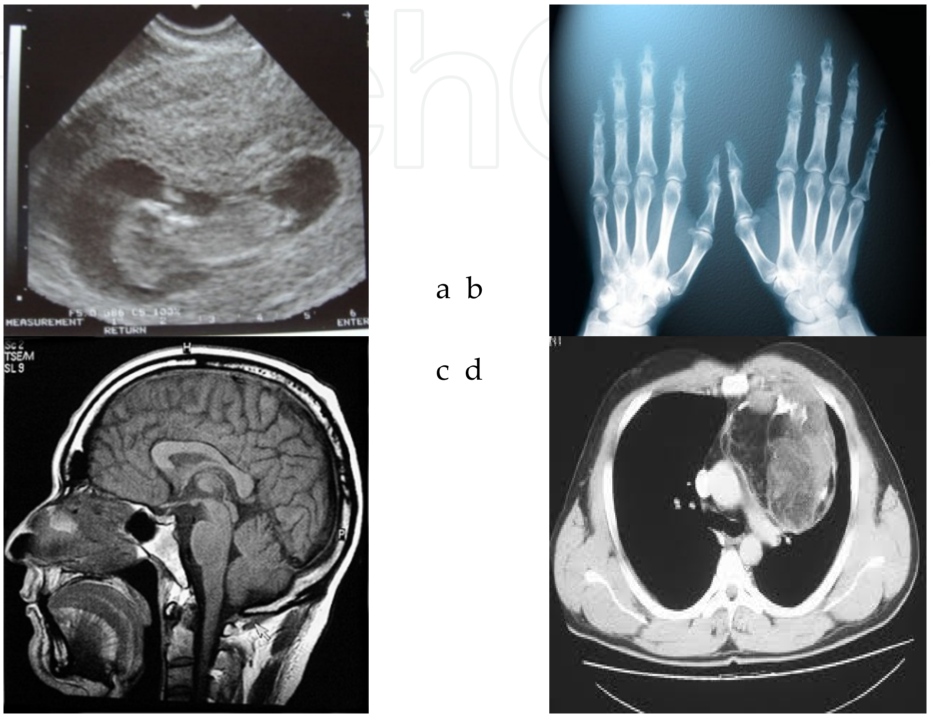


Figure 5. Examples of medical imaging. (a) Ultrasound. (b) A-rays. (c) Magnetic resonance. (d) Computed tomography.

4.3.1. Classification and prediction

A medical research wants to analyze breast cancer data in order to predict which one of three specific treatments a patient should receive. In the example, the data analysis task is classification, where a model o classifier is constructed to predict categorical labels, such as treatment A, treatment B, or treatment C for the medical data. These categories can be represented by discrete values, for example, the values 1, 2, and 3 may be used to represent treatment A, B, and C.

The implementation methods discussed are particularly oriented toward show different tools for analyzes medical data.

4.3.2. Classification by decision tree induction

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flow-chart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each brand represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. The learning of decision trees from class-labeled training tuples is named decision tree induction. A decision tree can be viewed as a flow-chart-like tree structure, where each internal node (nonleaf node) represents a test on an attribute, each brand represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The root node is the principal node (highest node) in a tree.A typical decision tree is shown below (Figure 6).

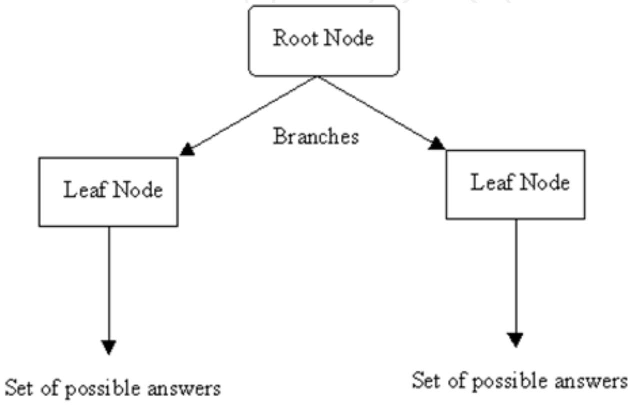


Figure 6. Decision Tree Induction

ID3 is an algorithm which generates a decision tree based on input data by looking at the amount of information contents contained in the various input attributes. At each step in the decision tree, it chooses the attribute which provides the biggest information gain and uses that attribute to classify data further. Its pseudo code is summarized as follows:

```
Input: Set S of positive and negative examples, Set F of features
ID3( F, S )
1. if S contains only positive examples, return “yes”
2. if S contains only negative examples, return “no”
3. else
    choose best feature f in F which maximizes the information gain
    for each value v of f do
    add arc to tree with label v, along with the sub tree for that new branch
```

Like an example, the input and output variables and their domains are specified in the list below:

1. Input variables (from clinical observations):
    a. Extent (Size of Spreading): {E1, E2, E3, E4}
    b. Hypoxia: {H1, H2}
    c. Surface (surface marker): {S1, S2, S3}
    d. LOH: {M1, M2, M3}
2. Final result/outcome:

    Outcome: {P (progressed to cancer), NP (didn’t progress to cancer)}

The ID3 algorithm as implemented and the following decision tree are generated (Figure 7):

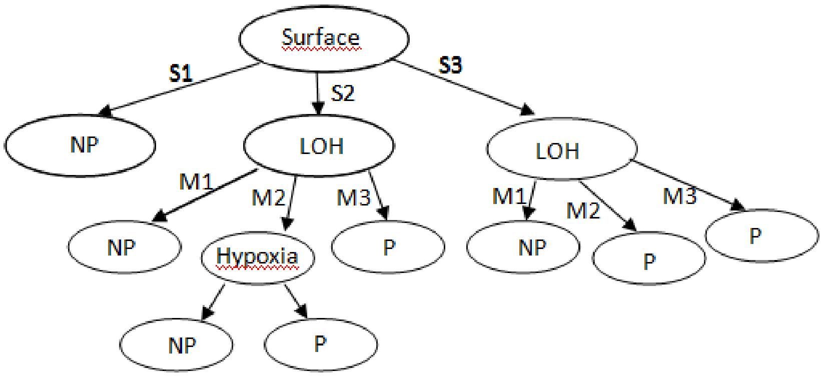


Figure 7. Decision Tree Induction for the medical data.

The decision tree method produces a reasonably good estimate on the outcome based on the inputs. This estimate is about 92% accurate, which is above the acceptable level of accuracy as proposed by the clinical researchers.

4.3.3. Classification by Back-propagation

An artificial neural network (ANN) is a computational model that is inspired by the structure and functional aspects of biological neural networks. They are usually used to model complex relationships between inputs and outputs and find patterns in data. In other words, we wish to infer the mapping implied by the data. The cost function is related to the mismatch between our mapping and the desired outcome. One very commonly used approach to train neural network from input examples is the back-propagation algorithm. Back-propagation algorithm is a common supervised-learning method that teaches an artificial neural network on how to perform a given task. The neural network is modeled as a set of neurons which take inputs, apply certain weights to each input and propagate the result forward into the next layer of units. Each unit in a particular layer is essentially a linear function of the input units from its previous layer. Eventually, the data gets propagated into the output layer where the results are presented.

Another important aspect is this algorithm is able to learn by propagating the errors in the output layer backwards into the inner layers by adjusting the weights between the input and hidden layer and between hidden and output layer in order to reduce the error on the output. The algorithm continues to do this until either the maximum number of epochs is reached or the errors at the output are within an acceptable range. This technique is also referred to as "back-propagation", as denoted in its name. A very typical neural network consists of 3 layers – input, hidden, and output layer. In practice, it is possible to have more than one hidden layers. The back-propagation algorithm used for this project is based on such a 3-layer neural network as illustrated in the figure 8.

The pseudo code for the back-propagation algorithm is as follows:

```
Initialize the weights in the network (randomly between -0.5 and 0.5)
Do
For each example e in the training set
```

```
O = neural-network-output(network, e) ; forward pass
T = desired output for e
Calculate error (T - O) at the output units
Compute delta_wh for all weights from hidden layer to output layer
Compute delta_wi for all weights from input layer to hidden layer
Update the weights in the network to reduce error
Until all examples classified correctly or stopping criterion satisfied
Return the network
```

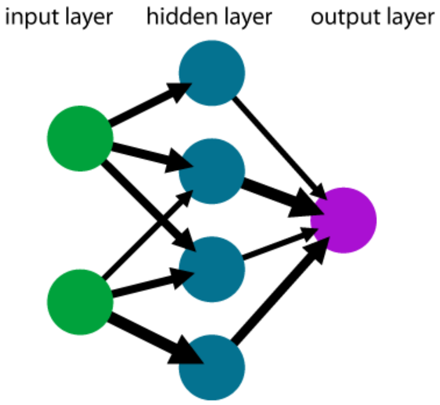


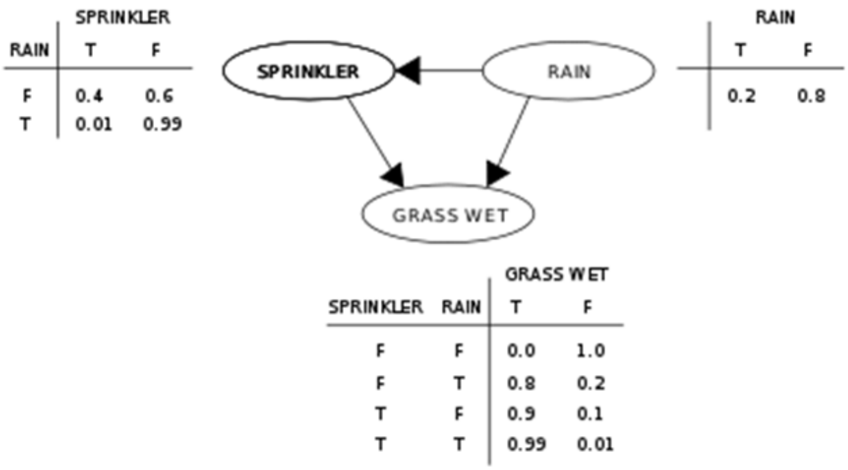
Figure 8. A simple neural network.

The output from the neural network is a simple binary value {0, 1} representing whether or not the patient's tumor progresses into malignant cancer. the classification boundary value to be the half-way point 0.5, so if the neural network's output value turns out to be above 0.5, it is categorized as 1; and values below 0.5 gets categorized as 0. The next important step is to determine the appropriate number of hidden variables in the neural network to avoid both under-fitting and over-fitting. The number of hidden variables should be strictly less than the number of inputs to the neural network, which is 4 in this case.

4.3.4. Classification by Bayesian networks

The naïve Bayesian classifier makes the assumption of class conditional independence, that is, given the class label of tuple, the value of the attributes are assumed to be conditionally independent of one other. This simplifies computation. When the assumption holds true, then the naïve Bayesian classifier is the most accurate in comparison with all other classifiers. However, dependencies can exist between variables. Bayesian networks specify joint conditional probability distributions. They allow class conditional independencies to be defined between subsets of variables. They provide a graphical model of causal relationships, on which learning can be performed. The learning can be perfomed in the graphical model of causal relationships, that they provide. Trained Bayesian belief networks can be used for classification.

A belief networks is defined by two components –a directed acyclic graph and a set of conditional probability tables (e.g., Figure 9). Each node in the directed acyclic graph represents a random variable. The variables may b discrete o continuous-valued. They may correspond to actual attributes given in the data to form a relationship (e.g., in the case of medical data, a hidden variable may indicate a syndrome, representing a number of symptoms that, together, characterize a specific disease). Each arc represents a probabilistic dependence. If an arc is drawn from a node *Y* to a node *Z*, then *Y* is a parent or immediate predecessor of *Z*, *Z* is a descendant of *Y*. Each variable is conditionally independent of its no descendants in the graph, given its parents, as is possible see in Figure 9.



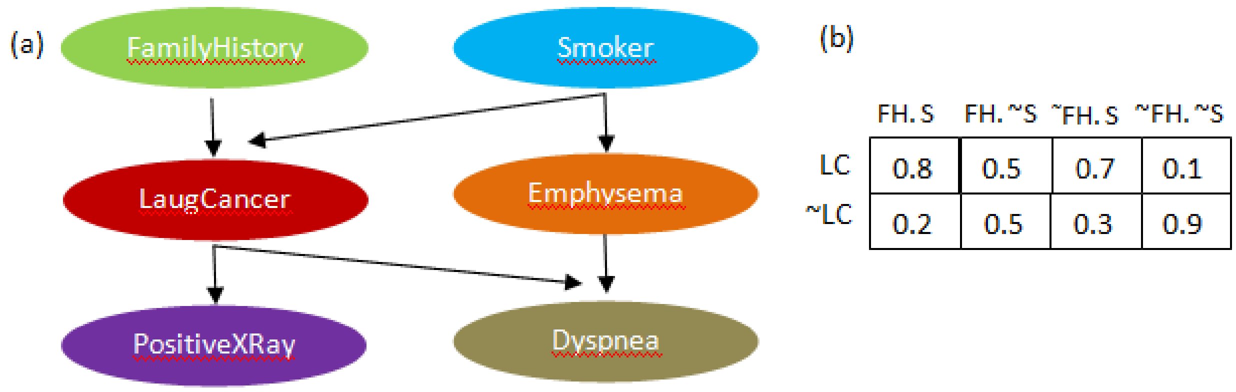
	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Figure 9. A example of Bayesian Network.

The Figure 10 is a simple Bayesian network for six Boolean variables. The arcs in figure 10 (a) allow the representation of causal knowledge. For example, having lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker. The arcs also show that the variable *LungCancer* is conditionally independent of *Emphysema*, given its parents, *FamilyHistory* and *Smoker*.



	FH. S	FH. ~S	~FH. S	~FH. ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

Figure 10. A simple Bayesian network: (a) A proposed casual model, represented by a acyclic graph. (b) The conditional probability table for the value of the variable LungCancer (LC) showing each possible combination of the values of its parents nodes, *FamilyHistory* (*FH*) and *Smoker* (*S*).

A belief network has one conditional probability table (CPT) for each variable. The CPT for a variable Y specifies the conditional distribution *P(Y|Parents(Y))*, where *parents(Y)* are the parents of *Y*. Figure 6 (b) shows a CPT for the variable *LungCancer*. The conditional probability for each knows value of *LungCancer* is given for each possible combination of values of its parents. For instance, form the upper leftmost and bottom rightmost entries, respectively, we see that

$P(LungCancer = yes \mid FamilyHistory = yes, Smoker = yes\ ) = 0.8$

$P(LungCancer = no \mid FamilyHistory = no, Smoker = no\ ) = 0.9$

A node within the network can be selected as an "output" node, representing a class label attribute. There may be more than one output node. Various algorithms for learning can be applied to the network. Rather than returning a single class label, the classification process can return a probability distribution that gives the probability for each class.

4.3.5. Visual data mining

Visual data mining discovers implicit and useful knowledge from large data Visual data mining have the capacity to find implicit and useful knowledge from great amount of data sets using data and/or knowledge visualization techniques. The human visual system is controlled by the eyes and brain, the latter of which can be thought of as a powerful highly parallel processing and reasoning engine containing a large knowledge base (Figure 11). Visual data mining essentially combines the power of these components, making it a highly attractive and effective tool for the comprehension of data distribution, patterns, clusters, and outliers in data. the eyes and brain, the latter of which can be thought of as a great highly parallel processing and reasoning engine that contain a large knowledge base (Figure 11). Visual data mining combines the power of these components, making it a highly attractive and effective tool for the comprehension of data patterns, clusters, distribution and outliers in data.

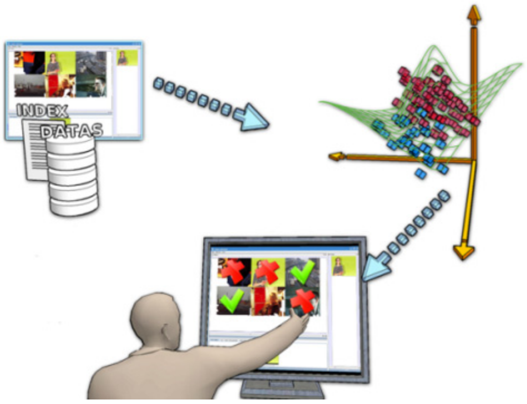


Figure 11. Human interact and processing large knowledge base.

Visual data mining can be viewed as an integration of two disciplines: data visualization and data mining. It is also closely related to computers graphics, multimedia systems, human computer interaction, pattern recognition, and high-performance computing. In general, data visualization and data mining can be integrated in the following ways:

Visual data mining can be viewed as an integration of two disciplines: data visualization and data mining. It is also closely related some disciplines: human computer interaction, pattern recognition, high-performance computing, computers graphics and multimedia systems. Data mining and data visualization can be integrated in the next ways:

- Data visualization: Data in a database or data warehouse can be view at different levels of granularity of abstraction, or as different combination of attributes or dimensions. Data can be presented in various visuals forms, such a boxplot, 3-D cubes, data distribution charts, curves, surfaces, link graphs, and so on. An example represented below:
  Data visualization: Data in a database or data warehouse can be view at different levels of granularity of abstraction, or as different combination of attributes or dimensions. Data can be presented in various visuals forms, such a data distribution charts, boxplot, curves, 3-D cubes, link graphs, surfaces, and so on. An example represented below:

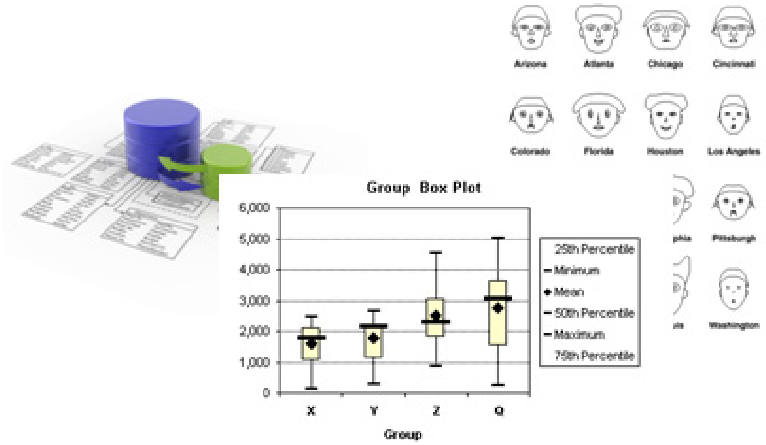


Figure 12. Boxplots showing multiple variable combinations in datasets.

- Data mining result visualization: Visualization of data mining results is the presentation of results or knowledge obtained from data mining in visual forms. Such forms may include scatter plots and boxplots, as well as decision tree, clusters, outliers, generalized rules and so on (Figure 9).
  Data mining result visualization: It means use techniques with which is possible the visual representation of results or knowledge that is obtained from data mining process. Such vicual forms may include scatter plots and boxplots, decision tree, clusters, outliers, generalized rules and so on (Figure 13).

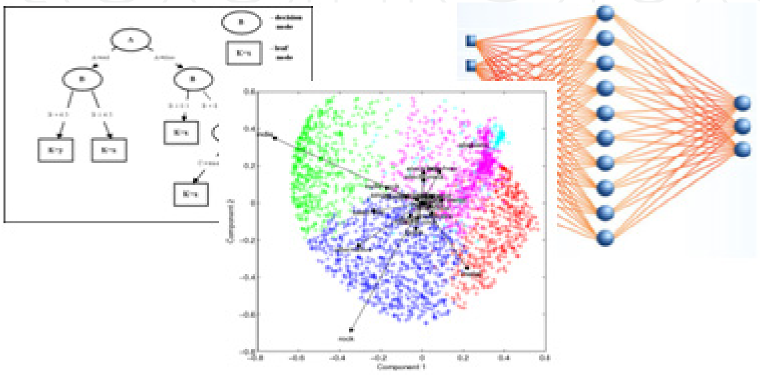


Figure 13. Visualization on data mining results.

- Interactive visual data mining: In visual data mining, visualization tools can be used in the data mining process to help users make smart data mining decisions. For example, the data distribution in a set of attributes can be displayed using colored sectors (where the whole space is representing by a circle). This display helps users determine which sector should first be selected for classification and where a good split point for this sector may be.

  The data mining process can be supported by visualization tools to help users to make smart data mining decisions. For example, in a circle that represents a whole space, the data distribution in a set of attributes can be displayed using colored sectors. With this visual representation the users can determine which sector should first be selected for classification and where a good split point for this sector may be.

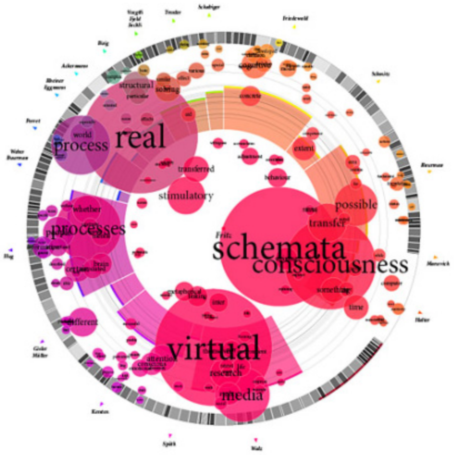


Figure 14. Example for circular data representation.

5. Analyzing people profile

The concept also is used to describe to the set of the characteristics that characterize to somebody or something. In the case of the human beings, the profile is associate to the personality. On the other hand, the word profile also is used very many to designate those particular characteristics that characterize a person and by all means they serve to him to be different itself from others. Your profile is built on other people's impressions and opinions, from the first time they hear your group's name or come into contact with one of its members. To some extent, you can control what people think and feel about your group, building a strong profile that will help you achieve action success. See Figure 15.

Orkut is a system of social networks used in Brazil by 13 million users, many of them, create more of a profile, and generate different relationships from their different profiles, this takes to think that they develop Bipolar Syndrome, to be able to establish communications with people of different life styles, and when they doing to believe other users that they are different people (Zolezzi-Hatsukimi, 2007).

The false profiles are created for: to make a joke, to harass other users, or to see who visualizes its profile. As the profile is false, the friends of this profile are also generally false, making difficult the tracking of the original author (Ochoa et. al., 2011). Using the tool of Data Mining denominated WEKA, it was come to develop a denominated "Ahankara" Model which perits reaize prediction of profiles in users of Orkut, which al-lows to understand the motivations of this type of profile and to determine if it has generated Syndrome Bipolar, to see figure 3 (Ponce et al., 2009). The model obtained Ahankara once used WEKA to look for the relations that us could be of utility to process the data. see Figure 16.

Figure 15. The profile show characteristics of a person or a group of people

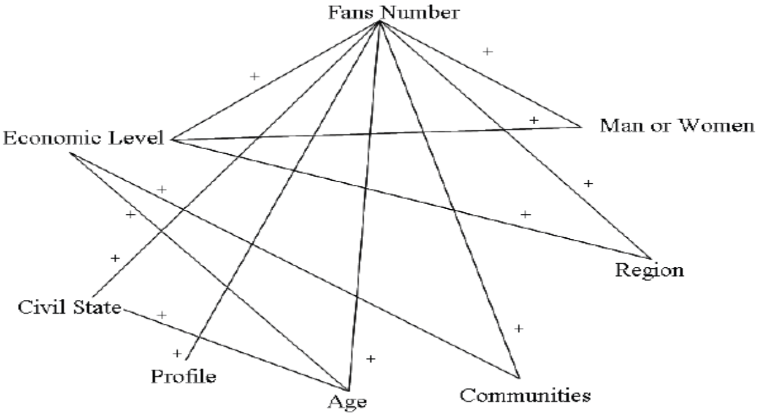


Figure 16. Ahankara Model

Waste. It is something that we produce as part of everyday living, but we do not normally think too much about our waste. Actually many cities generates a waste stream of great complexity, toxicity, and volume (see figure 17). In the management of solid waste have the problem relates to the household waste is the individual decision-making over waste generation and disposal. When the people decide how much to consume and what to consume, they do not take into account how much waste they produce (Ochoa et al., 2011).

E-commerce is the term use to describe the consumers that use the Internet for making purchases, usually refers typically to business to business type activities rather than consumer activity. It maybe more appropriate to refer to consumer activity in relation to purchasing goods and services on the Internet as on-line shopping (see figure 18). E-commerce is the term use to describe the consumers that use the Internet for making purchases, generally refers activities thath involve some business between two or more entities, rather than only consumer activity. This is more related to purchasign goods and services on the Internet rather than on-line shopping (see figure 18).

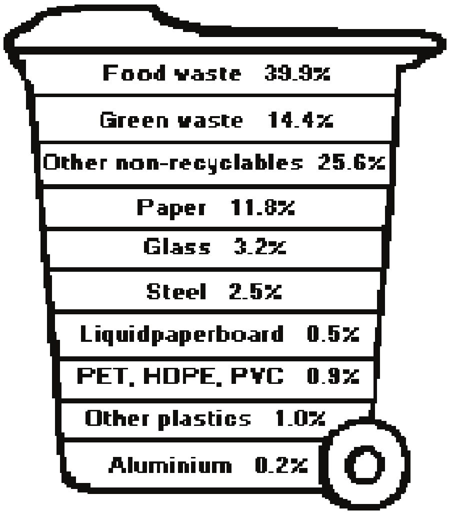


Figure 17. Example of composition by weight of household garbage

One of the important factors in the world of E-commerce is that it is much more than just a change in the way payments are made; E-commerce may not involve money at all. It gives customers the choice of making a wide range of transactions electronically rather than over the telephone, by post or in person. The E-commerce is not only a different way that the people use to pay for any thing, because, it is not simply money; this implies transactions that could be done by telephone, by post or in person, which the costumer can done electronically.

Figure 18. The people can buy services or things online

The major benefits of E-commerce are that it can help organizations to:

- improve working processes and service delivery;
- understand their customers better; and
- reduce costs through elimination of paperwork and bureaucracy.

Some of the most important benefits of E-commerce for the organizations are:

- The service delivery and the working process are improved.
- The organizations can understand their customers.

- They can reduce the costs caused by paperwork and bureaucracy.

In the e-commerce we has two different profile, the buyers and salesman profile, in this case we work with the buyers profile, In (Cocktail Analysis and Google, 2011) is sow a research about the buyers of fashionable clothes, in this work we can see that 42% of the people they have bought some article of clothes by Internet. They describe five different profiles only for the buyers of clothes, also it shows the relation of the purchases online with those of the physical stores. Data mining process can be used to determine the buyers profile, the enterprise can use this information to realize market studies in order to offer to the people specific products to them on the basis of its profile of purchases. Also the analysis of profiles is very important like dominion application of the data mining, can help to determine landlords us of conducts, habits, or of a single person or of a group, these data allow us to make predictions and can be used of diverse ways.

6. Data mining for E-comerce

The e-commerce is one of the profound changes that internet has induced in the people's lifestyle and in the way of doing business and transactions. The way that the consumers buy has been modified, appearing trends, patterns and preferences in specific groups. Some characteristics that can affect the consumerism by internet are: gender, age, social status, economic status, financial status, studies, culture, technology, knowledge of technology, geographic location, politics and others. In the early years of e-commerce, buying online was an erudite activity strictly dominated by "techies" and semi-technology literate individuals. These individuals were mostly made up of 20 to 35 year old males. This demographic were more comfortable and in tune with Internet's capabilities. But in recent years, the numbers of females making the technology leap to shop online is surging. Females are starting to harness Internet to make their lives easier and efficient (Christopher, 2004) . In the early years of e-commerce, buying online was an erudite activity. The individuals were mostly made up of 20 to 35 year old males. In recent years, the numbers of females making the technology leap to shop online is surging. (Christopher, 2004) . Data Mining (DM) has been applied successfully to find the patterns that the consumers create in the navigation trough the different web sites giving the opportunity to the enterprises to offer a better service.

6.1. Trends in E-commerce

In the e-commerce, the behavior of the consumers creates trends that change in the time for different variables. (Audette, 2010), mention three important trends in 2010 that should be considered by the people involved in the e-commerce (brands, retailers, and others).

6.1.1. Consumer focus is on price

The consumer always is looking for the lowest prices, it means, the best product for the best price or sometimes only the best price.

The consumer also looks for special offers that can balance that price was not the lowest. The offers can be the free shipping.

6.1.2. Riding the next wave: Video and visual search

It's very important now a day, the visual experience in the e-commerce because it is more attractive for the consumer and can be a reason to decide to buy something. The people spend a lot of time watching videos. A case is Mexico where the viewers watched 5 hours of video in YouTube in September 2011, and the audience has grown 17% to reach 20.5 million viewers, representing 85% of the total online population, according to a study by comScore. The next graphic shows video properties that prefer the viewers in Mexico.

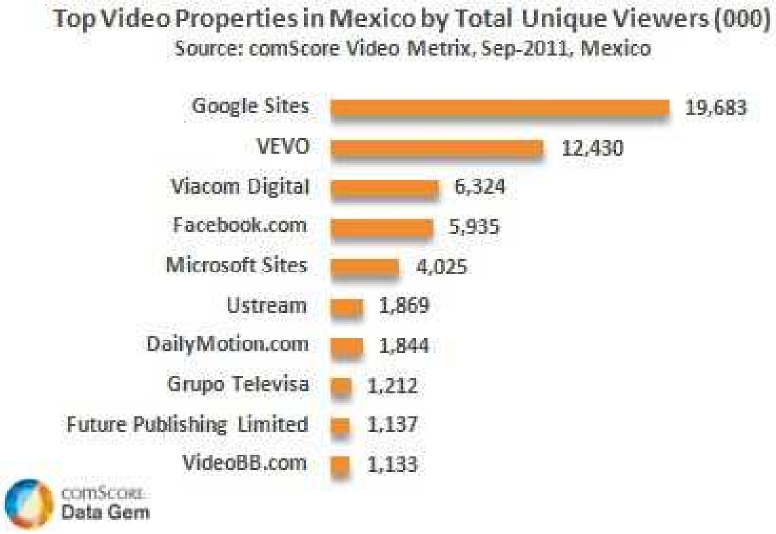


Figure 19. Top Video Properties in Mexico by Total Unique Viewers 2011

The video experience can improve the process information in a 30%, according Bing, and this can be explained because 65% are visual learners.

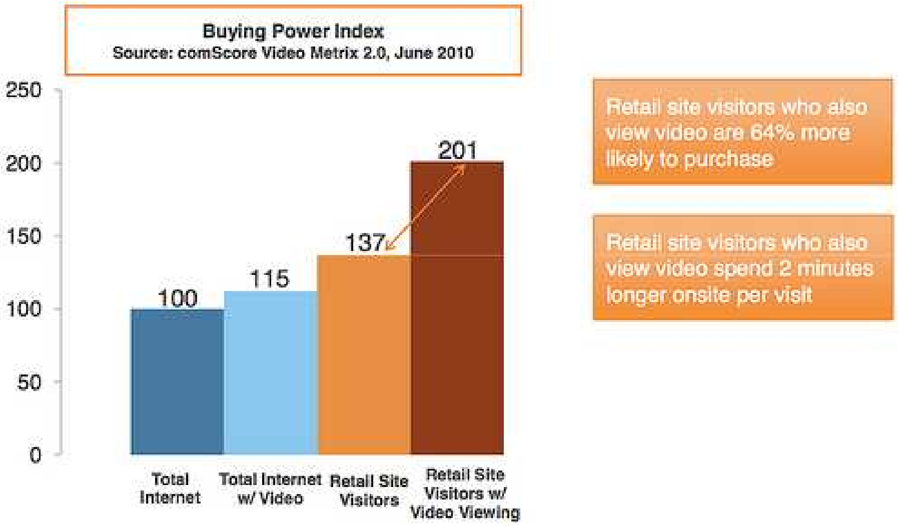


Figure 20. Buying Power Index 2010

6.1.3. Trend in technical SEO

SEO (Search Engine Optimization) is a technique which helps search engines find and rank your site higher than the others millions in response to a search query. This is based primarily in text. Google Instant had a little noticeable effect in the ecommerce clients.

6.2. Data mining and E-commerce

Data Mining (DM) have been applied to study the behavior of the users of different services (entertainment, mail, e-commerce, social network, among others) that internet provides. Many enterprises like Amazon and eBay have invested many resources to understand the consumers. Authors like (Sankar et al., 2002) explain why Web Mining, concept used for the first time by (Etzioni, 1996), is considered like sub-field of Data Mining. They say that Web Mining can be defined as "the discovery and analysis of useful information from the World Wide Web". The source of data can be the server, client, proxy server, or data bases of some enterprise. The web mining is divided in: Web content mining, Web structure mining, Web usage mining (Sankar et al., 2002). The principal tasks/phases of Web mining are: Information retrieval (resource discovery), information extraction (selection/preprocessing), Generalization (pattern recognition/machine learning), Analysis (validation/interpretation).

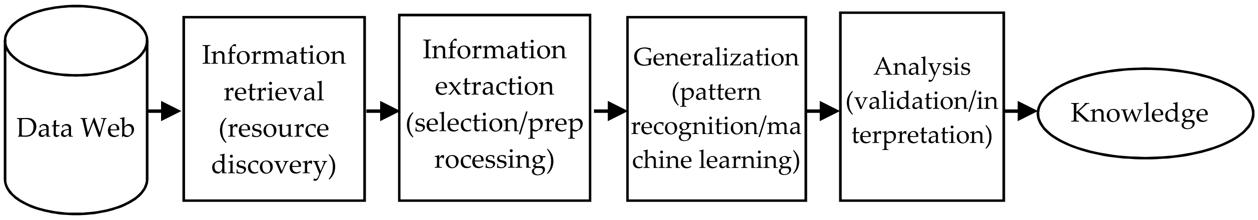

Data Web
Information retrieval (resource discovery)
Information extraction (selection/prep rocessing)
Generalization (pattern recognition/ma chine learning)
Analysis (validation/in terpretation)
Knowledge

Figure 21. Tasks of Web Mining

The Data Mining, or in this case Web mining, which is known, needs some problems with certain characteristics to obtain the major benefits. Those characteristics are (Ansari, Suhail, 2000):

- Large amount of data
- Rich data with many attributes
- Clean data collection
- Actionable domain
- Measurable return-on-investment

The e-commerce has every characteristics being a "Killer Domain" of Data Mining (Ansari, Suhail, 2000). The attributes more important in the e-commerce are RFM (Recency, Frequency and Monetary). Examples of these attributes are date, time, duration session, quantity, purchase (Ansari, Suhail, 2000). Other attributes are IP address, URL, error code, among others; however these are common in logs that are not created for analysis (De Gyves Camacho, 2009). The attributes (columns) related with time and date are used to find important hidden patterns. One of the applications of Web mining is the learning of Navigation patterns (Web usage Mining).

6.3. Artificial immune system applied to web mining

The Artificial Immune System (AIS) is a bio-heuristic based in the Natural Immune System (NIS). One of the characteristics that make interesting the NIS are: highly distributed, highly adaptive, self-organising, maintain a memory of past encounters, and learn of new encounters. Some algorithms that have been proposed to use in data mining, are based in theories like negative selection, clonal selection and immune network. However new algorithms have been created inspired in other characteristic or theories. In order to approximate a solution of the learning of navigation patterns an immune-inspired algorithm is proposed which is based in the immune network and was developed by (Timmis et al., 2000). The AIS has some characteristics that can be improved but is good for a first approximation. This algorithm is proposed to clustering the similarities of the users' behavior and according of the pattern, in a next step, suggest the best structure of the web site to the consumers to improve their experience. In this way, the companies con offer a better service, adapted to the consumers' necessities and finally increase sales.

7. Data mining to mobile ad hoc networks security

Mobile radio technologies, for both voice and data communication, has experienced a rapid growth and diverse concepts have been introduced in networking. However the concept of ad hoc network is not new, the paradigm started from the beginning of late 90's and gradually became popular with the wide range of deployments of IEEE 802.11x based WLAN, despite regularly ad hoc networks are based on single-hop peer-to-peer networking between several wireless devices, in different specialized scenarios such as control applications, logistics and automation, surveillance and security, transportation management, battlefields, environmental monitoring, unexplored and hazardous conditions, home networking, etc. multi-hop wireless networks are used. Multi-hop wireless ad hoc network consists of a number of self-configurable nodes (e.g. IEEE 802.11-based WLAN, 802.16-based WiMAX, ZigBee, Bluetooth, etc.) to establish an on-demand network using multiple hops paths if required where no network infrastructures pre-exist. The basic block of multi-hop ad hoc networking can be divided into four major specialized categories – Mobile Ad hoc Networks (MANET), Wireless Mesh and Hybrid Networks (WMN), Vehicular Ad hoc Networks (VANET) and Wireless Sensor Networks (WSN) (Kamal, 2010). MANET is the most theoretically researched arena of ad hoc networking which is a collection of autonomous and mobile network objects of any kind with truly dynamic and uncertain mobility that communicate with each other by forming a multi-hop radio network and maintaining connectivity in a decentralized manner. Nowadays, MANET has become a practical platform for pervasive services, i.e., the services that are requested and provided anywhere and anytime in an instant way. This kind of service is very valuable for mobile users, especially when fixed networks (e.g. Internet) or mobile networks are temporarily unavailable or costly to access. A generic concept of the general-purpose pure MANET is shown in Figure 22.

In Figure 22, let's suppose that node A wants to send data to node C but node C is not in the range of node A. Then in this case, node A may use the services of node B to transfer data since node B's range overlaps with both the node A and node B. In MANET, no fixed infrastructure, like base station or, mobile switching center is required. Instead, every possible wireless mobile host within the perimeter of radio link acts as an intermediate switch and participates in setting up the network topology in a self organized way.

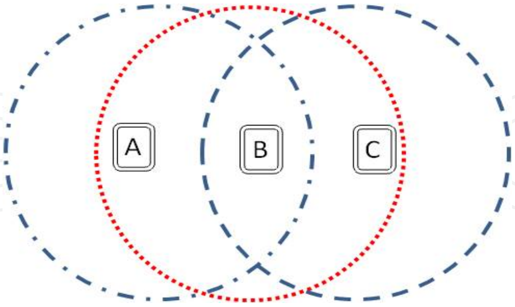


Figure 22. A Simple MANET

7.1. Data mining to deal with vulnerabilities of MANET

Despite the advantages, accord to Nakkeeran, the nature of mobility creates new vulnerabilities due to the open medium, dynamically changing network topology, cooperative algorithms, lack of centralized monitoring and management points and yet many of the proven security measures turn out to be ineffective (Nakkeeran et al. , 2010).

Despite the advantages, accord to Nakkeeran, the nature of mobility creates new vulnerabilities due to the open medium, dynamically changing network topology, cooperative algorithms, lack of centralized monitoring and management points and yet many of the proven security measures turn out to be ineffective (Nakkeeran et al. , 2010).

All these mean that a wireless ad-hoc network will not have a clear line of defense, and every node must be prepared for encounters with an adversary directly or indirectly. In order to avoid such circumstances requires the development of novel architectures and mechanisms that protect wireless networks and computer applications. Hence diverse research scopes do exist. Unfortunately, investigations are principally targeted towards routing, scheduling, address assignment, developing protocol stack etc. These are mainly functional properties. However, as nomadic and ubiquitous computing reaches its full potential, semantics and security will play the leading role, because the flexibility in space and time induces new challenges towards the security infrastructure. Due to in the case of the securtiy infraestructure, the flexibility in space and time will generate new challenges. Therefore, the traditional way of protecting wired/wireless networks with firewalls and encryption software is no longer sufficient. A very recurrent solution are Intrusion Detection Systems (IDS) (Mishra, 2004). Generally IDS can be defined as the detection of intrusions or intrusions attempts either manually or via software, through the use of schemes that collects the information and analyzing it for uncommon or unexpected events. Intrusion Detection (ID) is the process of monitoring and analyzing the events which occurred in a digital network in order to detect signs of security problems (Shirbhate, 2011). Then the ID is data analysis process, for this reason, as well as the growth of volume of existing data and insufficiency of data storage capacity leads us to the dynamic processing data and extracting knowledge. So the nature solution is utilizing data mining techniques, for example, anomaly detection techniques could be used to detect unusual patterns and behaviors, link analysis may be used to trace the viruses to the perpetrators, classification may be used to group various cyber attacks and then use the profiles to detect an attack when it occurs, prediction may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations (Khalilian, 2011). Data mining can improve variant detection rate, control false alarm rate and reduce false dismissals (Jianliang, 2009).

7.2. Intrusion detection methodologies

If we want to categorize intrusion detection methods, we will recognize two main aspects for grouping approaches, which one group refers to type of attack according to the kind of input information the analyze includes host based, network based, wireless and Network Behavior Analysis (NBA). Another group of approaches refers to solutions techniques which are misuse detection, anomaly detection methods and hybrid methods (Khalilian, 2011).

a. Host based methods.
   This methods are based on data source category; consequently, its data comes from the records of various activities of hosts, including system logs, audit operation system information, etc. the main architecture for this kind of methods is similar to network based.

b. Network based methods.
   These systems analyze network packets that are captured on a network. Network packet is the data source for network intrusion detection system.

c. Wireless methods.
   Wireless intrusion detection system monitors wireless network traffic and analyzes its wireless networking protocols to identify suspicious activity involving the protocols themselves. It cannot identify suspicious activity in the application or higher-layer network protocols such as TCP, UDP that the wireless network traffic is transferring. So each node is responsible for detecting signs of intrusion locally and independently, but neighboring nodes can collaboratively investigate in a broader range.

d. Network Behavior Analysis.
   NBA which examines network traffic to identify threats that generate unusual traffic flows, such as distributed denial of service (DDoS) attacks, certain forms of malware such as worms, backdoors, and policy violations. NBA systems are also deployed to monitor flows on an organization's internal Networks, and are also sometimes deployed where they can monitor flows between an organization's Networks and external networks such as the Internet.

Figure 23 shows the basic architecture for NIDS in data mining, which is very similar to the other detection methods

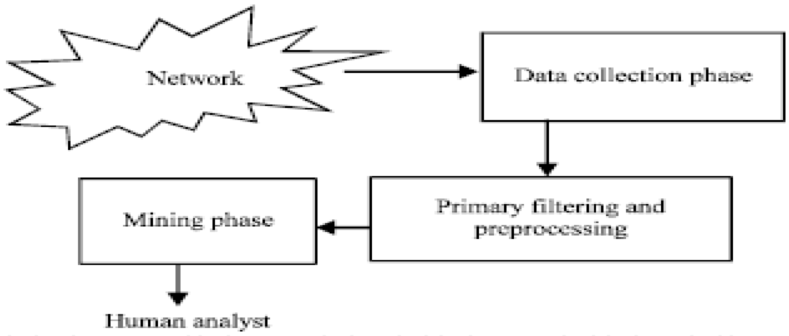


Figure 23. Basic NIDS architecture

e. Misuse based methods.
Misuse detection which the main study is the classification algorithms relies on the use of specifically known patterns of unauthorized behavior. In misuse detection related problems, standard data mining techniques are not applicable due to several specific details that include dealing with skewed class distribution, learning from data streams and labeling network connections. The problem of skewed class distribution in the network intrusion detection is very apparent since intrusion as a class of interest is much smaller i.e. rarer than the class representing normal network behavior. In such scenarios when the normal behavior may typically represent 98-99% of the entire population a trivial classifier that labels everything with the majority class can achieve 98-99% accuracy (Dokas, 2002). It is apparent that in this case classification accuracy is not sufficient as a standard performance measure. ROC analysis and metrics such as precision, recall and F-value have been used to understand the performance of the learning algorithm on the minority class. A confusion matrix as shown in Table 1 is typically used to evaluate performance of a machine learning algorithm.

Confusion matrix (Standard metrics)		Predicted connection label	
		Normal	Intrusions (Attacks)
Actual connection label	Normal	True Negative (TN)	False Alarm (False Positive)
	Intrusions (Attacks)	False Negative (FN)	Correctly Detected Attacks (True Positive)

Table 10. Standards metrics for evaluations of intrusions (attacks)

In addition, intrusions very often represent sequence of events and therefore are more suitable to be addressed by some temporal data mining algorithms. Finally, misuse detection algorithms require all data to be labeled, but labeling network connections as normal or intrusive re-quires enormous amount of time for many human experts. All these issues cause building misuse detection models very complex.

f. Anomaly based methods.
Misuse detection system unable to detect new or previously unknown intrusions occurred in computer system or digital network. Novel intrusions can be found by

anomaly detection which the main study is the pattern comparison and the cluster algorithms ((Khalilian, 2011). The basic idea of clustering analysis originates in the difference between intrusion and normal pattern; consequently, we can put data sets into different categories and detect intrusion by distinguish normal and abnormal behaviors. Clustering intrusion detection is detection for anomaly with no supervision, and it detects intrusion by training the unmarked data.

Most anomaly detection algorithms require a set of purely normal data to train the model, and they implicitly assume that anomalies can be treated as patterns not observed before. Since an outlier may be defined as a data point which is very different from the rest of the data, based on some measure. In statistics-based outlier detection techniques the data points are modeled using a stochastic distribution and points are determined to be outliers depending upon their relationship with this model. However, with increasing dimensionality, it becomes increasingly difficult and inaccurate to estimate the multidimensional distributions of the data points. However, recent outlier detection algorithms are based on computing the full dimensional distances of the points from one another as well as on computing the densities of local neighborhoods. Nearest Neighbor (NN), Mahalanobis-distance Based Outlier Detection and Density Based Local Outliers (LOF) are approaches for recent outlier detection algorithms.

g. Hybrid methods.

Through analyzing the advantages and disadvantages between anomaly detection and misuse detection, a mixed intrusion detection system (IDS) model is designed. First, data is examined by the misuse detection module, and then abnormal data detection is examined by anomaly detection module. The intrusion detection system (IDS) is designed based in the advantages and disadvantages of the models of intrusion detection that are: anomaly detection and misuse detection. The first step is examine the data with de misuse detection module and in the second step, the anomaly detection module analyze the atypical data detected.

7.3. New trends in safety MANET

The ultimate goal of the security solutions for wireless networks is to provide security services, such as authentication, confidentiality, integrity, anonymity, and availability, to mobile users. The final goal of the security solutions for wireless networks is to offer security services to mobile users. This services are authentication, integrity, confidentiality, anonymity, and availability.This kind of schemes depend on cooperation amongst the nodes in a MANET for identifying nodes that are exhibiting malicious behaviors such as packet dropping, packet modification, and packet misrouting, so most of this methods assume that this problem can be viewed as an instance of detecting nodes whose behavior is an outlier when compared to others. Some novel solutions incorporate mobile agents (Nakkeeran et al., 2010) to provide solution against security issues in MANET networks. With the help of home agent and mobile agents, it gathers information from its own system and neighboring system to identify any attack and through data mining techniques to find out the attacks has been made in that networks.With the help of home agent and mobile agents, it is possible to extract information from both, own system and neighbor system, to identify any attack and with data mining techniques try to find the attacks that has been perpetrated in such networks. Home agent is present in each system and it gathers information about its system from application layer to routing layer.

Each system have a home agent, which should obtain iformation about the system from application layer to routing layer.

Mobile agents are a special type of agents defined as "processes capable of roaming through large networks such as the ad hoc wireless network, interacting with machines, collecting information and returning after executing the tasks adjusted by the user". Mobile agents are a special type of agents defined as "processes capable of roaming through large networks such as the ad hoc wireless network, interacting with machines, which return the colected information when finishing the execution of the tasks of the user". Often such proposals provide the three different techniques to provide suffice security solution to current node, Neighboring Node and Global networks.

Frequently, the proposals afford the three differents techniques that are used to offer security solution to current node, Neighboring and Global networks.

The trust together cooperation are another kind of novelty solution (Li, 2009); the idea of them consists in an algorithm to can help us identify the outliers, which are generally the nodes that have exhibited some kind of abnormal behaviors. Given the fact that benign nodes rarely behave abnormally, it is highly likely that the outliers are malicious nodes. Moreover, a multi-dimensional trust management scheme is proposed to evaluate the trustworthiness of the nodes from multiple perspectives. There are many techniques that have been discussed to prevent attacks in wireless ad hoc networks but most of them have in common that are based on cooperation and on methods based on the principle of anomalies.

8. Conclusions and another specific application domains improved with data mining

Diverse applications based on Data Mining have the objective to learn the patterns of the users' interaction with the Web or Data Repository. The data includes user profiles, registration profiles, user queries, and any data generated by the users' interaction with the web. This is useful, for example, to restructure the web page according the preferences of the users. This means that the web site is going to provides information, special offers, and others, that can be interesting for the consumers according their patterns of interaction or according the hour of the day. Also, this can be used to design offline strategies. One process to make projects of web usage mining is described in (DAEDALUS, 2002). To find the patterns in Web usage mining, the techniques used are: Clustering and Classification, Association rule detection, Path analysis, Sequential patterns detection. The use of any technique mention above to analyze automatically the data implies difficulties by the complexity of the problem (heterogeneous data, and others) and the limitations of the existing methodologies. To overcome that difficulties and limitations is necessary to use other techniques and methodologies like soft computing. Some algorithms developed to address Data Mining using techniques of Soft Computing are revised by (Mishra el al., 2004). Other application of Data Mining using evolutionary algorithms were proposed by (Ochoa et al., 2011) obtaining good results. In addition, we described another specfic applications domains such as: Deterining Euskadi ancesters based on family names and compare the anthropmetry of the individuals to found patterns of their ancesters; Organizational Models to supporting little and medium business related with Regional Development; Organizational Climate to identify cases of Burnout Syndrome characterized by high expectatives of productivty and Organizational Culture (Hernández et al, 2011); Identification of the use of new languages related with the songs from Eurovision for exameple Udmurt language in the entry from Russia to Eurovision Song Contest'2012; Analysis of Pygmalion Effect on people from Pondichérry in India whom be considering more closely culturally of Francophonie because the French influence in their past lifes; Zoo applications to classify more vulnerable species in an interactive map (see figure 24) or identify the adequate kind of avatars on a roll multigame players associated with cultural aspects in this case Brazilian people and their selections of features related with spcific skills (see figure 25).

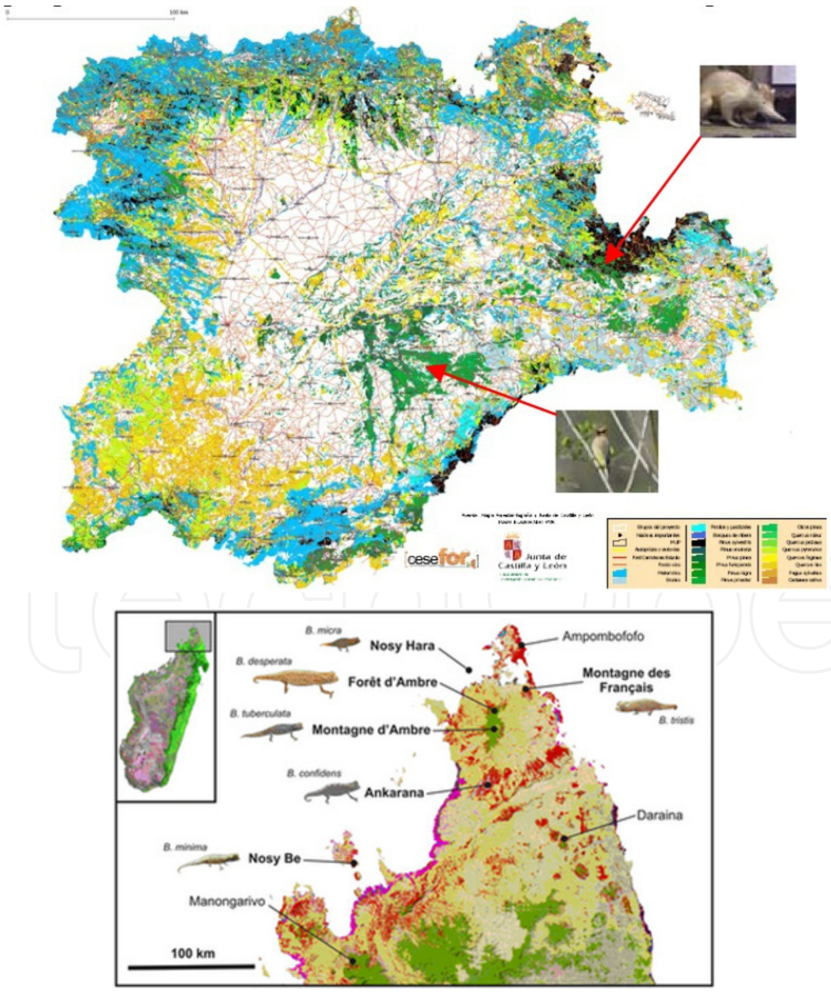


Figure 24. Interactive map based on data mining, locating the habitats of species of reptils, birds and mammals specifying the ubiquity of their behavior –changes provoked by the human- during the time.

Figure 25. Cultural Avatars related with traditional aspects and antropometry from Brazlian people used in Multi player games according of specific skills used on the online game.

In addition Tatebanko traditional Japanese Dyoram is using new ideas based with a Hybrid Algorithm conformed by the use of Data Mining and a Bioinspired Algorithm to built 3D scenario (Ochoa et al., 2012) including issues of specific time and location by each one.

Author details

Alberto Ochoa, Daniel Azpeitia, Petra Salazar, Emmanuel García and Miguel Maldonado
Juarez City University, México

Rubén Jaramillo and Jöns Sánchez
LAPEM, México

Javier González and Claudia Gómez
ITCM, México

Julio Ponce, Sayuri Quezada, Francisco Ornelas and Arturo Elías
UAA, México

Edgar Conde and Víctor Cruz
Veracruzana University, México

Lourdes Margain
Universidad Politécnica de Aguascalientes, México

9. References

Ansari, Suhail; Kohavi, Ron; Mason, Llew and Zheng, Zijian. Integrating E-Commerce and Data Mining: Architecture and Challenges. WEBKDD'2000 workshop: Web Mining for E-Commerce -- Challenges and Opportunities, 2000.

Audette Adam. Founder and Presidente of AudetteMedia., Nov 29, 2010 http://searchengineland.com/3-important-trends-to-watch-in-ecommerce-56890

Cocktail Analysis and Google (2011). El comportamiento del Comprador de Moda OnLine. http://tcanalysis.com/

Cheng, Ri; Kai, L.; Chun, B.; Shao-Yu, D.; Gou-Zheng, X. Study on Partial Discharge Localization by Ultrasonic Measuring in Power Transformer Based on Particle Swarm Optimization. *International Conference on High Voltage Engineering and Application*. (2008). 600-603.

Christopher, James. E-Commerce: Comparison of On-line Shopping Trends, Patterns and Preferences against a Selected Survey of Women. Kingston University. MSC Business Information Technology Program. November 2004.

DAEDALUS – Data, Decisions and Language, S.A.: Minería Web: Documentos básico DAEDALUS. White Paper, C-26-AB-6002-010, Noviembre 2002. http://www.daedalus.es

De Gyves Camacho, Francisco Manuel. Web Mining: Fundamentos Básicos Doctorado en informática y automática Universidad de Salamanca. Informe Técnico, DPTOIA-IT-2006-003. Mayo 2009.

Dokas, P., et al. *Data mining for network intrusion detection*. in *In Proceedings of the NSF Workshop on Next Generation Data Mining*. 2002. Baltimore, MA.

Etzioni, O. The world-wide web: Quagmire or goldmine?, Communications of the ACM, vol. 39, pp. 65-68, 1996.

Hernández, Alberto et al. Aplicación de la minería de datos para la toma de decisiones: El Caso de la cultura organizacional en una tienda del IMSS, XVI Congreso Internacional de Contaduría, Administración e Informática, 2011.

Jianliang, M., S. Haikun, and B. Ling. *The Application on Intrusion Detection Based on K-means Cluster Algorithm*. in *Information Technology and Applications, 2009. IFITA '09. International Forum on* 2009. Chengdu IEEE, Press.

Kamal, J.M.M., *A Comprehensive Study on Multi-Hop Ad hoc Networking and Applications: MANET and VANET*, in *Faculty of Computing, Engineering and Technology*. 2010, Staffordshire University: Stafford. p. 155.

Khalilian, M., et al., *Intrusion Detection System with Data Mining Approach: A Review.* Global Journal of Computer Science and Technology (GJCST), 2011. 11(5): p. 29-34.

Kohonen T. Engineering Applications of Self Organizing Map. *Proceedings of the IEEE*. (1996).

Li, W., J. Parker, and A. Joshi, *Security through Collaboration in MANETs Collaborative Computing: Networking, Applications and Worksharing*, E. Bertino and J.B.D. Joshi, Editors. 2009, Springer Berlin Heidelberg. p. 696-714.

Mishra, A., K. Nadkarni, and A. Patcha, *Intrusion detection in wireless ad hoc networks* Wireless Communications, IEEE 2004. 11(1): p. 48-60.

Nakkeeran, R., T. Aruldoss Albert , and R. Ezumalai, *Agent Based Efficient Anomaly Intrusion Detection System in Adhoc networks.* IACSIT International Journal of Engineering and Technology, 2010. 2(1): p. 52-56.

Ochoa, Alberto; Castillo, Nemesio; Yeongene, Tasha & Bustillos, Sandra. Logistics using a new Paradigm: Cultural Algorithms. Programación Matemática y Software, Vol. 1. No 1. Dirección de Reservas de Derecho: 04-2009-011611475800-102. 2011.

Ochoa, Alberto et al. New Implementations of Data Mining in a Plethora of Human Activities. In Knowledge-Oriented Applications in Data Mining, ISBN 978-953-307-154-1, 2011.

Ochoa, Alberto et al. Developing a Traditional Tatebanko Dyoram using Cultural Algorithms V Workshop Hybrid Intelligent Systems at MICAI'2012 to publish.

Orihuela, José Luis. Nuevos Paradigmas de la Comunicación. Retrieved March. Vol. 12, España (2002).

Ponce, Julio; Hernández, Alberto; Ochoa, Alerto et al. Data Mining in Web Applications. In Data Mining and Knowledge Discovery in Real Life Applications, ISBN 978-3-902613-53-0, 2009.

Rubio-Sánchez, M. *Nuevos Métodos para Análisis Visual de Mapas Auto-organizativos*. PhD Thesis. Madrid Politechnic University. (2004).

Salaveria, Ramón (2009). El Impacto de Internet en los Medios de Comunicación en España. Comunicación Social Ediciones y Publicaciones. Pp. 11-15, 2009.

Sandoval, Rodrigo, Saucedo Nancy Karina. Grupos de Interés en las Redes Sociales: El caso de Hi 5 y Facebook en México. Tecnociencia Chihuahua. Vol. IV, No. 3, 2010.

Sankar K. Pak, Varun Talwar, Pabitra Mitra. Web Mining in Soft Computing FrameWork: Relevance, State of the Art and Future Directions. IEEE Transactions on Neural Networks Vol. 13, No. 5 pp 1163-1177, September 2002.

Shirbhate, S.V., V.M. Thakare, and S.S. Sherekar, *Data Mining Approaches For Network Intrusion Detection System.* International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2011. 2(2): p. 41-44.

Timmis, J, Neal, M and Hunt, J. An Artificial Immune System for Data Analysis.*Biosystems. 55(1/3)*, pp. 143-150. 2000.

Zolezzi-Hatsukimi, Z. Implement social nets using Orkut, Proceedings of CHI'07, Nagoya, Japan, 2007.