Hello from Mars: A Sociolinguistic approach to Mars 500 Twitter

Alberto Ochoa[1], Emmet Kaudallis[2, 3], Arturo Hernández[4] & Sandra Bustillos[1]

[1]Instituto de Ciencias Sociales y Administración, UACJ; México, [2]ISTC-CNR, Rome; Italy, [3] Kaunas University, Lithuania, [4] CIMAT, México.
cbr_lad7@yahoo.com.mx
(Paper received on September 09, 2010, accepted on October 20, 2010)

Abstract. The present paper discusses a research related to Sociolinguistics using Weka, a tool that mine information of the structure and content of speech of a Italian Cosmonaut, Diego Urbina Vallessacchi whom participate in Mars 500 project (A simulation of a travel to Mars) with the purpose of discovering hypostasis and parataxis in his speech and in his followers, which consists of relation in formal and informal use of language induced by the relation with another speakers in different languages (He write daily in three languages: Italian Spanish & English), this phenomena has been documented recently, but with few detailed research with truly information, for this purpose we record speeches in Twitter related with a social networking of followers (1600) whom send messages to him during the time of this mission (Mars 500), to explore a detailed sociolinguistics analysis.

Keywords Sociolinguistics, Speech Analysis, Multilanguage analysis.

1 Introduction

Social Data Mining Systems allow the analysis of the society's behavior. These systems do that by mining and redistributing the information on computer files storing the social activity. Although, we generate two general questions to evaluate the performance of such systems: (1) is the extracted information of any value? And (2) is possible to determine if a set of physical separated people can show a similar way of thinking about likes and preferences based on their language spoken?

We made an analysis that provides positive answers for both questions. We live in an age plenty of information. The Internet offers endless possibilities. Web sites to experience, music to listen, chats rooming, and unimaginable products and services offering to the consumer an endless options varying in quality. People are experiencing difficulties to manage the information: they can not and do not have time to evaluate the whole options by themselves, unless the situation seriously forces them to do that. In this paper we try to analyze a group of individuals following a Twitter user, and if he can understood conversations with many people at same time, because this Italian Cosmonaut send these messages in three different languages. A task to

manage information which several internet users must do is "the subject management", searching, evaluating and organizing information resources for a specific subject sometimes Users search for professional interest subjects, some other times just for personnel interest. Our approach to this problem combines social data mining with information about sociolinguistics. In the daily life, when people desire forming part of a social group, without having the knowledge to chose among different alternatives, they trust frequently on the experience and opinions of others. They look for advice in their linguistic-social group with certain likes and ways of thinking. When evaluating the offered perspectives by similar persons to them, or from recognized experts on a subject. For instance, a Usenet of users of Italian origin can recommend certain type of food and where to buy the ingredients also, when registers of these activities exist, these can be analyzed. For our research we need this information to understand how these sites on the web are populated and conformed. Social data mining can be applied to analyze the records generated on the web [5] (answering the question: Which are the most visited sites for the most of people?), online conversations [7] (Which are the sites where people purchase "thematic" things for a community.

This paper is organized in five sections. In section one, we introduce our paper. En section two, we describe our Sociolinguistic approximation focusing in Social Data Mining. In section three we discuss the application of WEKA to confirm the hypothesis of our research. In section four, we discuss the tests made to the analyzed information. In section number five, we discuss the results generated for the tests, and finally on the last section, we give the conclusions of our research.

2 Sociolinguistic approximation

Distinction between emotional grammatical and is not clear but it is possible to be conceited that emotional is pejorative and that thinks that the hypostatical style is superior. An analogy can be realized that "While a masculine oration usually is like a game of Chinese boxes, one fits within the other, a feminine one is like a Rep necklace them united by a thread of Greek is and other similar words", is for that reason that parataxis is common in British prose and the hypostasis are common in Renaissance prose.

2.1 Social Data Mining

The motivation to make an approach by means of applications with Data Mining is based on previous works of Social Data Mining in this research area. This research area emphasizes the role of the collective analysis of conduct effort, rather that the individual one. A social tendency results from the decisions of many individuals, joined only in the location in where they choose to coexist, yet this, still it reflects a rough notion of what the researchers of the area find of what could be a correct and valid social tendency [6]. The social tendency reflects the history of the use of a collective behavior, and serves like base to characterize the behavior of future descendants [3] or another new speaker of a language.

3 System Development

The system will be able to analyze the behavior for three samples of messages from the daily information write by Diego Urbina Vallessachi of a twit list recorded in Twitter: Italian, Spanish & English was used for this purpose, by means of WEKA use, which has demonstrated being an efficient tool for searching hiding parameters that must be discovered [5]. The compiled information was analyzed to discover behavior patterns that share these individuals, and based on their gender (the females tend to send more large messages and with more emotional support), we determine if this behavior was an innate or induced tendency by their linguistic community.

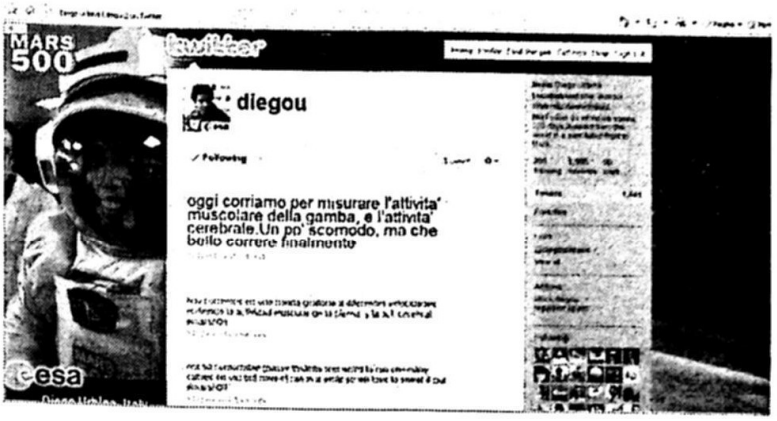

Figure 1. *Twitter's Diego Urbina from the Mars 500 Project.*

The name of Data Mining derives from the similarities between looking for valuable information in great data bases - for example: to find information of the tendencies of the society behavior in great amounts of stored Gigabytes – and mining a mountain to find a vein of valuable metals. Data mining automates the process to find predictable information in great data bases (See Figure 1). Questions that traditionally required an intensive manual analysis now can be directly and quickly answered from data [3].

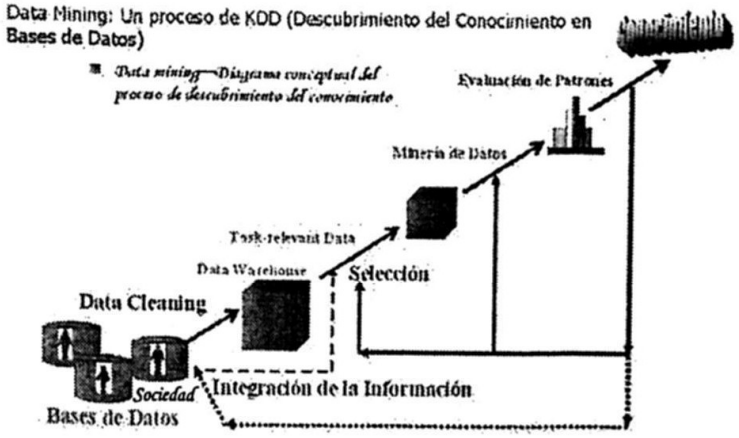

Figure. 2. *Data Mining process.*

The society information inside a *Data bases* is cleaned and stored in a *Data Ware House*, then is mined by means of a loop back *selection* and *patterns evaluation* process processes.

4 Applied tool

Use Data mining tool WEKA to analyze data. First, we proceed to develop a model that allows explain the behavior by three samples of people, and how affects their speech style. Figure 3 and 4 discover the existent relation among hypostasis and parataxis parameters, used by the different languages, the speakers communicate with the Italian Cosmonaut.

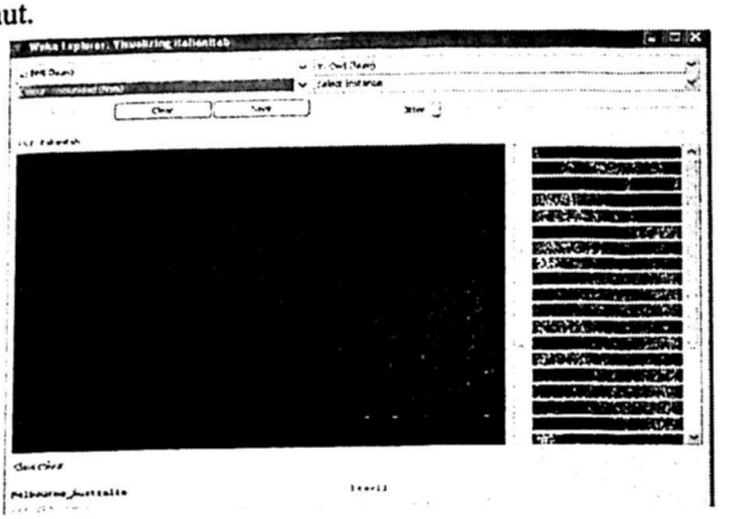

Figure. 3. *WEKA justifying the relation among Hypostasis found in several messages from Mars 500.*

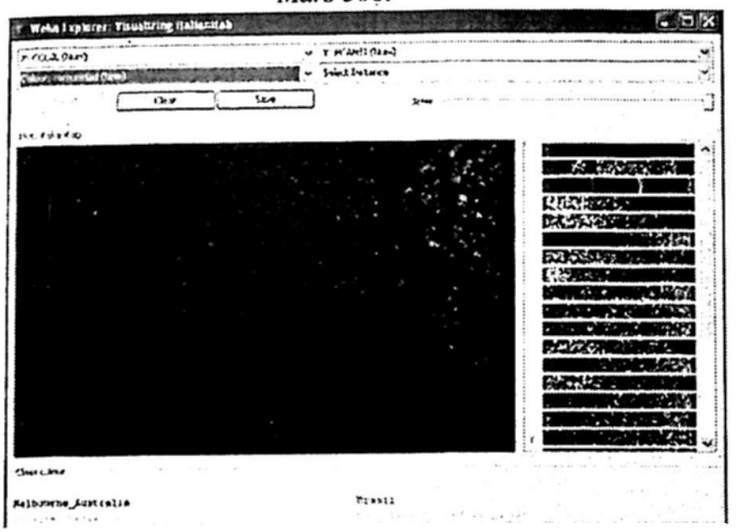

Figure. 4. *Relation of Parataxis found in users of Twitter following to Diego Urbina.*

We found in both cases that the Italian speakers showed a higher hypostasis and lowest parataxis regarding speakers of Spanish & English. This can be explained by the use of informal speech of Italian because they try to assimilate more easy at peo-

ple with commonly ancestors (the mother of Diego Urbina is Italian native speaker), and purchase decision is highly influenced induced by the language Community.

5 Results

We took in consideration a sample of 587 messages send and received (-212 Italian Messages, 190 Spanish Messages & 185 English Messages-) in Diego Urbina's Twitter during four months (04/06/10 – 05/10/10) conformed by three samples (Sample 1 with Italian Messages, Sample 2 with Spanish Messages and Sample 3 with English Messages), and using their conversations in a social networking, to identify different behaviors (See Table 1).

Table 1. Distributions of demands by category and sorted by three analyzed samples.

	Sample 1	Sample 2	Sample 3
Language	Italian	Spanish	English
N	212	190	185
Imperatives	12%	36%	26%
Directives declaratives	5%	6%	7%
Directives of Simulation	11%	4%	5%
Interrogatives Directives	2%	0%	1%
Interrogatives Postscripts	35%	16%	28%
Joint Directive	15%	3%	11%
Explosive Questions	2%	11%	4%
Information Questions	16%	22%	17%
Mechanisms of attraction of the attention	2%	2%	1%
Total	100%	100%	100%

The use of Data mining in social aspects has demonstrated being key part to corroborate the linguistics tendencies of a group established within of a common social networking, we found variations depending the intention of message and the linguistic resource used in different Languages, see Table 2.

Table 2. Contributions realized to the speech by a social networking according at different words including turns used by language.

	Volume of Speech		
People	Total of Emitted Words	Total of Turns	Average of words in turn
Italian	788	127	5.9
Spanish	567	93	6.1
English	492	88	4.2

6 Conclusions

There are an important number of questions that deserve additional research. One will be to find new information sources to mine about the use of these three languages and French language during all time of the Mars mission (520 days).

An area with great potential is the electronic usage of media, specifically, digital music [1]. In [6] is shown a system that learns of the user preferences based on the music listened, after songs are selected to be play on a shared physical environment, based on the preferences of the whole people present, this software has a narrative script to realize recommendations to another users in a free text.

Acknowledgements

We want to thank to ISTC-CNR for its economic support to purchase Social Data Mining books, and to permit use Databases related with Twitter users in Italian Language whom was in communication with the first Italian Cosmonaut Diego Urbina Vallessacchi during the emulation of this isolation experiment.

References

1. Amento B. Specifying Preferences based on User History. In Proceedings of CHI'2002, ACM Press. (2002)
2. Fiore T. Visualization Components for persistent Conversations. In Proceedings of CHI'2001. (2001)
3. Padméterakiris, A. & Ochoa A. Implementing of a Data Mining Algorithm for discovering Greek ancestors, using simetry patterns. Central Asia CCBR (Data Mining Workshop); Astana, Kazakhstán. (2005)
4. Pirolli, P. Life, Death and Lawfulness on the Electrical Frontier in Proceedings of CHI'97. (1997)
5. Tabrizi-Nouri H. & Ochoa A. Explain mixtured couples support with Gini Coeficient. CACCBR (Data Mining Workshop); Astana, Kazakhstán. (2005)
6. Toriello, A. & Hill W. Beyond Recommender Systems: Helping People Help Each Other. HCI in the new Millennium, Addison Wesley. (2001)
7. Winograd T. An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests. In Proceedings of CHI'97. (1997)

Implementing mahalanobis distance to select element in a dyoram gift card

Jesús Rodríguez [1], Cecilia Morales[2] & Raúl Holguín [2]

Instituto de Ciencias Sociales y Administración, UACJ; México, Scholarships MCs Planning and Urban Development at Autonomous University of Ciudad Juarez
(Paper received on September 09, 2010, accepted on October 20, 2010)

Abstract. The paper discusses the importance of Social Modelling in the build of a Dyioram Gift Card selecting a specific number of societies based on their features from a repository of 1077 Societies, Mahalanobis Distance determine the correct classification of societies to approach the cultural diversity in the social system described in Memory Alpha, using at formal methodology regarding to the construction of analysis based on social data mining analysis. A case of study is presented regarding to the construction of a Dyoram Gift Card using data obtained from the diversity of cultural patterns described in Memory Alpha. Some futures conjectures and open future analytical works in the Social Modeling studies are described.
The intention of the present research is to apply the computational properties; in this case of corroborating them by means of mining of data to propose the solution to a specific problem, adapted from the modeled Literature about of Societies. Combined to this, we analyzed the selection and location of diverse societies with respect to their social similarity of its neighbors, in a novel & popular representation denominated Dyoram Gift Card. The set of study allowed to analyze the individual characteristics without affecting the resultant Dyoram (what represents the adequate matching), and emulate the distances that separate each one of them and as a set matching characteristic social, linguistics & cultural which specify a position in the Dyoram. By means of this is possible to predict the best position in the Dyoram, redistributing to the individuals that conform it, this article tries to explain this representation of the social behavior.

Keywords: Cultural Algorithms, Pattern Recognition and Social Modelling.

1 Introduction

The Cultural Algorithms (CAs) are an approach of Evolutionary Compute[1], which uses the culture like a vehicle to store excellent and accessible information to all the members of the population during many generations. Like in a human society, the cultural changes act as advances the time, this one provides a line bases for the interpretation and documentation of individual behaviors within a society [2]. CAs were developed to model the evolution of the cultural component on the time and to

Special Issue in Advances in Artificial Intelligence and Applications
Research in Computing Science 51, 2010, pp. 115-121

demonstrate how this one learns and acquires knowledge. In agreement with this conception, the cultural algorithms can be used to lead the process of the self-adaptation within evolutionary systems in a variety of diverse social & cultural areas of application to analyze Textile Heritage, Commerce, Social Networking, Languages, Architecture and Dioramas in different representation [9].

The cultural algorithm base can be described by means of the following pseudo code.

```
Begin
t=0;
Initialize POP(t); /* Initialization of population */
Initialize BLF(t); /* Initialization of believing space */
Repeat
Evaluate POP(t);
Vote (BLF (t), Accept (POP(t))));
Adjust (BLF (t));
Evolve(POP(t), Influence(BLF(t)));
t = t +1;
Select POP(t) from POP(t-1);
Until (Term condition is reached)
End
```

Figure 1. - Pseudo code base of Cultural Algorithms.

Initially a population of individuals that represents the solution space, which is represented like a set of solutions within the space search, is generated randomly to create the first generation. In our example, the solution space contains a list of the attributes that can be used in the classification procedure. The space of beliefs is emptiness. For each generation, CAs will be able to involve a population of individuals using "frame" Vote-Inherit-Promote (VIP). During the phase of Vote of this process, the members of the population are evaluated to identify their contribution to the space of beliefs being used the acceptance function. These beliefs allow contributing in most of the solution of the problem and are selected or put to voting to contribute to the present space of beliefs. The belief space is modified when the inherited beliefs are combined with the beliefs that have been added by the present generation, this is made using a reasoning process that allows updating the space of beliefs. Next, the space of beliefs updated is used to influence in the evolution of the population. The belief space is used to influence on the rest of the population and the acceptance of its beliefs is modified. During the last phase a new population is reproduced using a basic set of evolutionary operators. This new population could be evaluated and the cycle continues successively, until all the population has the same space of beliefs [6]. Cycle VIP finishes when a condition of completion is introduced. The condition of term usually is reached when only a small change or no is detected in the population through several generations or when certain knowledge in the space has emerged from beliefs, as it is possible to be appreciated in Figure 2.

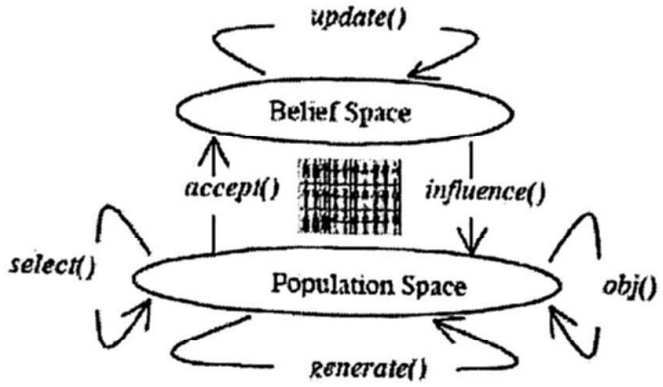

Figure 2. - Conceptual diagram of Cultural Algorithms.

2 Dyoram: the social net representation.

A social network is a social structure that can be represented making use of different types from diagrams. Both more common types are the graph and Dyoram. The graph is a collection of called objects vertices or nodes that are connected by called lines edges or arcs. The nodes represent individuals which sometimes are denominated actors and the edges represent the relations that exist between these actors. The social networks can be classified in: Dyadic, Valued, Transitive &. Directed. The representation of a social network can consist of one or more graphs where these graphs conceptualize the network, that is to say, the representation is made mainly on the basis of the relations that exist between the actors who conform the network. In this article, we focused our attention in a practical problem of the Literature related to the Modeled one of Societies, the select of elements to organize a Dyoram, which allows to include the location that keeps a society with respect to others, the capacity to establish the locations in the Dyoram, allows to establish "the adequate matching with the resultant card gift" for the given set of societies. The solution to this problem could be given by a sequence of generations of agents, denoted like "community". The agents, first realize a matching with the profile of person who was selected to bring this gift, the profile has different attributes with ranges of intensity and magnitude associated with her or his personality, after using Mahalanobis distance is determinate the elements (societies) selected to represent the components of personality associated with the person whom receive the Dyoram Gif card, finally is write a "Narrative Script" which justify the selection of a society according a ons specific behavior in the profile, for example "You are very kin as the Kelemane Society" and the representation show in the Dyoram is kept in a Repository for future new gifts [3]. A Dyoram characterizes a social networking, where the actors conforms the network according to their roll and to the location that each one have within the same. The development of a social networking requires on the one hand, of the conceptual development, and by another one, development of measures of mathematical discreet that allow ontological support to explore the human systems from the data. But it is necessary to prioritize the conceptual development and categories of the system in the social networking, and parallel think about the mathematical model.

3 Distributing elements within a Dyoram Gift Card.

From the point of view of the agents, this problem of optimization is very complex, on account that select a specific number of elements and a location of each one, with respect to the other elements selected. In the algorithm proposed for the cultural change, the individuals in the space of beliefs (beliefscape) through their better paradigm (BestParadigm) are put to zero to represent the fact that the culture increases the amount of expectations associated with the location of a society with respect to the others, giving an incentive to the behavior associated with the best paradigm (BestParadigm). For it we selected 1077 societies described in [4] and characterized their social behavior with base in seven attributes: emotional control, ability to fight, intelligence, agility, force, resistance, and speed, these characteristics allow to describe so much to the society as to the individual. The development of the tool this based on our desire to share the intuitive understanding about the treatment of a new class of systems, individuals able to have empathy, a reserved characteristic in alive people, which will be reactive with its decisions [5].

Formally, the Mahalanobis distance of a multivariate vector $x = (x_1, x_2, x_3, \ldots, x_N)^T$ from a group of values in this case the attributes of each society with mean $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ and covariance matrix S is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \qquad (1)$$

Mahalanobis distance (or "generalized squared interpoint distance" for its squared value) can also be defined as a dissimilarity measure between two random vectors and of the same distribution with the covariance matrix S:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})} \qquad (2)$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called the *normalized Euclidean distance*:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i^2}} \qquad (3)$$

where σ_i is the standard deviation of the x_i over the sample set. The Dyoram Gift Card resulting after to apply Mahalanobis Distance is presented in Figure 3 which included 33 Societies and its Narrative Script justify the selection of each one from the Repository with 1077 Societies.

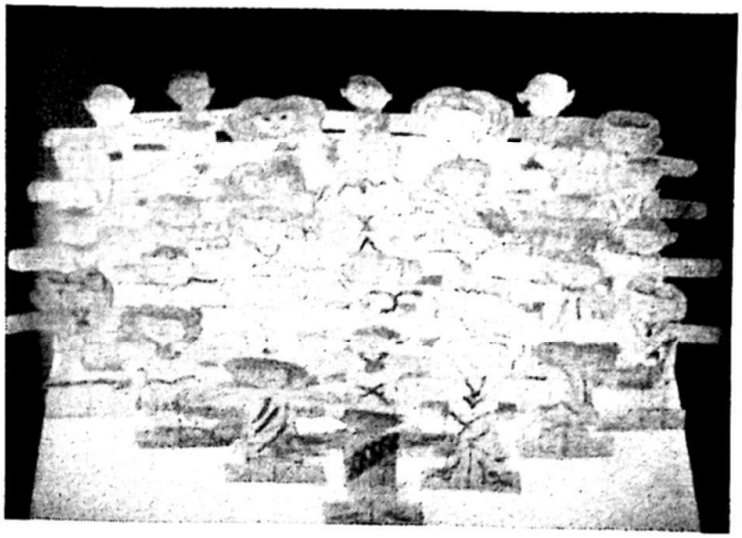

Figure 3. *Dyoram Card Gift resultant using Mahalanobis Distance [10].*

4 Experiments.

In order to be able similar the most efficient arrangement of individuals in a social network conformed a Dyoram Gift Card, we developed an atmosphere able to store the data of each one of the representing individuals of each society, this with the purpose of distributing of an optimal form to each one of the evaluated societies. One of the most interesting characteristics observed in this experiment was the diversity of the cultural patterns established by each community. The scenes structured associated with the agents cannot be reproduced in general, since they only represent a little while dice in the space and time of the different societies. These represent a unique form and innovating of adaptive behavior which solves a computational problem that it does not try to clustering the societies only with a factor associated with his external appearance (genotype), trying to solve a computational problem that involves a complex change between the existing relations. The generated configurations can be metaphorically related to the knowledge of the behavior of the community with respect to an optimization problem (to conform to cluster culturally with other similar societies, without being of the same quadrant [4]).

The main experiment consisted of detailing each one of the 87 communities, with 500 agents, and one condition of unemployment of 50 generations, this allowed us to generate different scenes from best dyoram possible, which was obtained after comparing different cultural and social similarities in each community, and to determine the existing relations between each one in relation with the Mahalanobis Distance (the colors indicated a cluster with specific intensity, the size of cluster determine the magnitude of societies) from the profile of person whom receive the gift. The developed tool classified each one in a different location according to the cluster more closely at profile used (To see Figure 4).

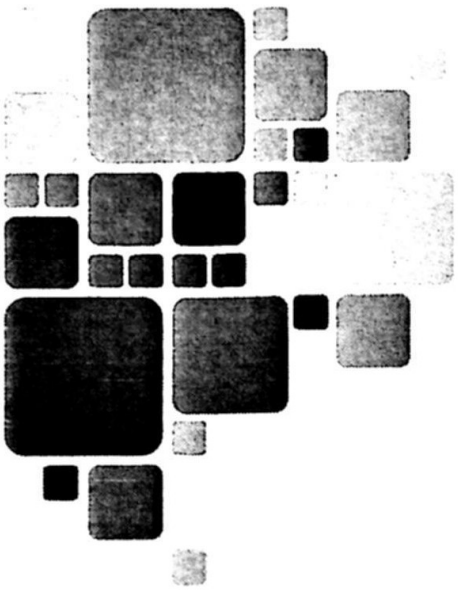

Figure 4.Clusters constructed by means of the use of Cultural Algorithms and Mahalanobis Distance [10].

5 Conclusions

Using Cultural Algorithms we improved the understanding substantially to obtain the change of "best paradigm", because we appropriately classified the agent communities basing to us on an approach to the relation that keep their attributes, this allowed us to understand that the concept of "selection and matching" exists with base in the determination of the function of acceptance on the part of the rest from the communities to the propose location for the rest of same ones. The Cultural Algorithms offer a powerful alternative for optimization problems and redistribution of clustering. For that reason, this technique provides a quite comprehensible panorama with the cultural phenomenon represented [8]. This technique allowed us to include the possibility of generating experimental knowledge created by the community of agents for a given dominion of application.

The analysis of the level and degree of cognitive knowledge of each community is an aspect that is desired to evaluate for the future work. The answer can reside between the similarity that exists in the communication between two different cultures and as these are perceived. On the other hand to understand the true similarities that have different societies with base in the characteristics that make them contributor of cluster and it as well allows him to keep his own identity, demonstrates that the small variations go beyond phenotypes characteristics, and are mainly associate to tastes and similar characteristics developed through the time [7].

Importantly, the cultural algorithm (CA) is a powerful tool, yet neglects various elements of cultural analysis, this being an opportunity to innovate new algorithms rescuing the complexity and the chaotic social and cultural relations.

A new Artificial Intelligence can take care to analyze retail these complexities that each society keeps, without forgetting that still they need to us methods to understand original and the particular thing of each society.

Also this experiment opens the possibility to analyze in future work, how to minimize to one society per quadrant to make a four societies bookmark and get a minimum unit form of visual representation of a diorama, using a grand prix model.

Acknowledges

We appreciate the permission of Alberto Ochoa Ortiz Zezzatti to use various published materials related with Cultural Algorithms.

This research was supported in part with Conacyt Scholarships from the MCs Planning and Urban Development at Autonomous University of Ciudad Juarez: Cecilia Morales and Raúl Holguín.

6 References

1. This algorithm differs from genetic algorithms (GA) in the appearance of the (GA) is a heuristic search mechanism, where adaptation and evolution based on natural selection.
2. Desmond, A. & Moore J. (1995). "Darwin – la vida de un evolucionista atormentado". Generación Editorial, São Paulo, Brazil.
3. Ochoa A. et al. (2007) "Baharastar – Simulador de Algoritmos Culturales para la Minería de Datos Social". In Proceedings of COMCEV'2007.
4. Memory Alpha (2007). memory-alpha.org (Star Trek World).
5. Callogerodóttir Z. & Ochoa A. (2007) "Optimization Problem Solving using Preda tor/Prey Games and Cultural Algorithms" NDAM'2003, Reykiavik; Iceland.
6. Tang Hué et al. (2006) "The Emergence of Social Network Hierarchy Using Cultural Algorithms", VLDB'06, Seoul, Korea.
7. Vukčević I. & Ochoa A. (2005) "Similar cultural relationships in Montenegro" JASSS'2005, England.
8. Zuckermann Dennis (1991). "Culture and Organizations", London: McGraw-Hill.
9. Correa Christian et al. (2007) "Algoritmo Transgénico", In Proceedings of COMCEV 2007 Aguascalientes Ags.
10. Ochoa A. et al. (2009). "Dyoram's representation using mosaic image". The International Journal of Virtual Reality, 8, 1-4