Sociolinguistics analysis of a social networking conformed by Italian ethnicity

Alberto Ochoa-Zezzatti[1,2], Sandra Bustillos[1], Iván Álvarez[1], Alessadra Tagliarducci[2] , Arturo Hernández-[3] & Rubén Jaramillo[4]

[1.] Instituto de Ciencias Sociales y Administración, UACJ; México
[2.] ISTC-CNR, Rome; Italy
[3] CIMAT, México.
[4] CIATEC Centro Conacyt (Doctoral Student), México.
cbr_lad7@yahoo.com.mx [1]

Abstract. The present paper discusses an investigation related to Sociolinguistics using WEKA, a tool that mine information of the structure and content of speech of Italians and Italians descendents with the purpose of discovering hypostasis and parataxis, which consists of relation in formal and informal use of language induced by the relation with another speakers, this phenomena has been documented recently, but with few detailed research with truly information, for this purpose we record speeches in a social networking whose are scholarship holders of the "RAI Internazionale" and participants in Miss Italia nel Mondo 2009, to explore a detailed sociolinguistics analysis.

Keywords Sociolinguistics, Data mining, Modeling of societies.

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?

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 describe how two groups of individuals with common ancestors can have similar conversations in a same language. A task to manage infor-

mation 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 ethnic-social group, familiar 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 or 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].

3 System Development

The system will be able to analyze the behavior for two samples of the Italian Communities, from the information of a sample of RAI Internationazionale scholarship holders and another with participants in Miss Italia nel Mondo, 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, we determine if this behavior was an innate or induced tendency by their family of Italian origin.
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].

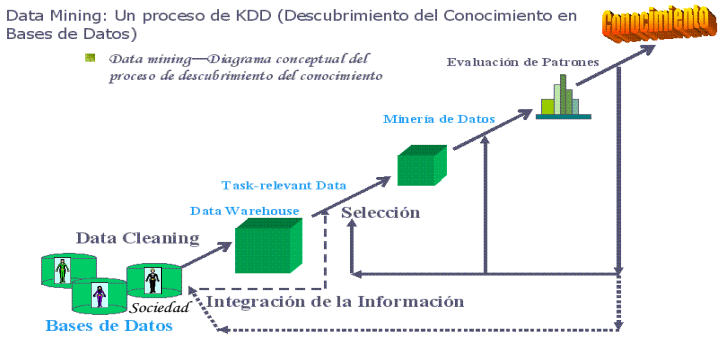

Fig. 1. Data Mining process. The society information inside a *Data bases* is cleaned and stored in a *Data Ware House*, then is mined by means of a loop back *selection* and *patterns evaluation* process processes.

4 Applied tool

Use Data mining tool WEKA to analyze data. First, we proceed to develop a model that allows explain the behavior by two samples of people, and how affects their speech style. Figure 2 and 3 discover the existent relation among hypostasis and parataxis parameters.

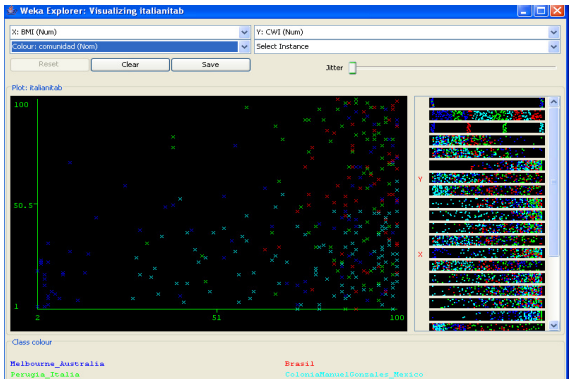

Fig. 2. WEKA justifying the relation among Hypostasis in Italian music

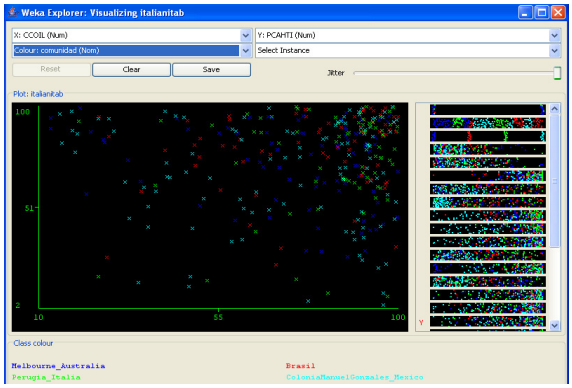

Fig. 3. Relation of Parataxis in users from an Italian Language Chat.

We found in both cases that the RAI scholarship holders showed a higher hypostasis and lowest parataxis regarding participants in Miss Italia nel Mondo. This can be explained by the use of informal speech of Italian because they resist losing their ancestors customs, and purchase decision is highly influenced induced by their relatives.

5 Results

We took in consideration RAI Internazionale scholarship holders from Italy (Sample 1 – Female & Male) and Sample 2 from Participants of Miss Italia nel Mondo 2009, and using their conversations in a social networking, to identify different behaviors (See Table 1).

Table 1. Distributions of demands by category and sort of the 2 analyzed samples.

	Sample 1		Sample 2
Category	F	M	F
N	43	47	50
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 with common ancestors, we found variations depending on the use of Italian Language, see Table 2.

Table 2. Contributions to the speech in a social network to Italian origin between two samples.

	Volume of Speech		
People	Total of Emitted Words	Total of Turns	Average of words in turn
Sample 1 (Male)	788	127	4.9
Sample 1 (Female)	567	93	6.1
Sample 2	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 Italian Language.

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 the use of Databases related with Miss Italia nel Mondo organized in Jesolo, Italy during July, 2009.

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)