Italianitá: Discovering a Pygmalion effect on Italian Communities Using Data Mining

Alberto Ochoa[1,2], Alán Tcherassi[2], Inna Shingareva[3], A. Padméterakiris[4], J. Gyllenhaale[5] &, José Alberto Hernández[6]

1. Facultad de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Av. Ramón López Velarde #801; C.P. 98000 Zacatecas, México,
2. Computer Institute (Postdoctorale Program), State University of Campinas; Postal Box 6176, 13084-971 Radamaelli – SP, Brazil.
3. Artificial Intelligence Institute, Kazaksthan University; Astana, Kazaksthán.
4. Larissa University; Larissa, Grecia.
5. Manx University; Ramsey, Man Island.
6. Centro de Investigación en Ingeniería y Ciencias Aplicadas, Universidad Autónoma del Estado de Morelos; México

cbr_lad7@yahoo.com.mx [1]

Abstract. The present paper discusses an investigation related to the Social Data Mining field using WEKA, a tool that mine information of the structure and content of activities made by descendants of Italians with the purpose of discovering a Pygmalion Effect, which consists of a conduct change of a group that shares similar characteristics induced by the expectations of the same one, this phenomena has been documented since the Sixties, but with few detailed research with truly information, for this purpose we applied a questionnaire to people of four Italian communities whose are scholarship holders of the "RAI Internazionale", to explore their daily activities made on the Internet.

Keywords Pygmalion Effect, 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 like Usenet messages, log files, purchasing records and links of interest. 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?

© *A. Gelbukh, S. Torres, I. López (Eds.)*
Advances in Computer Science and Engineering
Research in Computing Science 19, 2006, pp. 57-68

We made an analysis that provides positive answers for both questions. First, a number of attributes about web sites give us as a result the prediction of the behavior on the use of specific computer skills.

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 sixties decade appears the first serious studies to understand the Pygmalion effect, which try to demonstrate how "normal" people are induced to behave in a different way, when they show pertaining to a particular group.

In this paper we try to describe how four groups of individuals with common ancestors can make computational activities and web purchases in a similar way. 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. Users can create information storage collections in the web for personnel use or to share with partners at work or with friends.

Our approach to this problem combines social data mining [20] with information about work spaces [4]. As the cluster of this People in Web [13], follows certain patterns, this can be analyzed by means of these techniques. 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/near 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 [16] (answering the question: Which are the most visited sites for the most of people?), online conversations [24] (Which are the sites where people purchase "thematic" things or for a community?), or web log files [13] (Which sites are the most visited?). By means of social data mining is taken the final move.

This paper is organized in five sections. In section one, we introduce our paper. En section two, we describe the ethnic-social effect called "Pygmalion Effect", we describe how can be discovered using data mining, we describe an approach named "Social Data Mining" also. 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 The Pygmalion Effect

By the end of the Sixties, a professor of psychology called Robert Rosenthal, made the following experiment: joined the teachers of a school and showed them a test made among the students, which indicates that some students were more "shining" than others. "Of these students we can wait for great results ", assured to them. In fact - and responding to the objectives of the experiment- that test was simulated by Rosenthal [17], to induce the teachers to think that certain students had more potential that the rest. Nevertheless, after eight months, those students indeed obtained better qualifications than the average of the class. Like teachers believed in "the supposedly shining" students, offered to them more attention, support, time and feedback. This abundance of conditions was soon translated in a better learning and - in better qualifications. Those children did not stand out being intelligent, but because their teachers believed that they were. Through its experiment, Rosenthal discovered that the expectations of the teachers were reflected in the performance of the students. His conclusion was the following one: while higher are the expectations that a person has with respect to other, more probable than this last one obtains positive results. This discovery put in evidence a phenomenon that is known with the name of "The Pygmalion Effect".

2.1 Data Mining

Data Mining, is the extraction of hiding and predictable information inside great data bases, is a powerful new technology with great potential to help to the companies or organizations to focus on the most important information in their Bases of Information (Data Warehouse). Data Mining tools predict future tendencies and behaviors, allowing businesses to make proactive decisions leaded by knowledge-driven information.

The automated prospective analyses offered by a product thus go beyond past events provided by retrospective typical tools of decision support systems. Data Mining tools can respond to questions of businesses that traditionally consume too much time to be solved and to which the users of this information almost are not willing to accept. These tools explore the data bases searching for hidden patterns, finding predictable information that sometimes an expert cannot find because this is outside expectations.

2.2 Justification

Most of the social groups, that immigrate to another country form communities whose share common characteristics.

As time pass, these characteristics are reinforced if the number of members is considerable, or are assimilated by a greater group [9]. Due that the Pygmalion effect is not considered completely an ethnic-social effect, the necessity to propose, the analysis of the information using Data Mining, with study aims, has allowed to discover

the "Italianitá" that they thought they had, and how this marked their activities and forms of using the Web.

2.3 Data Mining Applications in Social Aspects

One of the most transcendental aspects of the use of Data mining is denominated Social Data Mining, which tries to find different patterns in predefined clusters in the network, like the groups of discussion, Usenets, thematic chats among others. Other work has been focused on extracting information about online conversations such as the USENET PHOAKS [19] mining messages in the USENET newsgroup that recommend Web sites. Categorizing the users mentions to create lists of popular Web sites for each group. Where? [22] Has been analyzed the newsgroup information and the Usenet conversations and if they have been used to create visualizations of the conversations. These visualizations can be used to find conversations with the desirable characteristics, such as equality of participation or regular participants. In [6] also was extracted information of newsgroups and visualizations of the conversation subject, contributions of individual messages, and the relation among them were designed. Another research has been centered in extracting the information of web user records. The Log files [23] register information of the users, analyze this to find common connections between Web pages, and they construct diverse visualizations of these data to help user navigation through Web sites. Persecuting the navigation metaphor, some investigators have used the term "social navigation" in order to characterize the work of this nature [11]. Finally, a different technical approach [2] uses the register of activity - e.g., a sequence of visited URLs during a session like the basic unit. Based on this, they have developed techniques to calculate similarities between the trajectories of sequences and to make recommendations - for example, to similar pages to the visited ones.

2.3.1 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 [3]. 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 [21]. The social tendency reflects the history of the use of a collective behavior, and serves like base to characterize the behavior of future descendants [8]. The Data Mining approaches for social aspects look for analogous situations in the behavior registers [14]. The investigators look for situations where the groups of people are producing computer registers (such as documents, USENET messages, or Web sites and links to groups with a specific profile) like part of its normal activity. The potentially useful information implicit in these files is identified; and the computer techniques to display the results are designed. Thus the computer discovers and makes explicit the "social tendencies through the time" created by a particular type of community.

The systems that analyze social aspects with Data Mining do not require expert users in no new activity, due to this, the investigators in the subject try to explore the information of the users preference implicit in the existing activity registers.

3 System Development

The system will be able to analyze the behavior for each one of the samples of the Italian Communities, from the information of the RAI Internazionale scholarship holders, by means of WEKA use, which has demonstrated being an efficient tool for searching hiding parameters that must be discovered [18]. The compiled information was analyzed to discover behavior patterns that share these individuals, and based on their gender and age, we determine if this behavior was an innate or induced tendency by their family of Italian origin.

3.1 Methodology

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. Both processes require to examine an immense amount of material, or to investigate intelligently until finding exactly where the values reside (see Figure 1). If data bases of sufficient size and quality are available, the Data Mining technology can generate new opportunities of interpretation when providing these capacities.

3.1.1 Automatic Prediction of Tendencies and Behavior.

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 the data [14].

A typical example of a predictable problem is the marketing oriented to objectives (targeted marketing). Data mining uses data derived of previous promotional mailing campaigns to identify possible objectives to maximize results of the investment in future mailing. Other predictable problems include forecasting of future financial problems and other forms of breach, and identify the population segments that respond probably to similar events.

3.1.2 Automatic Discovering of Previously Unknown Models

The Data Mining Data tools sweep the data bases and identify previously hidden models in only one step. Other problems for models discovering include the detection of credit card fraudulent transactions and to identify abnormal data that can represent keypunch errors in the load of data.

The Data mining techniques can generate benefits for the automatization of existing hardware and software platforms and can be implemented into new systems as the existing platforms get updated and new products are developed[7].

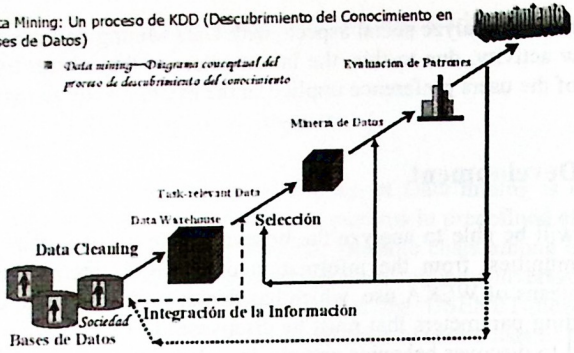

Fig. 1. Data Mining process. The society information inside a *Data bases* is cleaned and stored in a *Data Ware House*, then is mined by means of a loop back *selection* and *patterns evaluation* process processes

When the Data mining tools are implemented in high performance parallel processing systems, can analyze massive data bases in just few minutes. Faster processing means that the users can automatically experiment with more models to understand complex data [5]. High speed is practical for the user and makes possible to analyze immense amounts of data. The great data bases, as well, can produce better predictions.

The data bases can be huge as well on depth as well as on width.

More columns. So many times analysts must limit the number of variables to examine when manual analysis are done due limitations on time. However, variables that are suppressed because they seem without importance can provide information about unknown models. A high performance Data mining allows users to explore the whole data base, without a set of variables preselection [10].

More rows. Bigger samples produce less estimation errors and deflections, and allow users to make inferences about small but important population segments.

4 Applied Tool

We use a Data mining tool called WEKA to analyze data. First, we proceed to develop a model that allows explain the behavior showed by the four Italian communities, and how affects their computer activities and therefore their likes and purchase intention on the web. Figure 2 and 3 shows WEKA usage to discover the existent relation among four parameters associated to Italianitá.

We found in both cases that the RAI scholarship holders outside Italy showed a higher "Italianitá" regarding native Italians.

This can be explained by the Pygmalion effect because they resist losing their ancestors customs, and purchase decision is highly influenced by this effect induced by their relatives.

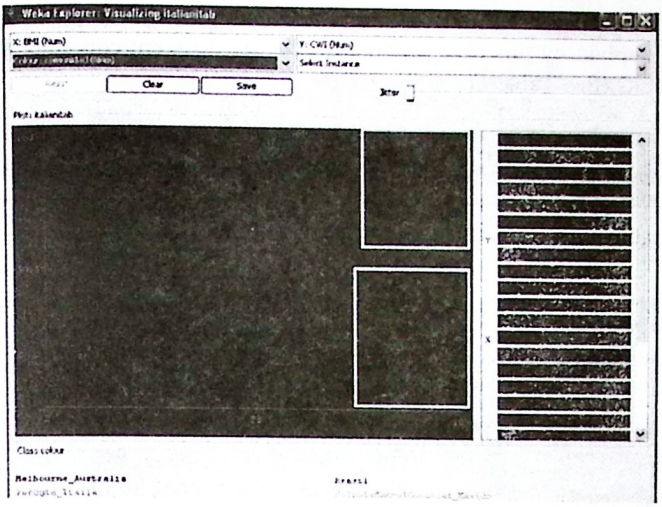

Fig. 2. WEKA justifying the relation among the "some Italian" web site creation with the relation to download Italian music (superior cluster). Users that download music but do not have the intention to create a "some Italian" web site form another group (inferior cluster)

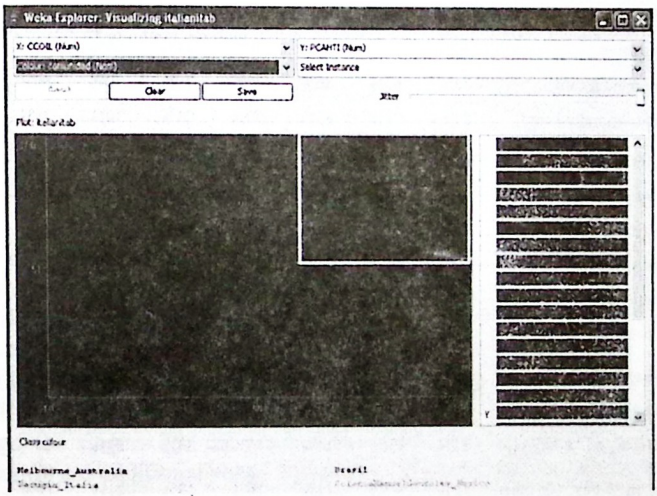

Fig. 3. Shows the relation to participate on a Chat on Italian with the purchase of items of Italian Origin

5 Results

We took in consideration RAI Internazionale scholarship holders of four Italian communities: Sample 1 (Melbourne, Australia), Sample 2 (Radamelli, Brasil), Sample 3

(Perugia, Italy) and Sample 4 (Manuel González Colony, Mexico City), to whose an instrument was applied by this organization, to identify different computer activities and purchase habits (See Table 1).

Table 1. Differences on computer skills by gender for the four Italian communities studied

	Sample 1		Sample 2		Sample 3		Sample 4	
	F	M	F	M	F	M	F	M
n	36	46	29	43	21	56	72	35
Online purchase of Italian books	22%	22%	21%	26%	29%	27%	14%	11%
Online purchase of any book	34%	31%	39%	21%	38%	37%	25%	26%
Having a PC	84%	87%	91%	80%	100%	96%	89%	91%
Creation of a "some Italian" web site	25%	50%	43%	63%	38%	66%	24%	29%
Write a Java Programm	39%	54%	4%	56%	29%	59%	7%	11%
Prepare a Power Point Presentation	78%	85%	75%	81%	95%	91%	66%	66%
Download documents on Italian with Acrobat	83%	93%	82%	98%	76%	93%	56%	66%
Sending photographs to relatives in Italy by means of e-mail	69%	91%	71%	79%	71%	87%	64%	66%
Install extra memory	31%	48%	29%	56%	48%	71%	21%	29%
Download Italian Music	72%	89%	89%	91%	81%	87%	80%	74%
Install an extra floppy drive	19%	57%	32%	40%	24%	66%	19%	20%
Send a static/silence greeting card	83%	67%	86%	74%	76%	70%	80%	60%
Send a animated/musical greeting card	81%	77%	89%	70%	76%	70%	79%	46%
Participate on Chats on Italian	61%	87%	75%	86%	71%	80%	73%	69%
Send an attachment by e-mail	91%	84%	100%	100%	100%	100%	87%	86%
Installation of a computers network	11%	17%	14%	44%	19%	60%	6%	6%
Purchase an italian origin item	80%	72%	71%	86%	71%	91%	76%	66%
Upgrade the PC's operating system	45%	47%	43%	72%	29%	79%	37%	34%
Research for online papers/assesments	88%	93%	100%	98%	90%	100%	98%	91%
Defragmentation of hard disk	48%	45%	75%	88%	52%	80%	47%	57%
Send a movie/video by e-mail	27%	33%	39%	65%	100%	62%	15%	31%
Purchase anything italian on "e-bay" (or other site)	35%	29%	18%	44%	33%	59%	28%	31%
Sell anything italian on "e-bay" (or other site)	19%	14%	11%	16%	19%	21%	10%	6%
Sum of PC Knowlege and Italianitá (mean)	5.6	7.2	12	14	11.1	14.9	9.7	9.4

The use of Data mining in social aspects has demonstrated being key part to corroborate the tendencies of a group with common ancestors (Pygmalion Effect), although on each group we identified factors that distorted the data analyzed in the answers (the factor of Lying is greater in women than in men), we found variations depending on the Italian community origin, see Table 2.

Table 2. Predictors to do computer activities and online purchase of books or items of Italian origin

Study 1	Women	Men
Skills for PC/Internet	0.43	0.15
Extraversion	0.23	0.10
Neuroticism	-0.30	0.19

Study 2	Women	Men
Skills for PC/Internet	0.44	0.26
Attitudes toward money:		
Power	0.14	0.04
Retention	0.06	0.33
Un confidence	-0.13	0.21
Anxiety	-0.25	0.26

Study 3	Women	Men
Skills for the PC/Internet	0.48	0.35
Psychopath	0.37	-0.15
Extroversion	-0.08	0.01
Neuroticism	0.21	-0.06
Factor of Lying	0.23	0.10

Study 4	Women	Men
Skills for PC/Internet	-0.02	0.11
Computers Anxiety	-0.07	0.21
Computers Attitude	0.07	0.21
Attitudes for the Internet	-0.03	0.04

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 users preferences. As we discuss earlier, researchers have investigated the hyperlinks structure, the electronic conversations and users' purchase records [12].

An area with great potential is the electronic usage of media, specifically, digital music. By analyzing what kind of music is someone listening, a system can deduce the songs, the singers and the genders the person prefers, and by using this information recommend additional songs and artists, to get the person in touch with people of similar interests. We made an approach on this direction with a system that allows users to view individual and group historical listening lists and define with this information new listening lists [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. Meanwhile the user preferences are extracted from a large number of sources; the idea to combine different types of preferences starts to be important. In PHOAKS [19] preferences are extracted from web pages since USENET messages and then presented to the users. Showing how the users visualize this information. PHOAKS keeps the track on what pages the users did click (other type of implicit preference).

Development of general techniques to combine different types of preferences is now a challenge. Panzanni [15] presents a method to give weight to different types of contributions, however, if this is the best combination of methods and how to determine the proper weights is still a complex idea. Such system will combine the people advantages – applying the judgment to select the initial system of collections – and of computers to apply analysis of techniques to provide remarked information and to store updated collections. A similar tactic will be to use a search engine. Finally, this discussion shows that even a very large system, manually constructed from "base" pages can be improved perceptibly by providing additional characteristics, grouping pages on sites, and offering a user friendly interface.

7 Future Works

We are planning to apply a similar methodology to identify Mexican way of being, attitudes and purchasing habits over the Internet from Mexicans living abroad, specifically in the United States and in the European Union. They represent more than thirty million persons, almost a quarter of the total Mexican population, that represents the first international income for the Mexican Economy and a very interesting target market to explore for business opportunities.

By using a different instrument and samples from different places of the world, we are planning to compare two societies without sea and with a high migration level. Our basic question is: Can these societies develop similar behaviours?

Acknowledgements

We want to thank to Mendoza R & Rodríguez L. for sharing their Pygmalion Effect Model called "Talucicé" that allows explain beliefs on certain societies about the relation of the birds singing and yellow flowers with death relatives and about bio luminescent insects with divine messages like the Catholicism of Oriental Timor. To Ochoa P. for his economic support to purchase Social Data Mining books.

References

1. Amento B. Specifying Preferences based on User History. In Proceedings of CHI'2002, ACM Press. (2002)
2. Broedbeck K. The order of things: Activity-Centered Information Access. In Proceedings 7thICWWW'98. (1998)
3. Bush, V. As we may think. The Atlantic Monthly. (July 1945).
4. Card K. et al. The Information Visualizer, an Information Workspace for the World-Wide Web. CHI'96. (1996)

5. Daurov T. & Sebastianni M. Modelling Kazakh costumes using data mining. CA CCBR; Astana, Kazakhstán. (2005)
6. Fiore T. Visualization Components for persistent Conversations. In Proceedings of CHI'2001. (2001)
7. Han Jiawei, Implementing data mining for discover conduct patterns, Am-Psychol, Jan 32[1]: (2001) 57-66.
8. Hé Z. .& Milodragovich K. Discovering chinese descendents in Palé Island using Data Mining. CACCBR; Astana, Kazakhstán. (2005)
9. Logan S. Discovering induced social patterns using an Intelligent Dyoram for displayed.CACCBR; Kazakhstán. (2005)
10. Maes P. Social Information Filtering: Algorithms for Automating <<Word of Mouth>>. In Proceedings of CHI'95. (1995)
11. Munro J. & Höök K. Social Navigation of Information Space. Springer, (1999)
12. Nieto M.; Ochoa A. Applying dependences model to Data Mining software.CIIC-'02;Soto La Marina, México. (2002)
13. Oki, B; Momoi, Kaori & Zhang Ziyi. Collaborative Filtering to Weawe an Information Tapestry. Communications of the ACM, 35, (1992) 51-60.
14. Padméterakiris, A.; Gyllenhaal, J. & Ochoa A. Implementing of a Data Mining Algorithm for discovering Greek ancestors, using simetry patterns. Central Asia CCBR (Data Mining Workshop); Astana, Kazakhstán. (2005)
15. Pazzani M. & Diggory Cedric Learning Collaborative Information Filters. In Proceedings of ICML'98. (1998)
16. Pirolli, P. Life, Death and Lawfulness on the Electrical Frontier in Proceedings of CHI'97. (1997)
17. Rosenthal Robert Explain the Pygmalion effect in the school. Hamburg, Germany. (1971)
18. Tabrizi-Nouri H.; Tañón O.; Ianevski S. & Ochoa A. Explain mixtured couples support with Gini Coeficient. CACCBR (Data Mining Workshop); Astana, Kazakhstán. (2005)
19. Terveen L. Using Frequency-of-Mention in Public Conversations for Social Filtering. Proceedings CSCW'96. (1996)
20. Tochi K. & Amento B. Experiments in Social Data Mining: The TopicShow System In Proceedings CHI'03. (2003)
21. Toriello, A. & Hill W. Beyond Recommender Systems: Helping People Help Each Other. HCI in the new Millennium, Addison Wesley. (2001)
22. Viegas F. Chat circles. In Proceedings of CHI'99, ACM Press, (1999), 9-16.
23. Wexelblat, P. Footprints: History-Rich Tools for Information Foraging. In Proceedings of CHI'99. (1999)
24. Winograd T. An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests. In Proceedings of CHI'97 (1997)