Creation of a competitiveness index based on sociolinguistic factors

Alejandro Lara Enriquez, Alberto Ochoa and Jorge Rodas-Osollo
Maestría en Cómputo Aplicado.
Instituto de Ingeniería y Tecnología.
Universidad Autónoma de Ciudad Juárez

Abstract. Sociolinguistics has become an increasingly important and popular field of study, as certain cultures around the world expand their communication base and intergroup and interpersonal relations take on escalating significance. Language use symbolically represents fundamental dimensions of social behavior and human interaction. The creation of the competitiveness index started as mean to validate, using social data mining, the influence of sociolinguistic factors among a specific linguistic group. To prove this theoretical model, structural equation model was chosen because of its ability to isolate observational error from measurement of latent variables.

We present a prototype Web based Decision Support System with data mining capabilities. The purpose of the presented system is to analyze differed social variables to determine specific indicators associated with European Parliament.

Original code was developed in Java for an intelligent agent to monitor main changes in diverse societies using data from different databases about competiveness based on information obtained from diverse organizations websites on the Internet. We conducted an experiment using our prototype system applied to European societies and their public policies, a region with a high development during the last 50 years. Preliminary results show the system could be used to model competiveness based on historical information and to identify critical future scenarios.

This system can serve as a base for the development of a prediction model. Using an analysis carried out in the translation of Natural Language Queries in Spanish to SQL involving the clause of grouping GROUP BY in Natural Language Interfaces to Databases (NLIBDs), the important role and the different ways to find them in the Natural Language. **Key words**: social data mining, sociolinguistics, and competitiveness

Keywords. Web based DSS, Decision Support, Processing of Natural Language, social data mining, sociolinguistics, and competitiveness.

1. Introduction

The European Parliament (EP, European parliament also colloquially or EP) is the parliamentary institution in the European Union (EU) directly represents the citizens of the Union and with the European Commission and the EU Council exercises the legislative function. Described as one of the world's most powerful legislators, the European Parliament is composed of 754 members representing the second largest democratic electorate in the world (after the Parliament of India) and the largest trans-national electorate (375 million eligible voters in 2009). Also, is the only institution directly elected by the citizens in the European Union.

He has been elected by universal, direct and secret suffrage every five years since 1979. However, participation in the European elections has fallen consecutively in each ballot from that date, and has been below 50% since 1999. In the last elections in 2009, voter turnout stood at 43% of European citizens are entitled, ranging from 90% in Member States such as Luxembourg and Belgium (where voting is compulsory) and 20% in Slovakia. The turnout was lower than 50% in 18 of the 27 Member States. The Parliament is considered the "first institution" of the European Union is mentioned first in the treaties and its President takes precedence over all other ceremonial authorities at European level which shares with the Council the legislative and budget, taking control over the European Union budget. The European Commission, the executive body of the EU, is accountable to Parliament. Specifically, the European Parliament elects the President of the Commission, approves (or rejects) the appointment of the Commission as a whole, and even as the body can remove her filing a motion of censure.

The current President of the European Parliament Martin Schulz is the Social Democrat who was elected in January 2012 and headed a chamber composed of a variety of parties associated in

groups. The two main groups in the European Parliament (together own 61% of the seats) are the European People's Party and the Group of the Alliance of Socialists and Democrats Progressive.
Since the founding of Parliament in 1952, its powers were extended several times, especially through the Maastricht Treaty and the recent in 1992 the Lisbon Treaty in 2007.
The European Parliament has two meeting places: The Louise Weiss building in Strasbourg, France, in which twelve plenary sessions are held four days a year and is the official seat of Parliament, and the complex of buildings of Space Léopold in Brussels, Belgium, which is the larger of the two and serves for committee meetings, political groups and complementary plenary sessions. The General Secretariat of the European Parliament for its part, the administrative body, is based in Luxembourg.
A part of the European parliament is chosen based on work proposals submitted from across Europe, not just the 27 plus Croatia which begins in January 2013 to join the European Union. Proposals will be ranked according to their total score and then vote for the other European representatives. Many contests competing for a better place as Eurovision have been studied with different perspectives: the compatibility between countries and political and cultural structures of Europe [4], the persistent structure of hegemony in the Festival of the Eurovision Song Contest [5], voting culture [6] and the analysis of the Grand Prix that evaluates many countries participating in different years and with many different types of countries competing with each other [7], among others. This research is novel because the type of behavior analysis when people from different cultures proposals are analyzed and participate in this setting mechanism proposed for the whole European continent. The objective is to estimate the final ranking of the proposals. The organization of this paper is as follows. The analysis of the 30 calls for proposals for incorporating a priori knowledge about voting patterns and relationships among potential winners is explained in Section 2. Then the problem statement is defined in Section 3. COPSO The algorithm is explained in detail in Section 4. In Section 5, our approach has been tested in the Call for Proposals 2011, with proposals from 43 countries including Israel. Experiments and analysis to estimate the classification of a specific proposal in the competition for proposals are explained in Section 6. The conclusions are set out in Section 7.

2. Analysis using Data Mining Proposals

Data mining is the search for global patterns and relationships between data in huge databases, but they are hidden within the vast amount of information stored in these repositories of information [3]. These relationships represent the knowledge of the value of the objects in the database. This information is not necessarily a true copy of the information stored in the databases. Rather, it is information that can be inferred from the database. One of the main problems in data mining is that the number of possible relationships extracted is exponential [2]. Therefore, there are a variety of machine learning heuristics have been proposed to the knowledge discovery in databases [2]. One of the most popular methods for representing the results of data mining is the use of decision trees. A decision tree provides a method for recognizing a given case for a concept. It is a "divide and conquer" strategy for the acquisition of the concept (example). Decision trees have been useful for a variety of case studies in science and engineering, in our case we use data mining to characterize the behavior of each country's historic election related to accepting proposals. Therefore, we selected the companies that have participated and characterized their behavior based on their votes already cast, allowing to describe both society and the individual's behavior. The purpose is to explain υij, voting (ie, the number of points) issued by the country's society i ≠ L in evaluating the performance of a public policy j ≠ L (i ≠ j, as a society can only vote for one specific proposal), where L is the total number of entries submitted from across Europe. Regardless of any other characteristic, the equation could be written simply vote

$$\upsilon ji\ uij = \alpha\ ij\upsilon ij + \quad (1)$$

Where α ij is a parameter υij commitment and a random disturbance. If the exchange of vows was "perfect", and any proposal to maintain its ability to receive votes α ij would equal 1. More

generally, this type of equation should contain variables k = {1, ..., K α representing characteristics (feasibility, proper timing etc.) of a given i, and variables that represent the different attributes of the i this proposal over its involvement in sending proposals to the European parliament.

$$v_{ji} = \alpha_{ij} v_{ij} + \beta \sum_{k=1}^{K} x_{ik} + \gamma \sum_{t=1}^{T_i} z_{it} + u_{ij} \tag{2}$$

where β and γ are parameters to be estimated. The party associated with the beta parameter is related to the performance attributes of a proposal (A useful public policy for the Company). The party associated with the gamma parameter is related to the performance of these proposals during the assessments in the European parliament. One problem has to do with the fact of what you want to calculate the part of the equation for the comment on the vote of a company i to j represents the proposal for a specific country.

This can be treated in several ways. First, and this is the easiest way, instead of using υij on the right side, you can use the vote in the previous proposal evaluation, say υij-1, although one might think that societies do not necessarily maintain its time commitment. An alternative is to use only half of the observations along all editions assessments proposed in Europe, therefore, υij appears on the right side of the equation is not used while the left side. The vote equation is estimated by linear methods. The influence on the order in which they appear in the list of proposals often described. The exogenous order in which proposals are made is included as a factor. Other variables include (a) a factor of innovation for new proposals, this variable is set to 1 for the person who submits a proposal from the same approach as a proposal for a similar country-, (b) the language in which the proposals is presented, (c) interest nature of the proposal such as being of ecological, and (d) whether the proposed uses specialized pubic policy.

The last group of variables includes linguistic and cultural distances between voters and proposals, and we can afford to dispense with the use of variables that characterize voters. Cultural differences among societies are represented by the four dimensions studied in [1]. These studies have identified the following four dimensions that explain the "cultural distance":

(A) Power Distance: Measures the extent to which the less powerful members of a society accept that power is distributed unequally, but focuses on the degree of equality between individuals;

(B) Individualism: Measures the degree to which individuals in a society are integrated into groups, but focuses on the degree of how a society reinforces individual or collective achievement and interpersonal relationships;

(C) Masculinity: refers to the distribution of gender roles in society, but focuses on the degree a society reinforces the traditional role of male labor male achievement, control and power;

(D) To avoid uncertainty: It is about a society's tolerance for uncertainty and ambiguity, and refers to man's search for truth.

Table 1: Correlations between Cultural Distances and Linguistic

	Language	Power	Indiv.	Masc.	U. A.
Language	1				
Power	0.205	1			
Indiv.	0.254	0.111	1		
Masc.	-0.092	0.031	-0.128	1	
U. A.	0.319	0.567	0.404	0.083	1

Table 2: Cultural Distances vs Contender Characteristics

	(a)	(b)	(c)	(d)
Quality	0.911	0.914	0.901	0.905
	(0.03)	(0.03)	(0.03)	(0.03)
Logrolling	0.028	0.022	0.018	0.016
	(0.01)	(0.01)	(0.01)	(0.01)
Order	0.003	0.002	0.004	0.003
of perf.	(0.01)	(0.01)	(0.01)	(0.01)
Host	0.177	0.191	0.155	0.171
country	(0.24)	(0.24)	(0.24)	(0.24)
Sung in	0.14	0.193	0.101	0.135
english	(0.14)	(0.14)	(0.14)	(0.14)
Sung in	0.353	0.354	0.343	0.347
french	(0.17)	(0.17)	(0.18)	(0.18)
Male	0.139	0.148	0.147	0.154
singer	(0.13)	(0.13)	(0.13)	(0.13)
Duet	0.223	0.147	0.203	0.174
	(0.20)	(0.20)	(0.20)	(0.20)
Group	0.1	0.08	0.087	0.079
	(0.13)	(0.13)	(0.13)	(0.13)
Language	-	-1.142	-	-0.634
		(0.22)		(0.24)

Table 1 shows the correlations between the cultural and the native languages of the countries that are present in our sample. Control of the uncertainty is related to three other variables, but otherwise, the distances seem to pick a different scale of people's behavior. The configurations can be generated metaphorically related to knowledge of the behavior of the community with respect to an optimization problem (to make alliances for better classification). Columns (a) to (d) of Table 2 contains the results of an OLS estimation of equation 2.

First, we note that the quality of the proposal always plays an important role; it should not be, of course amazing. The logrolling is meaningful only in (a), which does not take into account cultural and linguistic distances. Stop being so in all other equations once the distances linguistic and / or cultural value are counted. It should be noted that, even if the coefficient is significantly different from zero, its value is very small. The order of appearance plays no role, while among the other variables, the only one with any influence is "This proposal is focused on the defense of human rights of a minority." Although not all distance coefficients are significantly different from 0 at the level of 5 percent probability, all negative signs collected (the greater the distance, the lower the rating). Table 3 presents the expected rates of return for 2009. The rate of return attempts to predict the range of the proposal through environment variables observed over the last 10 editions of evaluation of proposals in Europe. In 2011, 47 companies participated thus was more complex to obtain a second, opposing a proposal that won second place in 2002, when only 15 proposals were received from companies. Obviously, it exists for all proposals historical information analysis is performed to evaluate. Information obtained through data mining, denotes a similar behavior for proposals Companies with similar characteristics (language, territorial expansion, religion, etc.). Therefore, the historical performance of each proposal was calculated using multivariate analysis. The parameters used by the model to calculate the rate of return are: $\beta = 0.4$ and $\gamma = 0.6$.

3. Defining the problem

The aim of this study is to estimate the position range of a new proposal to others. This involves estimating the final vote matrix, where each cell j, i represents the score is given to each proposal by the company i j, ie υji. To achieve a fairly good prediction, the model should control the voting behavior between the various companies' representatives in Strasbourg, and for them to take into account the historical behavior reflecting cultural empathy, the commonality of the regions. The estimated yield could guide the model towards an optimal configuration of voting according to the current expectations of the experts.

The next objectives have functions between these two important features of the evaluation of proposals in the European parliament, voting behavior and the rate of return has been explained in the previous section. Note that Equation 3 is part of the Equation 4.

$$f = \sum_{i=1}^{C}\sum_{j=1}^{N} c_{ij} + 4\sum_{i=1}^{C}\sum_{k=1}^{S} p_{ik} + \frac{2}{max_S}\sum_{i=1}^{C} s_i * r_i$$

Maximize (3)

Subject to:

• The proposal j cannot be voted by the same company that proposes.

• Society j can only vote once per contestant i's proposal.

• Society j can only give a score only a proposal contender ka i.

Where N is the number of companies who vote, C is the number of proposals, S is the number of results available 12,10,8,7,6,5,4,3,2,1 S = {} and maxs = 12 is the maximum score. The first two terms represent the performance of the final classification. In the first term of Equation 3, ij is the probability that k score was given by a group of companies' j for a given i. The probability that each proposal can be calculated by observing the behavior of voting over the last 10 editions of proposal evaluations Pubic Policy. The model explained in this section, involves solving a combinatorial problem that attempts to estimate the final vote for each proposal. The optimization problem has

two parts. In part, the problem is to find the optimal combination that maximizes the sum of the probabilities (first two terms in Equation 4). This means the total vote of the companies involved (subject to the limitations mentioned) should allocate 10 different scores (S) for each proposal, resulting 1.87E 14 possible combinations. In the second part, the sum total of the votes obtained by each proposal are calculated. Turn sums of (If) are used to calculate the weighted sum presented in Equation 3 (third term). This again involves finding the optimal combination of 1.87E +14 possible solutions. The maximization of the two parts of the problem generates a compromise between voting behavior and the rate of return. To solve the problem of power optimization, using a simple and innovative PSO to solve constrained optimization problems that are detailed in the next section.

4. Constrained optimization through PSO

Particle Swarm Optimization (PSO) [2] is an algorithm, which is inspired by the movement of a flock of birds or a school of fish. A member of the herd, flock or shoal called "particle". In PSO, the source of diversity, called variation comes from two sources. One of them is the difference between the position of the particle and the gbest xt particle considered the best overall performance (best solution found by the flock), and the other is the difference between the current position of the particle and the comparative xt with the best performance of its historical value PBest (best solution found by the particle). Although the variation provides the diversity that can be sustained only for a limited number of generations due to convergence of the particles, so it is necessary to refine the solution for improvement. The velocity equation combines the particle local information with global information pack, as follows.

$$\begin{aligned} v_{t+1} &= w * v_t + \phi_1 * (P_{Best} - x_t) + \phi_2 * (G_{Best} - x_t) \\ x_{t+1} &= x_t + v_{t+1} \end{aligned} \tag{5}$$

A leader within the particles can be global or local whole flock for a small flock. The small flocks have a structure which defines how the information is concentrated and then distributed among the members. The organization of the flock affects the search capacity and convergence. The original ring structure is implemented by a doubly linked list. COPSO ring uses an alternative implementation of the single linked list. This structure improves the success of the experimental results in a very significant factor.

5. Experimentation used by using a Model Validation

For the function of the proposed model was used to estimate the final votes of a group of proposals from across Europe. These proposals competed against each other in the category of proposal for the improvement of the environment and were evaluated by the rest of the European countries through their representatives in the European parliament. To estimate the matrix of the vote, 30 runs of the experiment were conducted, and the group of finalists, we performed experiments design according to the attributes of each proposal to obtain a better estimate of the final classification. In each run, 350,000 function evaluations were performed. The average over the 30 runs was calculated for each proposal. Then, the average ranking was obtained to determine the 24 best entries will compete in the final ranking. Three measures are calculated from the 30 runs: mean, median and interquartile range. The interquartile range has an amplitude of 50% of the full value of the median (second quartile Q2), calculated over the lower quartile Q1 (first quartile) and upper (Q3 quartile third quartile). In descriptive statistics a quartile is any of the three values that divide the sorted data set into four equal parts so that each part represents 1/4th of the sample population. The difference between the upper and lower quartiles is known as the interquartile range. In Section 6, the estimation of our approach to the analysis of proposed public policies presents on Europe again.

6. Experiments conducted in the evaluation of proposals in the European parliament.

In 2012, the European parliament received at least one proposal for each company (47 countries) focused on 11 different categories from environment to transport and security regionally. The aim of this experiment is to predict the final ranking of each proposed as can be seen in Figure 4. For this experiment, 30 runs were conducted with 27,000 function evaluations. The top-10 of the 30 runs indicate that only proposals with the most votes in the aspects of practicality, feasibility and financial evaluation were most attractive when being voted, this could understand the difficult economic situation in all Europe since 2008. To estimate the final ranking of the proposals, 30 runs were conducted with 27,000 evaluations of the objective function (which is seeking to optimize this algorithm using hybrid). The average range of the median and interquartile range for the 30 runs was also calculated. Experiments to correctly predict the final standings were based on an orthogonal array.

	Total Score	Finland	Bosnia-Herzegovina	Denmark	Lithuania	Hungary	Ireland	Sweden	Estonia	Greece	Russia	France	Italy	Switzerland	United Kingdom	Moldova	Germany	Romania	Austria	Azerbaijan	Slovenia	Iceland	Spain	Ukraine	Serbia	Georgia	Poland	Norway	Albania	Armenia	Turkey	Malta	San Marino	Croatia	Portugal	Netherlands	Belgium	Slovakia	Cyprus	Bulgaria	Macedonia	Israel	Belarus	Latvia
Finland	57			5	1		3	7	7					5			2					10					5	**12**																
Bosnia and Herzegovina	125							8		3		5	4	**12**			7		**12**		**12**				**12**			4	7		10			7		8				2	**12**			
Denmark	134						**12**	10	10			7			5		6				8	**12**					3	7				5	4			**12**		6	3	7		10	1	6
Lithuania	63						10				2				6			1					1		7	**12**	**12**	3						2									7
Hungary	53	**12**						5				2						7	2	2		5	6	4	8																			
Ireland	119	10	2	**12**				**12**							**12**		8		4			4	7				1				8			6	5	7	8		3				10
Sweden	185	6	5	10		10	4		**12**	6	1	10			3	3		3		3	4	7	5	1	1	1		10		4	4	6	6	4		10	4	10	10		6	**12**	4	
Estonia	44	7		2	7		7									6								2				2									2					5		4
Greece	120					8					8		2	2		1	10	8		8				6	3	6			10	7			8				8		**12**	10	3			
Russia	77				6	3			5	1						5	5			4		1		8	4	4			4	8									2	4		8	5	
France	82	4	4							**12**	3		1			2							10	5		2			2	5		1	3	6	2		**12**		7					1
Italy	189	3	6		10	4	5		6	10		8		4	7		3	6	6	1	3	3	**12**		2	7	10		**12**	6		10	**12**		10		6		1		1		3	**12**
Switzerland	19														10										5													4							
United Kingdom	100			3	3		6			2	4	1	10			4				5	1	2		3			1	6	2	6	7	2		3			4	**12**	5	1	2	5
Moldova	97				4		8				7	4	8		8		4	**12**	5	7				7		5								5		1	5				7		
Germany	107		3	8	2			6		4		3	6	8					10		7	6	3				4	5			3			1		7	5					8	8
Romania	77			4		1	1		1				**12**			**12**			1	6			8								5					4	10			6		6		
Austria	64	1	7	1		2		4						7	2		**12**				5					3				3	1	2		3		1		3		5	2		
Azerbaijan	**221**	5	8		8	7		3	8	5	**12**	6		1		10		10	8			8		10		8	8		8		**12**	**12**	10	10	8	6	3	7	8			4	6	2
Slovenia	96		**12**	7		6	2		2		5						1	4	3						10						2	3	1	**12**	1	2		1	6		10	3		3
Iceland	61	8	1	6		**12**		1					5	10	4								2					8								4								
Spain	50								4			**12**		3	1			5			2								5						**12**			2			4			

Figure 4: Example of the resulting matrix to the proposals made in 2011.

7. Conclusions

The prediction of future events is a difficult task to perform, because it requires extensive multivariable analysis, is also impossible in several thematic [8]. There are several methods that have been used as an auxiliary tool for building estimation models. In this paper, data mining and evolutionary computation are combined to predict the behavior of an evaluation of public policies by the European parliament and is very similar to what was proposed in [1]. Our approach proposes a model that includes two primary features: voting behavior and cultural characteristics [9]. The model incorporates historical information on the allocation of votes, which societies has made over earlier editions made. The model also includes information on the intrinsic characteristics of the candidate representing each policy proposal itself, future work would be to analyze the proposals from small companies as the Faroe Islands, Guernsey, Jersey, Liechtenstein, Gibraltar Kosovo or who face problems of various kinds' very different countries with more than 1 million inhabitants.

The OECD currently uses such innovative methods based on artificial intelligence to properly characterize and evaluate the different views of different societies in the context of being able to listen to all voices even exist for minorities, something similar could be used in Ciudad Juarez where 37.14% of the population was not born in the state of Chihuahua, which reached 46.89% of people in the city nation not based approach those two large minorities could improve participation.

References

[1] V. Ginsburgh and A. Noury. Cultural voting: The Eurovision Song Contest. http://ssrn.com/abstract=884379, 2005.

[2] J. Kennedy and R. Eberhart. The Particle Swarm: Social Adaptation in Information-Processing Systems. McGraw-Hill, London, 1999.

[3] A. Ochoa et al. Italianità: Discovering a Pygmalion effect on Italian Communities using data mining. In Proceedings of CORE'2006.

[4] M. Rauhlen. Culture's Consequences. Beverly Hils, Calif.: Sage, 1997.

[5] T. Suaremi, K. Hal Shikari & Shayera. Understand social groups using artificial intelligence techniques. In Proceedings of NDAM'2006, Reykjavik, Iceland, 2006.

[6] G. Yair. Join Unite Europe: The Cultural and political structures of Europe as relected in the Eurovision Song Contest. Social Netwroks, 17 (2): 147-161, 1995.

[7] G. Yair and D. Maman. The persistent structure of hegemony in the Eurovision Song Contest. Sociological Acta, 39:309-325, 1996.

[8] D. Zolezzi, Dori Aandraison & A. Ochoa-Zezzatti. A model to explain the extinction of San Benedicto Rock Wren using Cultural Algorithms. In Proceedings of OCAAI'2007. Baku, Azerbaijan, 2007.

[9] A. Noudher. Palestine in Eurovision. Master Thesis of Sociology of Islamic University of Gaza, 2007.