Study of Classification of Attributes of Quality in Argentine Sites of E-Commerce

M[A] BEATRIZ BERNÁBE LORANCA
Cuerpo Académico de Ingeniería de Software -Aplicaciones Estadísticas.
Facultad de Ciencias de la Computación, BUAP.
Calle 14 sur y Avenida San Claudio, Colonia San Manuel, Puebla.
MÉXICO.

JUAN OLIVETO; GUSTAVO LAFUENTE; LUIS OLSINA
Grupo de I+D en Ingeniería de Software
Departamento de Computación, Facultad de Ingeniería, UNLPam.
Calle 9 esq. 110, 6360 General Pico, La Pampa
ARGENTINA

*Abstract:-*The starting point of this work conforms it the results obtained in [10] where they are analyzed and they evaluate attributes of quality to be considered in projects operative Web, mainly for the characteristic of functionality of sites or applications with e-commerce.

In order to obtain groups or classes of Web sites and attributes of quality that reveal interrelation between sites and attributes, we have applied the Clusters Analysis (AC). This group, has leaned in the categorical data that are concentrated in a Binary table and which they come from evaluations e-commerce. Since it is desired to obtain a reduced number of clusters that simplify to the analysis of sites and attributes, some groups of clusters were created. Consequently a function was defined to know the most reliable number of clusters for the sites.

Lastly, we present conclusions and a summary of the obtained results and options for future research.

Key-words:- Analysis, Attributes, Classification, Cluster, E-commerce.

1 Introduction

The development of web sites as the electronic commerce, requires of the complicated combination of analysis, design and application of different methods and tools. Nevertheless, in a qualitative sense, it is possible to abstract and to include many of the same qualitative characteristics when e-commerce is constructed to an application. With respect to Functionality, this characteristic of quality is distinguished of a quality model. As well, Functionality, one comes off a set of hierarchic represented by a tree of characteristics and attributes and that are maintained by a specific categorización for the dominion[10]. Considering a sample of 71 Argentine sites e-commerce [10], it has been analyzed and evaluated 17 requirements that conform the quality tree on the aspect of functionality under the criterion: available (1) or nonavailable (0). From being from such evaluations a binary table of 17 variables by 71 cases was obtained. This table has been processed with the statistical technique of Nonsupervised Classification denominated Cluster Analysis. The AC is also a multivaried tool that has helped us to enter us to the essence of the phenomenon to throw classes of sites and attributes. In order to obtain this group, a simple mechanism has considered that allows to choose the best number of classes.

The investigation priority that follows to him this work, is centered in the analysis of the components of the classes. This induces to describe the property relation that exists between the sites that belong to a class in individual. For organization effects, the

present work is organized in 5 sections: Introduction, analyses of the problem, classification of variables, classification of sites, and conclusions.

2 Analisys of the Problem

The Cluster Analysis, first supposes that strong attachments between the variables exist that are going to form the profile of clusters. This process of verification we have successfully made it by means of the Analysis of Correlations. Because we have categorical data, we agree to initiate the process of grouping with Analisis Cluster Exploratory (ACE). This analisis will give information sufficient to determine the optimal number us of clusters for the data under study. Experimentally speaking, until this point, single it is possible to consider very intuitively as it could be the best solution respect to the number of clusters. Later, it is possible to propose a solution of cluster open, that is to say, to construct a mechanism that includes the procedure of rank of solutions [x1 xn]. This process indicates the degree of property of a variable to determining to cluster (Cluster Membership). The procedure that we have proposed (ACE), this constituted by a set of ordered steps. The ACE can also be seen like a Conglomerativo Mechanism, and that has like objective the obtaining of a guessed right set of classes of sites and attributes e-commerce. Those are 8 points that describe of general way the Conglomerativo Mechanism to obtain classes. MC proposed, the nonsingle one is applied to this case e-commerce, also is flexible for other situations that require categorización:

1. Formulation of the Problem
2. Selection of a Measurement of Similarity
3. Standardization of Data
4. Supposed of the Analysis
5. Selection of the Procedure of Grouping
6. Decision of the Number of Number of Clusters
7. Interpretation and Elaboration of the Profile of the Clusters
8. Validation of Obtained Conglomerates

2.1 Formulation of the Problem in E-Commerce Classification

Counting on the selection of variables (in this case by sampling) and verifying the correlation between the variables, we can to say that we have formulated our problem AC. For it we were based on that the set of variables selected must describe the similarity between the objects, thus, is possible to formulate some hypotheses on support of variables and that later must be proven.

In order to begin and to describe first perception that is had on relation between the variables, we observe in "Correlation_2004" [14], that variables 8 and 9 (Shipping and Cost information and payment information) have a correlation coefficient (CC) of 0,71. Following in order of magnitude, we observed the correlations of 0.53 between variables 4, 6 (Sales on-line and Secure transaction). The variables 7 and 16 (Shopping car of Searching Mechanisms Global) have a correlation coefficient of 0.52,) and Searching Mechanisms Global and Restricted (Variables 16,17) have CC of 0.5. The following pairs have smaller coefficients of 0,5, but they are even significant:
5,4 (sales off-line is against to sales on-line) (-0.49), 5,14 (Promotion at sales and prizes) (0.46), 2,17 (subscription client and restricted searching) (0.42) and finally 8.10 (information shipping and costos information and puncharse cancellations information) (0.42) (See Table A.1 Apendix A)

By the previous thing, we can affirm that the characteristics of Shipping and Cost information and payments information, are important sales along with online and safe transaction. Of equal way it must pay to it them attention to the searching mechanisms with the shopping car

2.2 Selection of a measurement of similarity and standardization of data

Remember that the objective of the AC is to group similar objects, then is necessary some measurement to evaluate the differences and similarities between objects. The similarity (similarity) is a measurement of correspondence or similarity between the objects that are going to be grouped. The strategy commonest consists of measuring equivalence in terms of the distance between the pairs of objects. The objects with distances reduced among them are more similar to each other, otherwise those that have greater distances will group within the same one to cluster. This way, any object can be compared with any other object through the similarity measurement. In the process of similarity between the objects of a AC, three methods are distinguished: Measures of Correlation, Measures of Distance and Measures of Association. By virtue of which the binary data of the table that is processed are measurable, the correlation and those of distance are

the optimal ones in this work. Certainly, the distance measures are quite sensible to the differences of scales or from magnitudes done between the variables, therefore, we have taken into account the dispersion between the attributes from quality, later to standardize of being necessary. Nevertheless, the binary table in question frees the high dispersion.

2.2 Supposed of the analysis

the AC it is a methodologic objective to quantify the characteristics of a set of observations. For that reason, it has forts mathematical properties, but not many statistical foundations. The requirements of normality, linearity and homocedasticity (so excellent in other techniques), have little consistency in the AC. Nevertheless, attention in other two essential questions for this type of analysis is due to render, as they are: the representativeness of the sample and the multicolinealidad. Because the sample of 71 Web sites is reliable, we can "ideally" supposition that the cluster analysis will be as good as it is it the representativeness of the sample, then is possible to reject the multicolinealidad.

The selection of the variables and calculate of the matrix of similarities, give rise to the partition process. With the purpose of justifying the grouping algorithm that is going away to use to form clusters (groups) we have rescued important points of the grouping procedures, then it is possible to make a decision on the number of groups that are wanted to form. In the following section one briefly describes two types of grouping procedures

3 Clasification of Variables

3.1 Selection of the procedure of grouping

The suitable election of the process of classification applied to our table of data, justifies in the theory [1]. Of this theoretical frame, we rescued the characteristics of two types of essential procedures: hierarchic and the nonhierarchic ones. The acquired experience of the hierarchic conglomerate (CJ) in other studies [2, 3], makes us think that he is of utility in this case e-commerce. The CJ is characterized by the development of a hierarchy or structures in tree form, but still, the results of the first stage can be nested with the results of the last stage, of such form that originates the tree. As well, methods CJ can be by Aglomeration and Division. We emphasized that the conglomerate by aglomeration, consists of Methods of Connection, Methods of Variance or Sum of the Squares of the Error and the Centroide Method. On the other hand, the Method of Connection can be simple, complete and average. After making and analyzing several tests to create clusters, we decided to apply the Metodo Jerarquico by Aglomeracio'n (MJA) for the grouping of the variables. In order to compare the groups of the MJA, complete such agglomeration with the connection methods complete/simple and the one of variance-Ward. Figures 1 to 7 below show the resulting clusters for aglomerativos CJ.

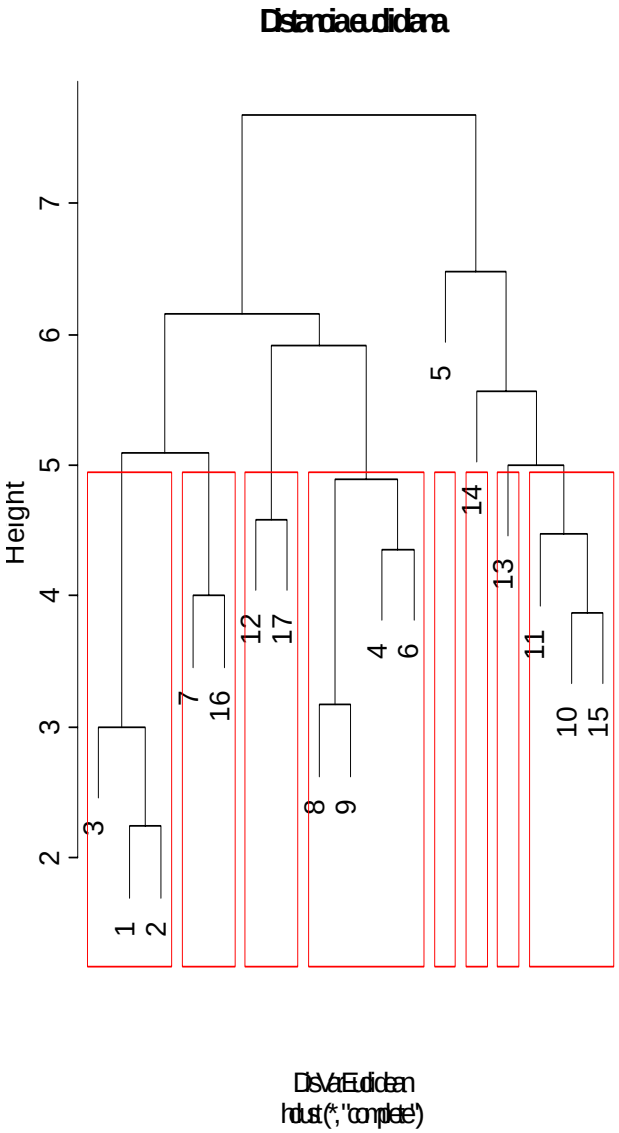

Figure 1

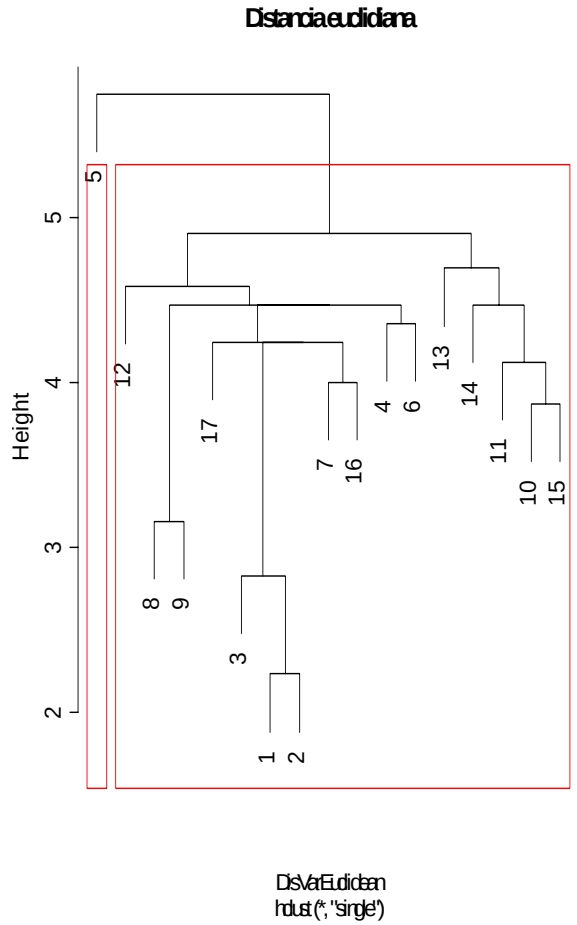

Figure 2

In figure 2 and 4 we showed the turn out to apply to the procedure (MJA) with the method of simple connection and euclidiana distance 6 and 4. This method is based on the minimum range or the rule of the next neighbor. The first two conglomerated objects are those that to each other have the smaller distance, then, the following shorter distance is identified, or that the third object is grouped with both first or that forms a new conglomerate of two objects. In each stage, the distance between two clusters is the distance between its two next points, and, in any stage, two clusters arise by the shorter simple connection between these. This process continues until all the objects are in a conglomerate. When processing itself the attributes of quality with this method, we observed that for a distance of 6 (Fig. 2), the simple single method throw 2 groups, one of them contains to variable 5 (Offline-Sales) and the other group this made up of the rest. It is obvious that with so single 2 clusters it is not possible to give answers on clasificaciòn of the attributes, but we do not lose to the variable 5. On the other hand, the simple connection with distance 4 (Fig. 4) to thrown 13 groups, which makes suppose that this conglomerate this outside the objective of the analisis that we persecuted.

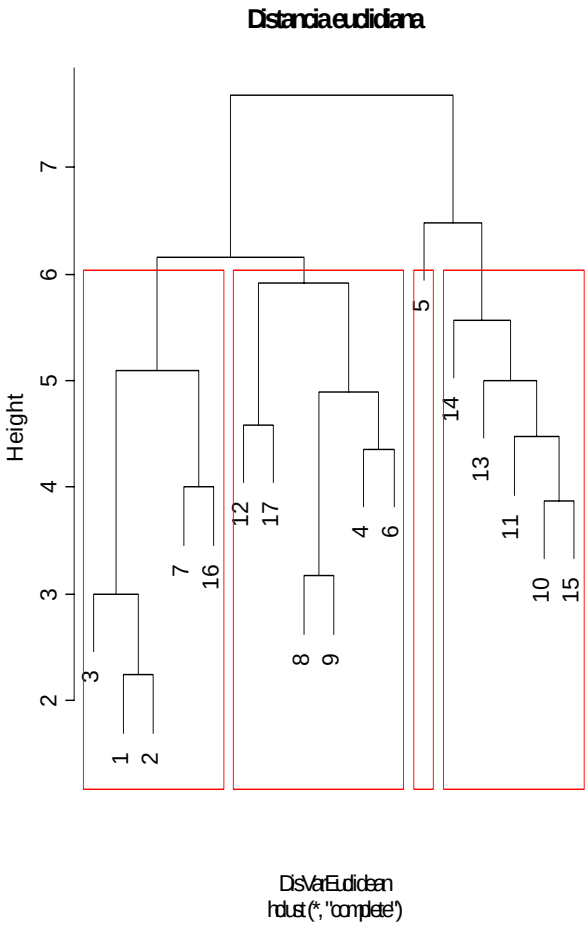

Figure 3

The clusters on Fig. 1 and 3 show the application of method complete connection with distance 5 and 6 respectively. The Fig. 3 shows a grouping of 3 main groups, but figure 1 is not excellent. It is important to mention that this method of complete connection is similar to the simple connection, unless this last one furthermost is based on the maximum distance or the strategy of the neighbor. In this case, the distance between two clusters calculates like the distance between its more distant points.

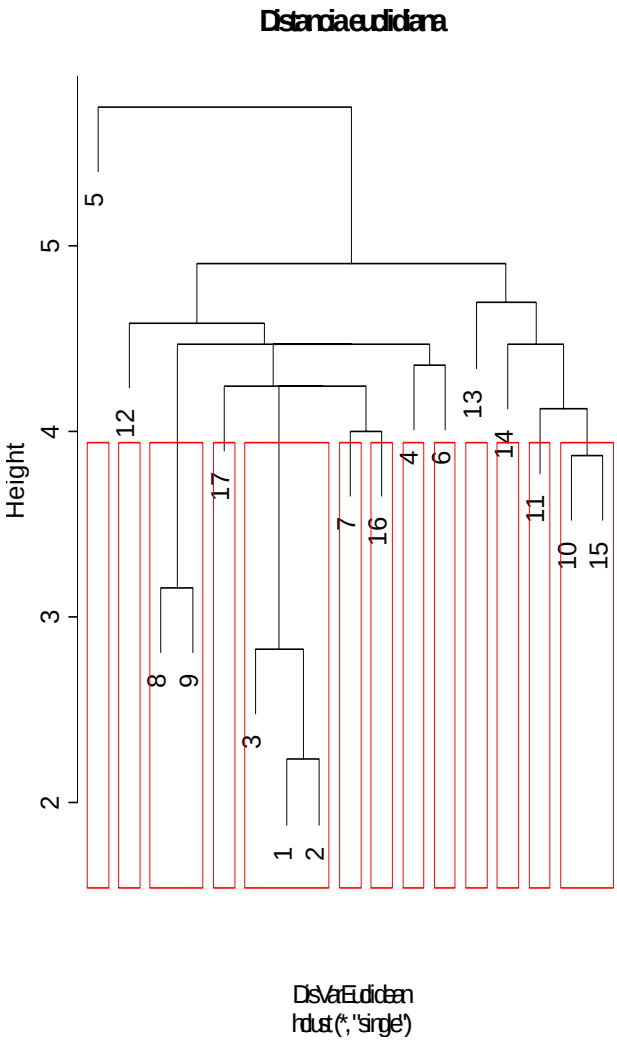

Figure 4

3.1.1 Metodos of variance

In this work we compared the different metodos from connection along with procedure of Ward, who is excellent metodo of variance. The Methods of Variance try to generate clusters in order to reduce the variance within the groups. In the procedure of Ward to create each conglomerate, the average ones for all the variables calculate. Later, for each object, square euclidiana for the averages of the groups is compute the range, as well, these distances are added to all the objects. In each stage, clusters with the smaller increase in the extreme total of the squares of the distances within the clusters are combined both. Of the hierarchic methods, the method of Connection Average and the Procedure of Ward have demonstrated a better performance than the others

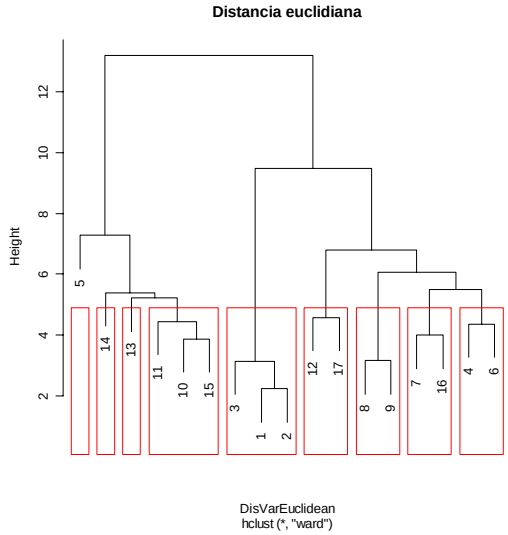

Figure 5

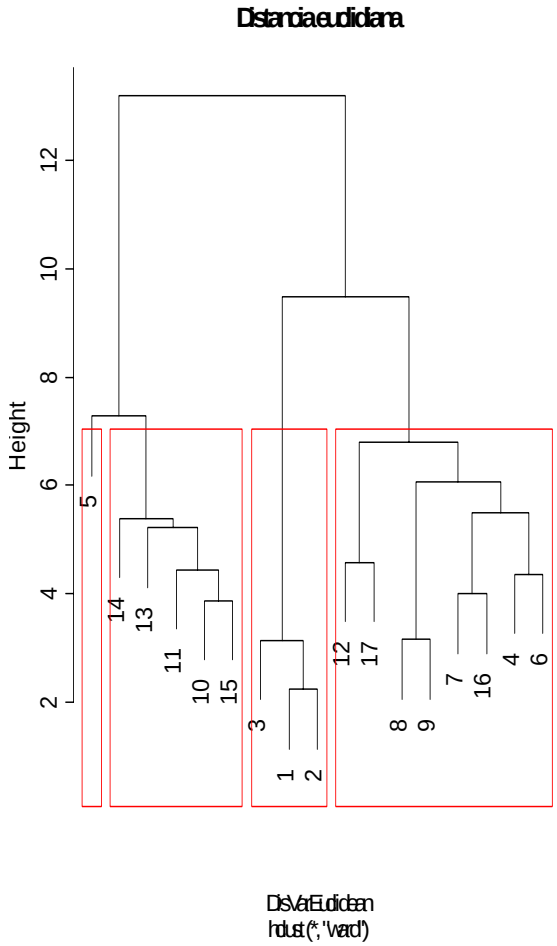

Figure 6

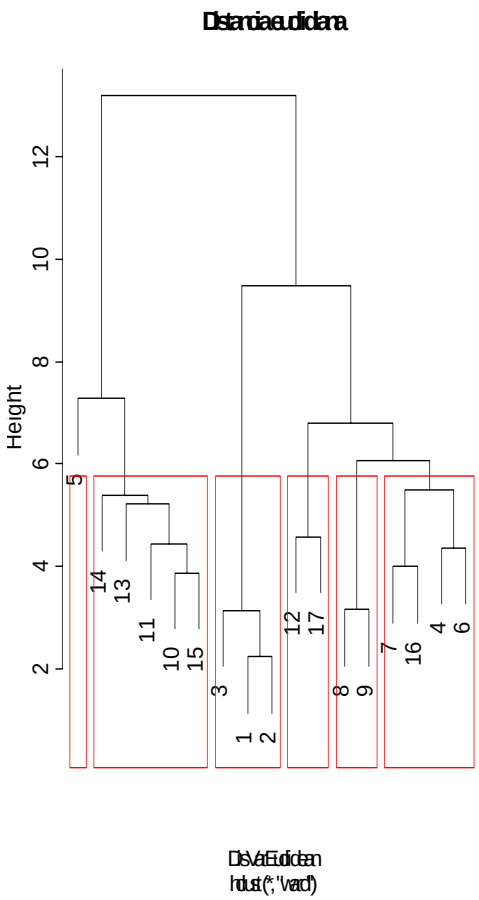

Figure 7

To Figs. 5, 6 and 7 have been applied to the Conglomerate Jerarquico Aglomerativo with the method of Ward and distances to them 5, 7 and 6 for each case. The best set of clusters is distinguished in the Figure 7, which is looked like the group of figure 4. Asociaciòn and analysis between clusters that we have obtained, we will present/display it in the conclusions. The Centroide Method was not used to group variable, but it is a good method of grouping for cases. This procedure also includes to I calculate of distances, single that the centroide method the distance between two groups is the distance between its centroides (average for all the variables). Whenever the objects are grouped, a new centroide calculates.

4 Analysis of Clusters of the Web Sites

With the objective to continue the grouping which we have obtained with the variables, and, stopped in the scene of the sites e-commerce, to obtain a classification of cases, we directed our attention to a second type of procedures of denominated clusters methods of nonhierarchic clusters. The Web sites, have been classified with this metodo, and in this section we raised a small teorico frame of reference to so justify the election of metodo. The nonjerarquicos clusters frequently are known like Average Grouping of K. These methods include the Sequential Threshold, Parallel Threshold and the Division for the Optimization. Two forms exist basic to know the way grouping of the objects at issue: Graph of Cara'mbanos and Dendograma. We have applied this last one to obtain the groups of Web sites, and for their interpretation, the Dendrograma, Lee of left to right. The vertical lines represent the united groups. The position of the line in the scale indicates the distances in which the groups are united. Because, in the first stages, many distances have similar magnitudes, it is difficult to determine the sequence in which some of the first clusters form. Nevertheless, it is evident that in last the two stages, the distances in which the clusters are combined are great. This information is useful to decide the number of clusters. Also it is possible to obtain data on the participation of the clusters of the cases if the number of groups is specified. Finally, to complete our study of classification, we resorted to the process of the Binary table to create clusters or groupings of Web sites. Since we have mentioned, we used the method of k-average determining the most suitable number of clusters to obtain. For this intention a procedure similar to the used one in [3] was used, in this case several groupings by means of the k-average method were obtained using the 17 proportionate variables and varying the number of clusters from 4 to 20

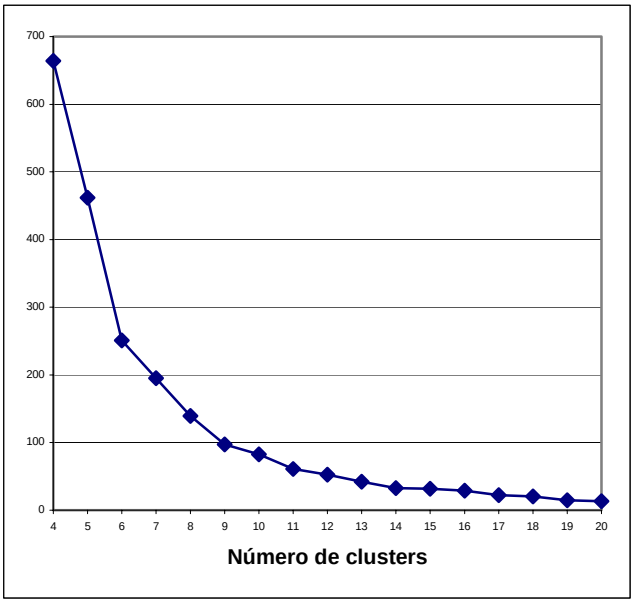

Figure 8

The best number of clusters was obtained using the extreme average weighed of the internal one of squares (within s square sum) in each conglomerate and graficando it against the number of clusters. Figure 8 sample that the approximate point of flexion is in number 9, reason why this one volume like the number of groups to analyze itself.

The nine resulting clusters by the method of the k-average are in Table 1.

Table 1

N	Web Sites
1	29,30,32,41,47,48,57
2	20,33,36,38,40,44,50,59
3	2,6,19,34,49,67
4	9,10,11,24,46,52,56,66,68
5	1,14,17,25,31,60,64
6	4,8,16,18,27,39,45,58,71
7	7,12,15,61
8	3,5,23,26,28,35,53,55,62,63,69,70
9	13,21,22,37,42,43,51,54,65

In addition to the method of the k-average, the method of hierarchic clusters was used to analyze alternative groupings.

Figures 9 and 10 are 2 examples of dendogramas that reflect the grouping of sites. The groups of sites compose the conglomerate as well. In future investigations the 9 clusters were analyzed.. This agrupaciòn was not studied with detail in this work, but with such dendogramas we stimulated the continuity of the study of categorizacion e-commerce. On the other hand, we will take care of the analisis of the resulting classes, nonsingle to identify property from a site to a group specify, also we are interested in redefining, enriching and to model the metodologic scheme of conglomerativa statistical that has appeared in this work.

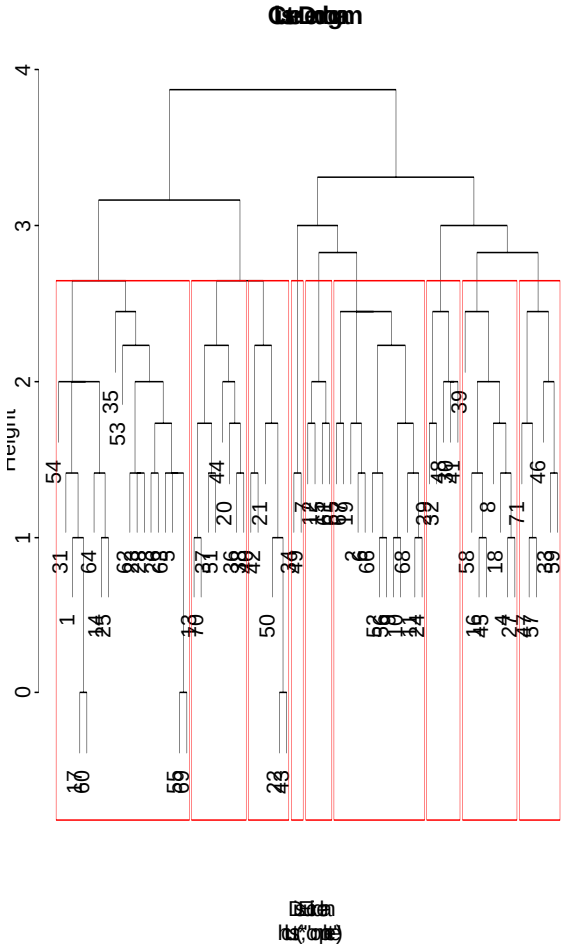

Figure 9

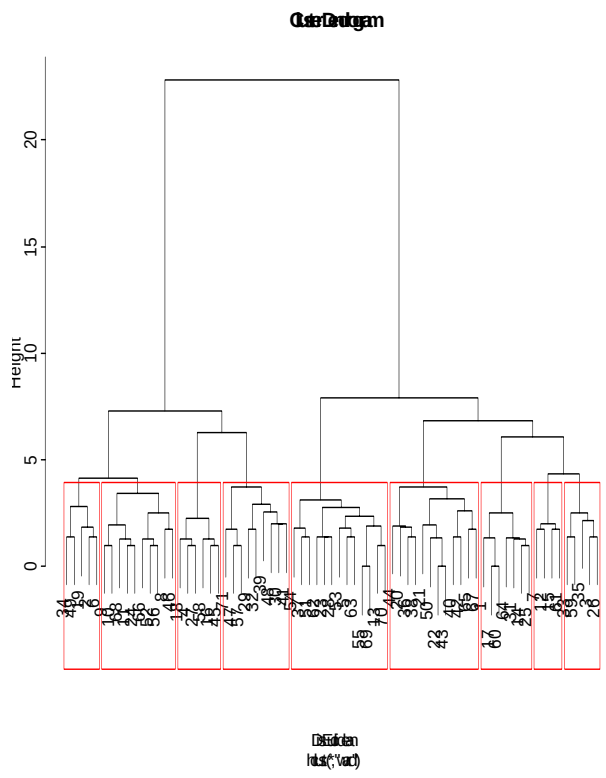

Figure 10

5 Conclusions

The classification methods have extended in diverse areas of investigation and commercialization. Although the dominion and enrichment of the qualifying algorithms are a difficult task, it is possible to include/understand the essence of each one and to apply them suitably. By such reason, in this paper we have taken care to apply classification nonsupervised (clusters) to a table of categorical variables. Since we took care of the analysis categorical of the variables, and, after analyzing the resulting clusters for the quality attributes on functionality in e-commerce, we chose 2 groups to conclude the importance of the characteristics. The conglomerate that used complete connection and euclidiana distance 6 I throw 4 groups like the conglomerate with method of Ward and euclidiana distance 7, even, the components of the groups are such except variables 7 and 16, which belong to a class that contains to attributes 1, 2 and 3 in the conglomerate with complete connection. Without a doubt some, the information of the product excels in the same way in this classification observed in [9]. The variable of subscription client is associated with searching mechanism restricted; shipping and costs information and payments this related to online sales and secure transaction. We remembered that the study [2], secure transaction is associet with customized count and searching mechanisms. It is clear thus, to notice coincidences of these variables with those of the mentioned study.

With respect to characteristics 7 and 16 (shooping car and seraching global), these are related very, for that reason, we understand that if exists an icon of shooping car, the site would have to count on a global mechanism searching. The same it happens to searching restricted and information for the client on its purchases. We emphasized the fact that when making a sale in line demands a safe transaction, which is possible to be appreciated in dendogramas. Finally, Promotion at sale, customized account and quik buy mechanism accompany by information of cancellations and promotion by prizes. It is not chance that the shopping car "is not joined" with these variables since when doing a quik shopping we did without of the car shooping. When comparing the groups of variables in [2] with those of this paper we distinguished some differences, and, we think that he is not strange, since also very important similarities are observed. The brief differences can obey to that on the one hand, in [2] resorted to a different software package to process the table, in this work we used R [6,12], and on the other hand, it is clear that the nature and size sample of the sites change in 4 years. Like later work, the priority is to describe with exactitude a mechanism that allows to obtain dispersed sets of sites to respond the degree of property from a new site to the sample on which we counted. Evidently such mechanism is inserted in the applied statistical process in this work. Finally, when combining models of evaluation in e-commerce with methods of

classification nonsupervised, we bet to the sprouting of a model (initially experimental) that speaks on the generalized behavior of Web sites of this category.

Appendix A

Table A.1 Functionality and Content-oriented E-Commerce Attributes and Variables.

1	**Product Information**		
	1.1	**Basic Product Description**	**IP-DES**
	1.2	**Product Image**	**IP-IMA**
	1.3	**Catalog**	**IP-CAT**
2	**Shopping Features**		
	2.1	**On-line Sales**	**CC-V-ON**
	2.2	**Off-line Sales**	**CC-V-OFF**
	2.3	**Secure Transaction**	**CC-TRANS**
	2.4	**Shopping Cart**	**CC-CAR**
	2.5	**Shipping and Cost Information**	**CC-IEC**
	2.6	**Payment Information**	**CC-IP**
	2.7	**Purchase Cancellation Information**	**CC-ICAN**
	2.8	**Quick Buy Mechanism**	**CC-MERC**
3	**Client Customization**		
	3.1	**Subscription**	**PC-SUS**
	3-2	**Customized Account**	**PC-CUEN**
4	**Promotion Policies**		
	4.1	**Promotion at Sale**	**PP-PROV**
	4.2	**Promotion by prizes**	**PP-PROVR**
5	**Searching Mechanisms**		
	5.1	**Global**	**MB-GLO**

References:

[1] Anderson, T.W, 1984, "An Introduction to Multivariate Statistical Analysis", 2nd Edition. Wiley.

[2] Loranca, M.B. & Olsina, L WSEAS International Conferences, ASCOM'04 Cancun, Mexico, May 12-15, 2004. (6th WSEAS International Conference on Algorithms, Scientific Computing, Modelling and Simulation). "Classification of Poputalion Zones Using Multivariate Statitistical Techniques"; Cancun, México, mayo 2004.

[3] Bernabé L.,B., López S., R.. 2004 "Application of Classification nonsupervised to Population Data". International Conference on Electrical and Electronics Engineering and X Conference on Electrical Engineering, ICEEE/CIE 2004, Acapulco Guerrero, México.

[4] Chatfield, C.; Collins, A.J., 1991, "Introduction to Multivariate Analysis", Ed. Chapman & Hall.

[5]. Dixon, W.J, 1990, "BMDP Statistical Software Manual", Vol I, II. Dixon,W.J Eds, University of California Press, Berkeley, California.

[6] J. Maindonald and J. Braun (2003), "Data Analysis and Graphics Using R: An Example-Based Approach", Cambridge University Press, ISBN 0-521-81336-0

[7]. Kalakota, R.; Whisnton, A.B.,1997, "Electronic Commerce: A Manager's Guide", Addison-Wesley.

[8]. Lafuente, G.H.; Oliveto, J.; Olsina, L.; 2000, "Requerimientos de Calidad en Sitios de E-commerce" Proceed. JUCSE 00, Nuevas Tendencias en Ingeniería de Software, Universidad Católica de Santiago del Estero, Arg., ISBN 950-31-0045-3

[9]. Lafuente, G.H.; Oliveto, J.;. Bernábe Ma B.; Olsina, L.; 2004, "Estudio de Atributos de Calidad en Sitios de E-commerce Argentinos" CACIC 2004, X Congreso Argentino de Ciencias de la Computación.

[10]. Olsina, L.; Lafuente, G.J.; Rossi, G.; 2000, "E-commerce Site Evaluation: a Case Study", Lecture Notes in Computer Science 1875, Proc. 1st International Conference on Electronic Commerce and Web Technology (ECWeb 2000) , Springer-Verlag, London-Greenwich, UK, pp. 239-252.

[11] http://estadistico.com

[12] http://www.r-project.org "Statistical Software Manual"

[14]http://www.cs.buap.mx/~bety/Investigacion.html