A Conceptual Model for Guiding the Clustering Analysis

Wagner F. Castilho[1,4], Gentil J. Lucena Filho[2], Hércules A. do Prado[2,3], Edilson Ferneda[2], and Margarete Axt[4]

[1] Brazilian Federal Savings Bank, Brasília, DF – Brazil
SRTVN 701, conjunto C, Bloco A – Sala 321
70.719-930 Brasília, DF – Brazil
[2] Graduate Program in Knowledge and Information Technology Management
Catholic University of Brasília (UCB)
SGAN 916, Módulo B
91.501-970 Brasília, DF – Brazil
[3] Embrapa Food Technology – CTAA
Av. das Américas, 29501 - Guaratiba.
23.020-470 Rio de Janeiro, RJ – Brazil
[4] Federal University of Rio Grande do Sul
Av. Paulo Gama, 110
90.040-060 Porto Alegre, RS – Brazil
castilhowagner@gmail.com, glucena@pos.ucb.br,
hercules@ctaa.embrapa.br,
eferneda@pos.ucb.br, maaxt2002@ufrgs.br

Abstract. Knowledge discovery from databases, in the descriptive approach, includes clustering analysis (CA) as an alternative to estimate how a set of objects is organized in the space of their dimensions. The main objective in this task is to find "natural" groups that could exhibit some meaning. Considering the strong subjectivity that underlies this process, an important issue refers to the relationships among the CA players when looking for a model that could adjust the data. In this work, a model for actions coordination that provides an order to drive the relationships among CA players is presented. This model is presented as a conceptual contribution towards the construction of a computational environment to support effective conversations in a subjective context.

Keywords: Knowledge Discovery in Databases, Data mining, Clustering analysis, Action coordination.

1 Introduction

Departing from a set of objects, Clustering Analysis (CA) looks for a category structure that can fit in this data set. The aiming is to find "natural" groups, based in arbitrary internal criteria, in such a way that the cohesion among the members of a group would be the maximum and among the groups would be the minimum.

Grossly, the process of CA includes two basic steps: generating a clusters configuration and interpreting them in order to find some meaning in them. The first step is

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 483–490, 2008.

carried out by means of an algorithm, usually based in some kind of distance, which generates clouds of points. In the second step, specialists analyze these clouds aiming to find some meaning in the clusters. The second step presents a strong subjective bias, since it depends on mental models of the people (human beings) involved.

In this work we propose a model to deal with these subjective aspects in which a protocol based on speech acts is applied. This model provides a decision support process to build consensus and better articulated actions on the issues related to clusters interpretation.

The judgements and decisions from people involved with the process and the way they communicate on the elaboration of these thoughts and coordinate to make decisions, take actions and procedures is crucial for the planning cycle, execution and evaluation of the results from CA. These aspects can also be considered for application of data mining, multivariatre analysis, among others, guiding the relation between the people involved on the process.

2 An Overview on the Clustering Analysis

The whole CA process can be organized in nine steps (see Fig. 1): *(i)* domain and data understanding, *(ii)* definition of objectives, *(iii)* selection of relevant and discriminant variables, *(iv)* data preparation, *(v)* weighting definition, *(vi)* algorithm choice and configuration, *(vii)* algorithm application, *(viii)* results evaluation, and *(ix)* knowledge building and refining data structures. Notice that we assumed to apply a weighted clustering algorithm, as defined in [1].

In the first step a shared space of understanding about the domain and the data structure is built to enable the communication between the domain specialist and the data analyst. The former is related to the specific field in which the CA is being applied and the latter is the responsible for managing the whole CA process. While the domain specialist holds the knowledge regarding to the application area, the analyst master the methods, techniques and tools for CA. In the ideal situation they develop a synergy aiming to find a model that better adjust to the data.

In the second step, departing from a shared understanding space, they are guided to focus on defining the analysis objective.

In the third step the selection of variables are carried out taking into account their relevancy and how discriminant they are according to the analysis objective. Techniques like principal components analysis or factorial analysis [2] can be applied to figure out how discriminant is the selected variables. For short, low discriminant variables are those which values change very slightly among the objects, having a small effect in the clusters definition.

The fourth step is focussed in sampling, cleaning, and structuring the data set. The adequate treatment of missing values is also part of this step.

In the fifth step the components for the algorithm weighting is defined. In the informed clustering algorithm [1] an information matrix expressing the previous knowledge regarding to the application context and the data must be supplied as a way to introduce a domain bias in the clustering algorithm. This information matrix is built from a relationship (or cause-effect) mapping of the involved variables.

In the sixth step the clustering algorithm is chosen, according to the analyst or domain specialist negotiated preferences. The algorithm must be prepared to receive the information matrix, since it will provide the homogeneity coefficient that has to be considered in the clusters´ definition.

In the seventh step, the selected algorithm is applied in order to find a clustering configuration that can be seen as a candidate to represent the data structure. Many configurations can be generated until the specialist accepts it, according his experience in the domain.

In the eighth step the clustering results are evaluated. According to Cormack [3], many techniques exists that can be used to evaluate the quality of the generated clusters. There are two kinds of evaluation techniques for CA: the quantitative and

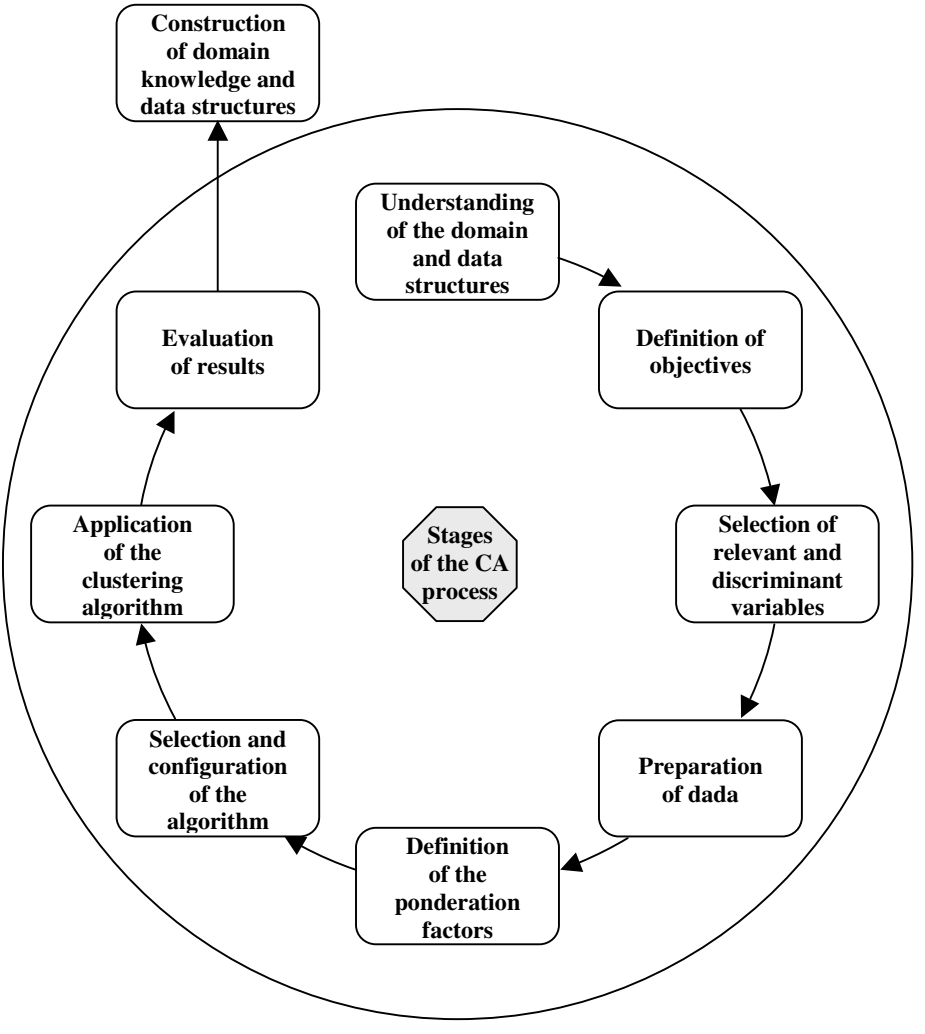

Fig. 1. Knowledge creation in clustering analysis

qualitative ones. As examples of quantitative techniques, Moreira [4], suggests the discriminant and the variance analysis. On the other hand, although, less precise, the qualitative approach cannot be ignored, since, by considering the huge amount of possible clustering configurations, one could argue that, in essence, the nature of the interpretation process is more qualitative than quantitative. According to this, in our view, the evaluation of results carried through this eighth step should consider both, the qualitative and quantitative approaches for this task.

The core of this paper is a roadmap to apply the qualitative approach that involves an intense and elaborated conversational agreement among the players. In the ninth step comprises the construction the knowledge that can include, beyond the application domain, the refinement of the own data structures. As it can be seen, this step is out the main cycle in Fig. 1. In a sense, this step can start another discovering cycle providing the input for the first step, in a spiral fashion.

3 The Actions Coordination Cycle

The conceptual basis for our proposal comes from [3], [4], [5], and [6], and is known as the actions coordination cycle. The actions coordination cycle has two phases: establishing a promise and promise accomplishment. The first one refers to the context creation and negotiation tasks, while the second one has to do with accomplishing the promise and the evaluation of the results derived from this accomplishment. There exist in the actions coordination cycle two agents involved when a promise situation occurs: the provider and the client.

The promise comprises the defined goals for the CA process. Precision and a explicit declaration for the customer is fundamental. Based on these defined (by the "client") and accepted (by the "service provider") goals, the results to be delivered should be marked with a statement of fulfillment in the form of a CA service accomplishment declaration. The client, once notified of this accomplishment declaration, should, in turn, declare a statement of satisfaction or dissatisfaction with the results just delivered, in accordance with his expectations presented at the begining.

An actions coordination cycle can be of two types, according to the nature of the speech act that starts it. It can be started by a request or by an offer. In both situations the provider and the client share a common space of interests and mutual commitments that is built from the expectations regarding the benefits that can come from the whole cycle. These expectations are supported by the reciprocal confidence that must permeate the relationship among the players.

Figs. 2 and 3 exhibit the schemas for the request and the offer cycles. In both cases a problem statement starts the cycle, beginning a context creation phase. In case of the request cycle, the problem statement is done by the client, based on his requirements for which satisfaction s/he depends on the provider. In case of the offer, the provider tries to meet what s/he figures out to be the client requirements.

Next, the negotiation phase starts after the request or offer statements have been posted and finishes with an acceptation statement. The acceptation statement in the request cycle is made by the provider and in the offer cycle is made by the client.

The next phase is the accomplishment, which begins with the promise statement and finishes with the accomplishment statement, always done by the provider. The

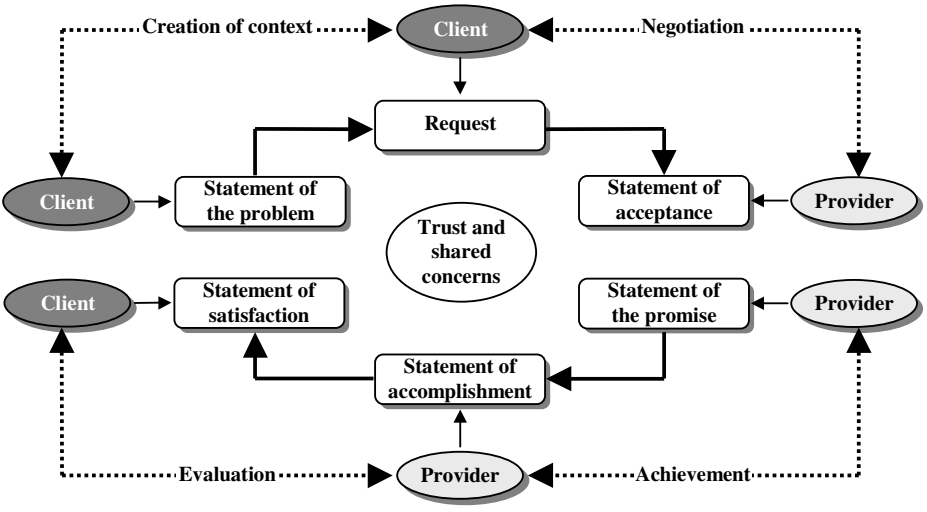

Fig. 2. The request cycle

fourth phase refers to the evaluation task and takes place after the provider declare the promise accomplishment, finishing with the satisfaction statement, always done by the client. This phase closes the request or offer cycles. However, not always these cycles end with the satisfaction statement. It may occur, depending how the previous phases were performed, that a client dissatisfaction statement may be expressed, closing those cycles in a non-effective way.

Notice that the differences between the request and the offer cycles are located in the upper side of the schemas. In the left-upper side of Fig. 2, the client behavior is characterized by thoughts regarding his necessities. Similarly, in Fig. 3, the provider is involved in thoughts related to the clients' necessities.

In the request cycle the client is in the two extremes of the context creation phase. He is responsible for the problem statement and for the sequence of speech acts (a conversation) that leads to the request. On the offer cycle, the provides plays a similar role, being in the two extremes of the context creation phase, when declaring the problem and the speech act that leads to the offer. These are the only important differences between the request and the offer cycles. In the lower sides of Figs. 2 and 3, the players' places and the nature of speech acts are the same.

The negotiation and evaluation phases are characterized by a bipolarity between the client and the provider, that are involved in a judgment sharing process in which an agreement with respect to the request or the offer is searched. Also, in this phase, a consensual evaluation of the promise accomplishment is desirable. These phases require parameters like action to be carried out, satisfaction conditions, and a timetable to accomplishment.

The context creation and the promise accomplishment phases are characterized by having only one player in their beginning and ending. For the request cycle, the context creation phase has the client in its both extremes and for the offer cycle this phase has the provider in its extremes. In addition, both cycles have the provider in the two extremes of the promise accomplishment phase.

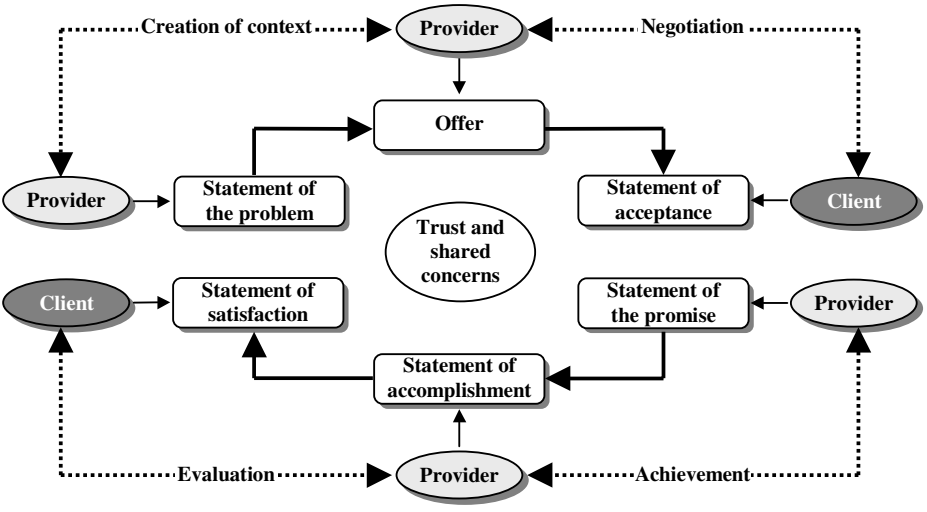

Fig. 3. The offer cycle

Notice that, in each phase of the actions coordination cycles it can be necessary to trigger new cycles in a commitment network, issuing, for example, new requests to other providers. This behavior was illustrated in Figs. 1 and 3 as entwined circles. In the heart of the cycles remains the shared confidence and concerns that are the basis for keeping the process cohesion. The weakening of these mutual feelings tends to provoke the process fragmentation.

4 Applying the Actions Coordination Cycle in CA

To approach the subjectivity in the CA process we propose to view it as an actions coordination cycle among the agents involved. The subjectivity in CA is mainly observed in the eighth and ninth steps of the process (results evaluation and knowledge building and refining data structures), since it is in those steps that human interpretations are more strongly present. However, it is important observe that, even in the other steps, there are different levels of subjectivity.

Ultimately speaking, the CA process, as any other process involving people, is a human process, that is, the subjectivity issue is not a peripheral one; it is central. So, we modeled the whole process applying the concepts presented in the previous section. An adapted schema from the actions coordination cycle to the CA process is shown in Fig. 4. It corresponds to the offer cycle in which the analyst plays the provider, while the domain specialist takes the place of a client. The analyst provides the knowledge creation from CA service.

The context creation phase corresponds to the domain and data structures understanding as a set up from the analyst to achieve a good interaction with the domain specialist. This interaction enables the next phase, the objectives definition. The analyst makes a first offer based in the necessities from the domain specialist and on the

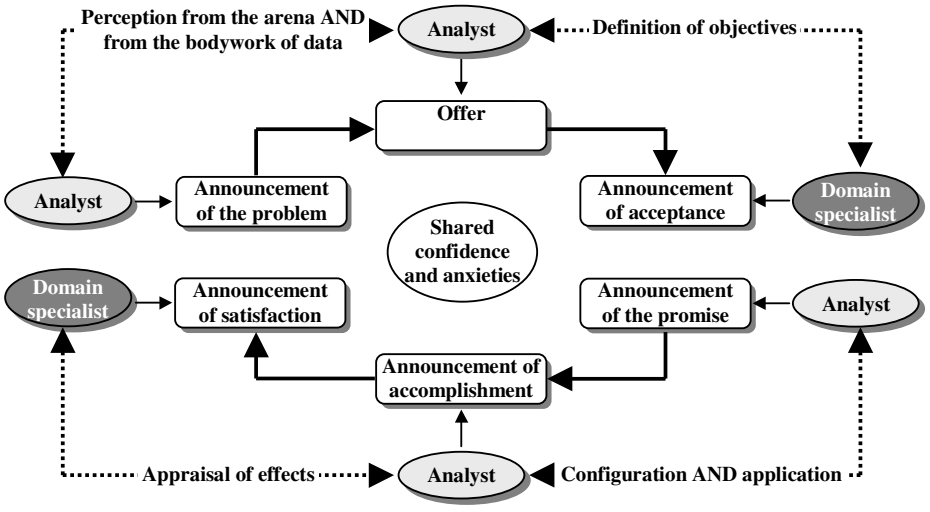

Fig. 4. The actions coordination cycle applied to CA

knowledge acquired regarding to the problem context. This phase begins with the problem statement to the analyst and ends with the first offer he does.

In the objectives definition phase a shared space of knowledge is created between the analyst and the domain specialist. This phase corresponds to the negotiation phase in which the negotiation focus is the objectives to be seek during the CA process. It ends after an interaction between both players in order to meet an agreement that leads to the acceptance statement from the domain specialist.

In the configuration and application phase, which corresponds to the accomplishment phase in the offer cycle, the analyst performs the variables selection, the data preparation, the definition of the weighting factors, the choice of the algorithm and its configuration, as well its execution. This phase requires a strong interaction between the analyst and the domain specialist and is completed with the results presentation to evaluation, after a promise accomplishment statement from the analyst.

In the results evaluation phase the analyst and the domain specialist put their knowledge, judgments, and beliefs in action looking for an enlargement of the shared knowledge.

The actions coordination cycle in CA problem can be repeated many times, by redefining objectives, renegotiating agreements, and so on, until a satisfaction statement is obtained from the domain specialist.

5 Conclusions and Ongoing Work

According to Echeverría [5] when we talk about coordinating common actions, we are talking about communication. Among humans language is a recursive coordination of behavior based on reflection and reasoning. The same author states that "conversations are the effective component of linguistic interactions – the basic language units" and emphasizes the importance of the actions coordination in a world in which the

auto-sufficiency is impossible. In this world, says Echeverría, we have to learn how to cooperate to coordinate actions. In this sense and in our point of view, the study and application of the actions coordination cycle in the CA process may help to promote a consensual understanding in a subjective learning context, enabling to feed a vast commitments network. The ongoing work includes both the application of this model for performance evaluation in public sanity companies and the development of an environment for conversation support in clustering analysis.

References

1. Castilho, W.F., Prado, H.A., Ladeira, M.: Informed k-Means: a Clustering Process Biased by Prior Knowledge. In: Seruca, I., Filipe, J., Hammoudi, S., Cordeiro, J. (eds.) ICEIS: Proceedings of the 6th International Conference on Enterprise Information Systems, Porto, Portugal, vol. 2, pp. 469–475. INSTICC Press (2004)
2. Dunteman, G.H.: Principal Components Analysis. Sage Publications Inc., USA (1989)
3. Cormack, R.M.: A Review of Classifications. JRSS, A 134, 321–367 (1971)
4. Moreira, T.B.S.: Financial and exchange crises in Asia in 1997-1998. Unb, Brasília, Brazil (2001) (in Portuguese)
5. Echeverría, R.: Ontologia del Lenguaje, 4th edn. Dolmen, Santiago, Chile (1997)
6. Flores, F.: Management and communication in the office of the future. PhD. Thesis, University of California at Berkeley (1981)
7. Flores, F.: Creando organizaciones para el futuro. Dólmen, Santiago, Chile (1996)
8. Kofman, F.: Metamanagement – The New Conscious Business. Antakarana Cultura Arte Ciência, São Paulo, Brazil (2002) (in Portuguese)