Contributions of KDD to the Knowledge Management Process: a Case Study

Paulo de Tarso Costa
de Sousa
Superior Electoral Court
SEPN 514 Ed Anexo II
55-61-340-3863
ptarso@tse.gov.br

Hércules Antonio do
Prado
Brazilian Agricultural
Research Corporation
Center for Agricultural
Research on Savannah,
Caixa Postal 08.223 –
CEP: 70770-000 -
Planaltina, DF, Brasil
hercules@cpac.embrapa.
br

Eduardo Amadeu
Moresi
Catholic University of
Brasília
Graduate Program in
Management of
Knowledge and
Information Technology
CEP: 70.790-160 -
Brasília, DF, Brasil
moresi@ucb.br

Marcelo Ladeira
University of Brasília
Campus Universitário
Darcy Ribeiro - Asa Norte
CEP: 70910-900 –
Brasília – DF - Brasil
mladeira@cic.unb.br

ABSTRACT

Knowledge Discovery in Databases (KDD), as any organizational process, is carried out beneath a Knowledge Management (KM) model adopted (even informally) by a corporation. KDD is grossly described in three steps: pre-processing, data mining, and post-processing. The latter is mainly related to the task of transforming in knowledge the patterns issued in the data mining step. On the other hand, KM comprises the following phases, in which knowledge is the subject of the actions: identification of abilities, acquisition, selection and validation, organization and storage, sharing, application, and creation. Although there are many overlaps between KDD and KM, one of them is broadly recognized: the point in which knowledge arises. This paper concerns a study aimed at clarifying relations between the overlapping areas of KDD and knowledge creation, in KM. The work is conducted by means of a case study using the data from the Electoral Court of the Federal District (ECFD), Brazil. The study was developed over a 1.717.000-citizens data set from which data mining models were built by applying algorithms from Weka. It was observed that, although the importance of Information Technology is well recognized in the KM realm, the techniques of KDD deserve a special place in the knowledge creation phase of KM. Moreover, beyond the overlap of post-processing and knowledge creation, other steps of KDD can contribute significantly to KM. An example is the fact that one important decision taken from the ECFD board was taken on the basis of a knowledge acquired from the pre-processing step of KDD.

Keywords

data mining, knowledge management, knowledge discovery in databases.

1. INTRODUCTION AND MOTIVATION

Practitioners on Knowledge Management (KM) usually adopt some model to guide the process and this model describes the activity of knowledge creation. On the other hand, the Knowledge Discovery in Databases (KDD) practice has been supported by methods like CRISP-DM [1], that include an interpretation activity. The clearest overlap of KDD and KM is the knowledge creation phase in KM and the interpretation task in KDD. In this

paper we explore relations between KDD and KM phases by means of a case study in order to set a scene in which we show the specific importance of KDD to KM and try to make smoother the road of integration of both areas. It is shown how KDD can generate relevant input to KM not only from the data mining and post-processing tasks, but also from the pre-processing activities. Observations taken during tasks like problem definition, data acquisition and cleaning, and data and algorithm engineering can lead the analyst to be aware of some organizational failures. These failures, if not properly analyzed in a KM context, cannot receive the deserved importance, keeping the organization from benefiting from preventive or corrective actions. The discussion flows on the basis of an application on the Brazilian election domain, with focus on the organization data rather than in voting problems. CRISP-DM [1] method was applied to conduct the process while Weka [17] suite was adopted to build classification and clustering models. Specifically, a *k*-means algorithm was used for clustering and decision trees for classification. The reference model we used for KM is the generic one proposed by Stollenwerk [14] shown in Figure 1.

Figure 1 – Generic KM model of Stollenwerk

This model considers four aspects that compose the environment of KM: leadership, culture, technology, measures and compensations. Inside this environment, seven activities related to the generation of organizational knowledge are proposed. They

are: identification of abilities, knowledge capture, selection and validation, knowledge organization and storage, knowledge sharing, knowledge use, and knowledge creation.

This paper is organized in four sections. The first one discusses the importance of KM to KDD and vice-versa. The second one outlines the organizational context where the case study was developed. The third one depicts the applied methodology [13] based on the trade-off between the KM Model of Stollenwerk [14] and the CRISP-DM methodology. In addition, the case study is also described in this section. Finally, the fourth section presents some findings of this research and points out some future works.

2. KDD AND KM

According to Peter Drucker [3] the most important challenge posed to the knowledge society is the development of systematic practices to manage the self-transformation that is based on organizational learning. In the same direction Senge [12] argues that the manager should follow five disciplines to develop an organization that learns: systematic reasoning, personal control, mental models, sharing views, and team learning. Moreover, KM requires an intense human interaction to generate, storage and share organizational knowledge.

KDD can be seen as any other process that needs to access the burden of organizational knowledge in order to have each task performed efficiently. Davenport and Prusak [2] emphasize the fact that KDD technology is part of KM, working as a lever to it. This technology plays an important role as it brings a methodology for the expert to perform the extraction of knowledge from the available databases.

On the KM side, Stollenwerk [14] discusses the existence of many models to describe KM. The author stressed these models looking for a synthesized alternative and, in all of them, identified the processes: identification/conceptualization, acquisition, selection and validation, organization and storage, sharing and transference, application and use, creation, products and services engineering based on new knowledge, and evaluation of the benefits and the value of the knowledge. On the other hand, Sveiby [15] classify the KM approaches as regards information management and organizational learning and abilities management. Based on these ideas, Stollenwerk coined a generic model, which focusses on organizational learning and abilities management. Our work is based on this model, which we extended [13] to show the importance of KDD to KM.

3. ORGANIZATIONAL CONTEXT

The Electoral Court of the Federal District (ECFD) belongs to the Brazilian Justice and is in charge of the election process in the Federal District, ensuring law enforcement. The Court has performed important services to the community and the parties.

In the KM context, the ECFD is working to map its human potential to generate knowledge by using the internal abilities and a set of databases. Also, a mapping of the external requirements has been carried out that cause an important impact in defining changes in the internal process. As shown in Figure 2, KDD can be naturally integrated to the KM model of ECFD, that enables the activation of the mental processes [9] that, in turn, lead to pro-active interventions from the Court board.

Figure 2 - Relation of KDD and KM

It is important to emphasize how crucial the involvement and the support from the board to the knowledge creation process is. According to Terra [16], the definition of knowledge areas to be explored by the organization and the statement of general views that can conduct innovative projects are part of the fundamental role of the board.

4. CASE STUDY

The case study comprises the application of our methodology, distilled from an analysis on the relations of KDD and KM. This analysis, carried out by Sousa [13], produced an integrated model (see Table 1) by comparing the steps from KDD with the generic KM model of Stollenwerk [14].

Table 1 – Integration of KM and KDD

#	Integrated model	Generic models	
		KM Model [14]	CRISP-DM
1	Problem and context definition	Identification of abilities	Business understanding
2	Capturing the required inputs to the problem solution	Knowledge capture	Data understanding
		Selection and validation	
3	Processing the inputs to the problem solution		Data preparation
4	Defining techniques and algorithms for knowledge extraction		Modeling
5	Expert assessment and interpretation of data mining results		Assessment
6	Structuring the acquired knowledge	Organization and storage	
7	Publishing results	Sharing	Deployment
8	Effective use of the acquired knowledge	Application	
9	Global assessment	Knowledge creation	

4.1 Methodology description

The steps from this methodology and the corresponding descriptions are shown next.

1) Problem and context definition– identification of data, information, and knowledge needs. As proposed by Stollenwerk [14], when discussing the identification of abilities, the problem to be solved has to be outlined before the definition of the required knowledge that can include knowledge not yet available. The focus in this phase is the definition and specification of the problem, along with the proposition of a possible solution.

2) Capturing the required input to the problem solution – identification of the knowledge sources. The selection of these sources is closely related to the target problem. Organizational knowledge, information, or data required to the problem solution are captured and explored in this phase.

3) Processing the inputs to the problem solution – this phase corresponds to knowledge filtering, redundancies elimination, conflict solution, and adequate data treatment to feed the data mining process, beyond data normalization and summarization.

4) Defining techniques and algorithms for knowledge extraction – selection of algorithms and tools to be used on the data mining process. The model building task is carried out in this phase.

5) Expert assessment and interpretation of data mining results – in this step the expert evaluates the potential pieces of knowledge based on the model resulting from the data mining process. After evaluation and interpretation new problems may arise that can force the return to previously performed steps. Also, new knowledge related to the problem context can be identified.

6) Structuring the acquired knowledge – knowledge organization and storage. Information technology should provide the most suitable resources for knowledge storage and recovering.

7) Publishing results – strongly related to the previous step, it aims at providing support to knowledge sharing for the organization members, from whom a feedback is required to improve the process. The elaboration of distribution plans and knowledge monitoring and maintenance are also part of this step.

8) Effective use of the acquired knowledge – concerns the application of knowledge to generate new products and services or even to aggregate value to already existing commodities.

9) Global evaluation – this evaluation enables the organization in determining new necessities of knowledge, learning, innovation capacity, research, and publication of ideas. All of these aspects are dimensions of the knowledge creation process.

On the basis of the model resulting from the method above and clients records, we will now build the electors' profile, aiming at identifying a behavioral pattern when looking for services from ECFD.

ECFD is the Justice branch in charge of election process in the Federal District, including legal and operational issues, in order to assure law enforcement. Personal sources of knowledge

are spread out on the court structure, mainly in the notary's offices. Another important source of knowledge is the electors database in which historical information is recorded.

Although informally, ECFD, through an intuitive KM model, is working to know its human potential to generate organizational knowledge by taking advantage of its internal abilities. Beyond the mapping of abilities inside the organization, the identification of external necessities is also carried out. External factors influence decisively the changing of internal processes as they come, mostly, from the clients demanding new services. As shown in Figure 2, questions related to KM, specifically the knowledge creation, and KDD are strongly connected in a conceptual model in the ECFD. The results of this connection feed the activation of the mental process that, as described by Pereira [9], lead to the solution of organizational problems which, in our case, correspond to improvements on the electoral branches and the Court Secretariat and in public relationship as well.

4.2 The case approach

The text in this section deals with the sequence of the methodology as described in the previous section, although we adopted a more natural presentation with no itemization. For the sake of simplicity, we show only a part of the complete application. This is sufficient to make evident the importance of KDD to KM. We have shown the analysis that was performed over a subset of the complete electors database, taking into account those that were identified as having the problem of "coincidence" that is described later.

Firstly, specification should be made of the mission and objectives of the ECFD, as well as the identification of problems that will conduct the data mining process. This step was carried out with the support of experts from the target area of the organization. According to Pyle [11], identifying the problem, exploring the problem space, and recognizing a solution are the most important phases in a data mining application. By means of interviews with the experts, organizational problems and some hypotheses of solution were enlisted that should be explored during the data mining application. By providing answers to the posed problems, the triggering of new insights to improve the electoral process is expected.

Next, it is necessary to gather relevant data for the problem solution. From the data analysis and understanding, the experts will identify previously unknown entity relationships. Thus, it is important to identify sources of data, hardware and software platform, beyond the tools to transform the data. In this sense, the electors database with 1,717,000 entries as the main data source and the Weka suite for developing the data models were defined. As external sources of data we used the national census data. The software and hardware applied were a Dual Pentium III, with 1 GHz, 2 Gbytes of RAM, and an HD of 40 Gbytes.

Considering the legal restrictions for using such data, it was necessary to obtain a formal authorization from the ECFD board to develop the application. The authorization granted excluded the use of personal data, enabling just summarized analysis.

After defining ECFD mission and objectives, the data sources, and Information Technology support, the next step is to perform the data mining tasks that requires an intense interaction with the experts. Han and Kamber [5] emphasize the importance

of the expert to answer the many organizational questions that arise during the data mining process.

During the pre-processing phase important problems were observed that require interventions in some organizational processes. Namely, it was observed:

- a) Inconsistencies on data: invalid dates, attributes with incorrect values.
- b) Violation of referential integrity.
- c) Missing data.

A significant effort was applied to solve problems like: restoration of relationships, fulfilling of missing data (or even discarding records with such problem), and correction of mistakes in data. According to Pyle [11], quality on data is a pre-requisite to build effective models on data. In fact, it goes in the same sense of the expression GIGO – for Garbage In, Garbage Out. Also, in the data understanding and preparation phase there occur the combination phenomena [8] in which while new knowledge is created the experts develop new views that are stimulated by the patterns found in the data modeling task.

Some preliminary summarization was carried out that leads experts to be aware of important aspects of the population. The summarization task includes the following results:

- a) Distribution by age intervals, pointing out that the majority of electors are concentrated between the ages of 25 and 59 years.
- b) Distribution by instruction level. 30% of the population did not complete even the basic school years (aged 7 to 13).
- c) Distribution by civil status, including single, married, divorced, and widowed electors.
- d) Distribution of electors by sex shows a great incidence of female people.

In this point it is interesting to observe how new insights about organizational problems come up while performing typical pre-processing tasks, like data understanding, cleansing, and transformation. This fact makes clear how important the role of both data mining analyst and domain expert are during the process.

The generic KM model of Stollenwerk [14], after the selection and validation phase flows to the knowledge organization and storage phase in which the main subject is the already acquired knowledge. There is a gap between these two phases that we fulfill with KDD tasks that generates organizational knowledge from all of its steps.

Two tables, one of current data about electors and other with historical data concerning the same electors, were generated to feed the KDD process. The kind of KDD task to be performed was defined after agreement with the domain expert. A classification model based on the current data and a clustering model to describe the historical data were defined. Classification was applied to explain standard characteristic of electors, while clustering was applied to discover standard behavior of the electors involved in coincidences. A coincidence means that the elector count twice or more times in the data base with different keys. It could occur as a result of marriage, for example, when the women used to change their names and update the records. In some cases, instead of simply updating the records, some women

ask for a new document and the system was not prepared to cope with this problem in time.

The domain expert formulated some questions and explained hypotheses that we tried to verify by the data mining tasks. The first model created was a decision tree based on the first data set according to the following characteristics:

- 1) Chosen algorithm: C4.5 (J48);
- 2) 16.952 instances corresponding to the electors involved in coincidences;
- 3) Confidence: 90%;
- 4) Number of selected attributes: 7 attributes;
- 5) Target class: SEX;
- 6) Test mode: 10-Fold Cross-Validation;
- 7) Minimal number of instances by leaf: 2;
- 8) Expected index for model acceptance: 75%;
- 9) Pruning option: post-pruning.

The clustering task was performed over the same data set adopting the K-Means algorithm (SimpleKMeans option in Weka) with $k=8$ (number of clusters, based on the expert experience) and considering 11 attributes.

Six clusters were generated, with a distribution according to Figure 3.

Figure 3 - Centroids distribution

The clusters were internally analyzed in order to provide clues for the interpretation task. The analysis was based on histograms issued to show the most important attributes for each cluster. The experts accomplished the characterization of each cluster by applying previous knowledge and the results from the analysis.

The model evaluation was carried out by comparing them to the objectives defined at the beginning of the work. Possible deviations were analyzed to check the necessity to return to already performed phases.

Possible consequences of the knowledge acquired from the models interpretation were stressed. Actually, these consequences are the most important results from all effort, since they are the findings that will impact the organization.

In this moment a complete revision of the process is necessary in order to identify possible failures and to promote a

reflection regarding the work methods and the information organization that surely increase the stock of organizational knowledge.

The identified patterns represent inputs to the knowledge creation model of Nonaka [8], in which one search new knowledge by socialization, combination, publishing, and assimilation of the experiences. It can occur anywhere in the described process.

In this sense, Stollenwerk [14] propose that, after the selection and validation phase, the acquired knowledge should be organized, stored, and made available to be shared along the organization.

After being validated, the models are published to the target area in order to provide useful support for its activities. As the present case study is focused on the quality of the electoral records and the electors' behavior, both the areas of data management and public relationship took advantage of the results, improving their organizational processes.

5. CONCLUSIONS AND FUTURE WORK

The case study developed in this paper provided expert users with a new perspective about the importance of informational analysis and available knowledge to keep in pace with the organizational objectives. At the first phase of the process, in which the analyst looks for a definition and context of the problem, he can perceive the necessity of identifying available knowledge, information, and data that can help in that definition. The expert in this phase is a decisive player to arrive at the problem definition. Pereira ([9] *apud* Polya [10]) assigns a utmost importance to the problem understanding activity in which the expert define the relevant data and general conditions (e.g., technical conditions).

Like abilities, data and information have their gaps that must be identified and, possibly, corrected. Authors like McGee and Prusak [6], and Pyle [11] regard the identification of needs and problems as the most important of an information managing process. Also, in order to take advantage of existing resources, their organization and availability are conditions that must be achieved to share abilities and make them useful throughout the enterprise.

Each step of the methodology adopted in the case study has the knowledge creation feature that can produce new views. These views can be assessed and can generate useful feedback.

So far as this is a study of knowledge management focused on Sveiby [15] approach, in which knowledge represents objects and is guided by information science, the phases three and four of the integrated model (see Table 1) have a more technical than managerial characteristic, requiring a deeper involvement of the data mining expert. However, the business expert was directly involved in this process so far as he could perceive some new knowledge and the need of turning back to previous phases.

Based on the evaluation and interpretation of the data mining results, it was quite clear to the experts the necessity of information improvement, focusing on information quality and on the establishment of effective controls and maintenance. The data mining process contributed in leading to the discovery of electors'

profiles and the need of tools to get up-to-date data and an information quality control on them.

The ensuing results can be made available on intranet since it is the most popular place in which users can find the knowledge they need. Then, the necessity of a portal arises naturally. In this point, technology gets aligned with management to provide people with a more effective way to access just-in-time knowledge. Note that knowledge obtained from KDD arises from the model interpretation. Beyond this knowledge, a lot of lessons are held during all the process. From this point of view, knowledge sharing, supported by information technology, will provide a process of introspection that will contribute to the creation of new ideas for the services offered by organization to society.

By analyzing the discovered rules, one can define many paths of action inside and outside the organization. The actions to be taken can reduce operational costs to the organizational process maintenance and problem resolution. In short, sharing knowledge is one of the permanent actions that must be taken by organizations that adopt knowledge management [7]. Without knowledge sharing, it is quite impossible to provoke the mental processes to the creation of new knowledge.

6. REFERENCES

- [1] CHAPMAN, P. *et al.* The Current CRISP-DM Process Model for Data Mining. CRISP-DM Consortium. Discussion paper. Available in <http://www.crisp-dm.org> (Feb 2003).
- [2] DAVENPORT, T. H. and PRUSAK, L. Working Knowledge. HBS Press, 1998.
- [3] DRUCKER, P. F. Post Capitalist Society. New York, NY: Harper Business Books, 1993.
- [4] FAYYAD, U. M., PIATETSKY-SHAPIO, G., SMYTH, P. and UTHURUSAMY, R., EDITORS. Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press, 1996.
- [5] HAN, J. and KAMBER, M.. Data Mining: concepts and techniques. Morgan Kaufmann Publishers, 2001.
- [6] MCGEE, J.; PRUSAK, L. Managing Information Strategically. New York, NY: John Wiley and Sons, 1993.
- [7] MORESI, Amadeu Dutra Moresi. Information and Knowledge Management. In: Tarapanoff, Kira (Editor). Organizational and Competitive Intelligence. Brasília, DF (Brazil): Editora Universidade de Brasília, 2001. (In Portuguese).
- [8] NONAKA, I.; TAKEUCHI, H. The Knowledge-Creating Company. New York, NY: Oxford University Press, 1995.
- [9] PEREIRA, W. C. de A. Solving Creative Problems: triggering the thinking ability. Brasília, DF (Brazil): Departamento de Informação e Documentação/EMBRAPA, 1980. 54p. (In Portuguese).
- [10] POLYA, George. How to solve it: a new aspect of mathematical method. New Jersey: Princeton University Press, 1945.

- [11] PYLE, Dorian. Data preparation for data mining. Academic Press, 1999.
- [12] SENGE, P.M. The Fifth Discipline: The Art and Practice of the Learning Organization. New York, NY: Doubleday Books, 1994.
- [13] SOUSA, Paulo de Tarso Costa de. Data Mining applied to the induction of a Knowledge Management Model. Dissertação de Mestrado. Universidade Católica de Brasília, 2003.
- [14] STOLLENWERK, Maria de Fátima Ludovico. Knowledge Management: concepts and models. In: Tarapanoff, Kira (Editor). Organizational and Competitive Intelligence. Brasília, DF (Brazil): Editora Universidade de Brasília, 2001. (In Portuguese).
- [15] SVEIBY, Karl-Erik. What is Knowledge *Management*? Site: www.sveiby.com.au (Sveiby Knowledge Management). Visited in July, 2002.
- [16] TERRA, José Cláudio Cyrineu. Knowledge Management – the great enterprise challenge. [S.l]: Negócio Editora Campus, 2nd Edition, 2001. (In Portuguese).
- [17] WEKA 3: Machine Learning in Java. Versão 3.2.3. Open-source software available in <http://www.cs.waikato.ac.nz/~ml/eak/index.html>. Visited in Feb, 2002.