Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Management of R & D projects using a data mining

M. V. Rodriguez y Rodriguez[1] & N. F. F. Ebecken[2]
[1]*UFF – Federal Fluminense University, Brasil*
[2]*UFRJ – Federal University of Rio de Janeiro, Brasil*

Abstract

In a global environment, where there is intensive competition among companies, the innovation process, the technological development and the intangible administration values, come as a need for the warranty of competitiveness in this new context.

Thus, this work seeks a methodology for establishment rules to aid in the technological project priorities, using as a case object study the Research Center - PETROBRAS, an oil company in Brasil. For this work, technological projects data concerning physical, cost and customers evaluations, were used for the technological projects priority.

As future research is developed, some computational tools will be constructed, based on experts systems that can aid the technology projects priority, normally with high risk and high reward, besides susceptibility to a great number of external variables.

1 Introduction

This work is related to knowledge mining from databases, knowledge extraction, patterns analysis and data warehouse, using a database with hundreds of technological projects, concluded and evaluated by the customers. In this case, it constitutes an enormous potential source for obtaining metrics related to these technological projects.

Thus, the objective is to define the data interpretation process to obtain success and failed projects rules. It represents an opportunity to obtain hidden knowledge starting from data stored along years by the Company.

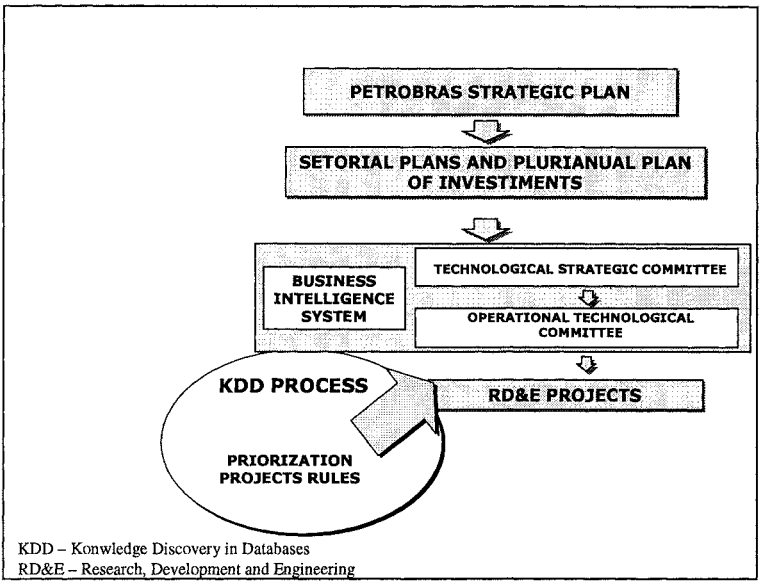


Figure 1: Process and methodology used.

The premise considered in this data analysis, is that the success of a technological project is directly related to the fact of it has ended and been implemented. Thus, the innovative character of this work meets in developing a methodology, which uses the process to obtain knowledge through data mining process, with the identification and presentation of the rules for supporting the technological projects priorities, as the schematic represented in fig. 1.

2 Knowledge discovery in database process (KDD)

The problem definition is the first stage of data mining process which represents, on average, 20% of the total invested time in the KDD process, but it means 80% in importance for the process as a whole, as represented in fig. 2 [8].

In the study presented, five months were invested in the problem investigation to be solved. For this, several interactions with the customers were needed, besides the potential computational tools identification to the problem solution and one of the most fundamental point of the whole process, that is: the data to be used for the study. Besides this, there were potential problems related to obtain an integrated and consistent database, faced during the data collection, selection, cleaning, integration and transformation.

In this phase some computational tools to be used were identified, but in this stage the problem identification was more important and the way it could be solved, considering the available computational tools for each problem.

First, some computational tools were obtained, but it was more important, in this step, to discuss what problems were to be solved and the way to obtain the

solution. In the case of that study presented, it was more important to know the business rule obtained and the methodology identification that permit prospect the remained data tendency.

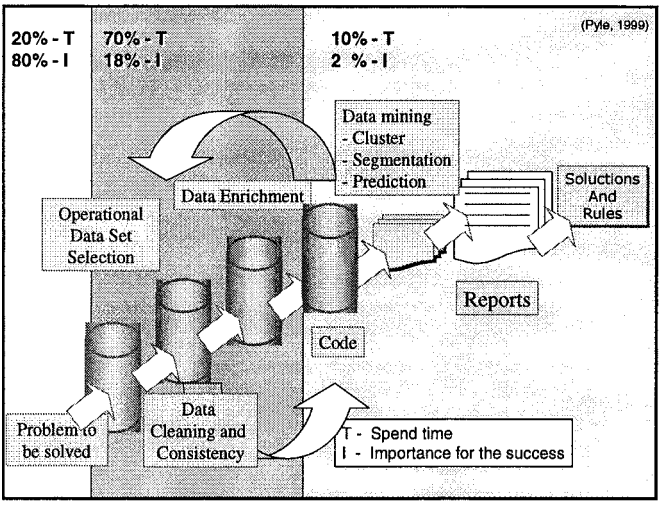


Figure 2: Knowledge discovery in database.

2.1 Data selection

In the data selection process the difficulties that happened were very different, mainly by consequence as the project data storage evolutionary process happened in a discontinuous way.

The data selection from the period 1996 to 2000, contained the following data: Evaluation Projects Reports available in paper; Annual Activities Summary Reports; Technological Projects Database in MS Access; and finally, Technological Projects using Ingres Database - Accompaniment System Project of PETROBRAS. This database was started from 1996, allowing that the feeding and it consults can be done starting decentralized in the Company.

The physical and cost project data were consolidated in a unique database contends: 619 project registrations with 37 data fields, (22.903 in total). The data includes a started project initiated in 1996 up to 2000 and finished projects that begun since 1996 up to 2004.

The data evaluation project for the customers were consolidated in a database contends: 714 project registrations with 52 data fields, (37.128 in total). These data corresponding a started project that begun since 1983 up to 1998 and finished project that begun since 1985 up to 1999.

In spite of the collected data storage during five sequential years, the data evolutionary identification and storage process were a consequence of the technological function administration process evolution. Thus, the database project stored in 1996 represents at the most 10% of the available database

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

project in the 2000, that is, the data structures and contents were increased a lot during this period. If compared the several years, when the data were collected, the related problems were identified as follows:

Lack of data integration: Some data were stored in MS Access database, others migrating for an Ingress database and other just in paper, as for example, the technical reports of projects evaluation.

Lack of current data: There were, even for the own data structure evolutionary process, non up-to-date data, or lost or even dirty. This fact is fully understood due to the gradual process with the distributed data feeders assuming this function along the database implementation during the last five years.

Data update time: Some data delayed a lot for to be up-to-date, per times, due to the own project characteristic this occurs several times after the effective project execution identification by the Company.

Lack of a long period vision: As the data structure construction were developed at every year, the vision in terms of the data structure were restricted to that moment needs, only considering that segmented vision and per times as a completely a non systematic process. For example: the relative questions to the projects evaluation are very pertinent to the execution process at the end of the project, but it didn't totally meet integrated with the initial prioritized project stage.

Different meaning: Some different data meaning processed, even for it be highly subjective, besides the quantification difficulties, as for example, the potential benefit of a certain project that depending on the understanding could assume any value. For example: a new technology that made possible an entire enterprise could be considered as tends a benefit in the value of the total enterprise, because without the same the enterprise would not be possible. On the other hand, to make possible this enterprise, not only this new technology that made it possible, it depends of several others components, like the physical and financial resources.

2.2 Cleaning, consistency, enrichment and coding

The next stage, after the data collection and selection, was the cleaning, consistency, enrichment and coding. It occurs, in this case study, according the following steps: Data cleaning and consistency - occurred at 2000 January to August; Data enrichment - occurred at 2000, May to August; Data coding - occurred in 2000, May to August. In this data adjustment process each stage happened according the following steps: cleaning and consistency, enrichment and code. However, this cycle happened for at once, since after the first cycle, were verified an increased result obtained at each new analyses cycle, for so much, were generated more than 30 decision trees and to each one tree generated a new enrichment and code could be happen in function of that analysis.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

3 Data mining

During the data mining process more than 50 decision trees were built, where starting from the top or root, when the data were divided in each level, forming new branches and finally the leaves, were created. In the data mining process, in a first moment, the decision tree were generated and pruned, after this, if it necessary, some additional knots were function of the final objectives.

For robustness verification effect the obtained rules in at least 40% of the valid data, representing 80% of overall database, the rules should be tested and validated in the rest of the database (20%), being obtained positive results and non significant alterations, that is to say, varying of 0-10%, for the selected rules, the result can be considered positive.

3.1 Solutions and alternatives

More than 80 rules were obtained from a database project using decision tree technology. From the 85 total rules initially obtained, some were discarded: obvious or little importance. From this total, 63 rules were selected and a consolidated group of rules was obtained, related to the projects concluded and not implemented, canceled or interrupted.

3.2 Analysis of the success and failed project rules

Related the rules obtained for the aid in the prioritization of the technological projects, especially those that obtained success, that is, concluded and implemented projects or in implementation phase, and those failed, the following analyses were performed [4; 6]:

Interestingness - As the *interest* of the rules: In this case, the obtained rules possess an usable logic when correlated to the project final situation, that is, success or failed, canceled or interrupted. In the case of success projects we considered that they were ended and implemented or in implementation process.

Minimum set of rules *covered*: A group of 63 rules was selected, collected from the largest number of examples of the data group. From these 63 rules, three rules were selected for projects that didn't obtain success, i.e., failed. In addition, for obtained success, eleven rules were selected.

***Action* rules:** The obtained rules were considered useful to the use, because they aid the users in the identification of possible failure and success projects. Using one of these rules, were observed that the projects with implementation time inferior to one year and with a total man-hour below to 138, developed with own resources, possess a 100% of chance to being ended and not implemented, canceled or interrupted.

Unexpected **rules:** For example, for success projects some of the obtained rules were related to the own resources of the Company. In this case, an unexpected rule result was obtained.

For non-success or failed projects the obtained rules presented an unexpected factor related to the customer participation in the process, (34 rules): the customer's participation reduces the superior man-hour limit from 630 to 504; representing that the customer project participation ends up reducing the worn-out resources in the same, probably for the fact of a better redirection to the customers needs, reducing losses of time and resources.

Completeness**:** The completeness is related to the total examples that were rules works, considering all data set collected and analyzed. In the identified rule, the completeness varies from 1.2% to 6%. In spite of this indicator to be low, the hit rate the rule occurrence is extremely high, that is, 100% in all cases.

The rule completeness is represented as following:

$$\text{Completeness} = \frac{\text{Number of covered examples for the Rule}}{\text{Total number of examples}}$$

The rule completeness was related to the total number of examples that the same ones collect in a data set. In the rules identified, the completeness varies from 63.7% to 3.5%. In this case, the completeness index is very high, especially, if considered the hit rate occurrence that varies from 80.3% to 100%, that is, quite high, turning a very robust rules group.

Precision: Considering the data used and the necessary precision degree to the selection projects process, the obtained and consolidated rules were adequate.

Comprehensibly: A degree of inferior comprehensibly is adapted for five (5,0) when used the formula for the complexity rules identification, which is: 0,6 x number of rules (=3) + 0.4 x clauses number (= 4 (5 + 4 + 3) / 3)), that is equal to = 1.8 + 1.6 = 3.4. Therefore, below the recommended value that is of 5.0 [4; 2]. Demonstrated that the rules have ***interest*** approaches, ***precision*** and ***comprehensibly*** then, the selected rules group is usable and possess a logic adapted for the human understanding.

The obtained rules possess an intelligible complexity degree to the human being. The rule complexity is measured by the clauses number that the same possesses associated to the total rules number, that is: A smaller complexity than five are considered appropriate for the human beings to have a perfect understanding of the presented rule.

In this case, to calculate the complexity were considered the formula for the rules complexity identification, which is: 0.6 x number of rules (=5 rules groups) + 0.4 x medium clauses number (= 2.54), that is the same to 3.0 + 1.02 = 4.02.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Thus, below the recommended value that is equal to 5.0 [4].

In this study, being demonstrated the interest approaches, precision and comprehensibility then, necessary so that the selected rules group is usable and possess a logic adapted for the human understanding.

The great advantage of the R&D projects definition selection approaches for a decision forum is the clarification of the Company goals and objectives [5].

Besides, the projects proposition considering these approaches guarantees the focus in the business for all the selected projects. An articulated and concise selection process, with defined approaches, facilitates the relationship among the project proposition, the decision and the execution, once were clear for everybody the reasons of the project acceptance or rejection [3].

Rules Complexity = 0.6* total rules + 0.4*total clauses

Therefore, with the rules obtained starting from the KDD process it is available an information group for the improvement of the technological current process project prioritization.

4 Conclusions and final recommendations

This case study was strongly built, starting from the experts, using tacit knowledge, facts and data registered during years, demonstrates that is possible starting from facts, data and the tacit knowledge to identify behavior patterns that cannot be clear in the first moment. The knowledge can be used to establish the priorities of technological projects subjected to great internals and externals variables, improving the results even in high risk investments level, (as in R&D projects).

For so much, it is necessary to work systematically in the data structure to be collected, taking into account this future vision to be considered, that will be not only for specific project direction and corrections, but for the improvement in the patterns behavior identification and the projects definition that tends to be success or failed.

Thus, considering this case study, the following observations are valid:

The customer's importance: It is of fundamental importance to identify and to classify the technological projects customers and the possible customers dreams for the solution of the subjects related to the processes, products and market.
This means more important than to have a high level technology; it is necessary to consider the client needs. By the other way, this can be observed considering the obtained rules when the customer participation is higher. The obtained results are more positive than in other cases where the customer participation project development is drop.

With facts and data, it is possible to identify important similarities for the improvement of the management processes. This means that the way as the

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

resources is demanded for the project execution and its implementation process can bring strong signals of the best way for conduct a technological project in the future. Thus, the systematic data collection and analysis can bring a great contribution in terms of the best management practices. As well as the customers, the projects contain certain similarities that can be explicit starting, in many cases, from a KDD process.

Rules validation: The need to build and to maintain a project database, turns the rules validation obtained by the data mining methodology of great importance, considering that the processes and the context change at each moment, and the obtained rules are the result of this dataset that represents these internal and external factors. Its systematic validation considering the dynamics and the continuous change of the market is of great importance to continuous rules improvement.

Tacit and hidden knowledge accumulation: The methodology proposal allows the hidden knowledge in data base and the tacit knowledge externalization, creating a continuous accumulative improvement process, being a powerful tool to aid the high risk and usually high reward project prioritization for the ones that obtain success.

The projects alignment to the technological guidelines: The use of strategic and operational technological committees acts as a sustainability and intelligence network technological base, constituted by experts, that has been presenting as a success formula in the technological projects management.

Database structured considering a holistic vision. Since the first phase, related to the data collection and the technological demands identification, starting from the CTE – Strategic Technological Committee, its important define an integrated vision of all the management project process. This is to guarantee the approach defined for the selection technological project process, since the initial economic stage evaluation to a final stage to the project execution and implementation.

As final recommendations, we can suggest:

The success in the extraction of the knowledge with the collected data: The data mining process obtained success when extracting a data group not initially prepared for such, resulting in a group of rules that can be used as help and support to technology project prioritization.

The need to structure a database to intend prospect knowledge: It was verified that greater gains are reached when the database construction considering an integrated and future view of data for these technological projects prioritization using data mining approach; resulting an enormous economy, as occurs the elimination of the collection of non-necessary data to the process without an effective return of the practiced investment.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Rules validation: The process of great importance is to consolidate and validate the rules, guaranteeing the rule robustness obtained starting from the data mining mass. Thus, the 20% data separating random process for posterior validation is of great importance for the end result process.

Validation with other database: After consisting and validating the rules for a specific data group, and also validating in other databases, it is necessary to include the expert opinion, that contains the tacit knowledge, and will improve the obtained rules.

In this study, the obtained rules were validated with a database related to the customers evaluation, (physical and costs project data), being necessary to complement the rules initially obtained. As example: It was obtained the rule that projects with <138 HHNS and with an implementation time <1 year would be concluded projects and not implemented, interrupted or canceled. The fact is that when validating this rule in the physical and costs database, projects that were in process appeared, didn't agree with the initially rule obtained. Analyzing with larger depth, it was verified that the projects in subject - in process – were used external resources to the Company, therefore, it was necessary to include in the rule that, the same, was only valid for projects accomplished with own resources.

Considering this study, seeking the improvement of the techniques used for the extraction of the knowledge of database, the following techniques must be still examined:

- **Structuring of an integrated system of projects**: the information architecture related to the management technological projects process similar to the CRM (Customer Relationship Management Process) architecture.

- **Inference tool or project prioritization verification:** the use of a tool that already allows the potential projects entrance and the exit of projects priorities, based on the obtained rules of the KDD process, where each rule can be dynamically updated. As well as only part of the explicit knowledge will be available, it can be increased periodically.

- **Automated selection tool of rules:** It is observed that the rules obtained using KDD process is delayed and complex. However, its automation can be obtained by genetic algorithms, where the rules formation approaches obtain an adequate function definition.

References

[1] Adriaans, Pieter e Zantinge, Dolf, Data Mining, Addison - Wesley Press, England, 1998.

[2] Fayyad, Usama e Shapiro, Gregory e Smyth, Padhraic e Uthurusamy, & Ramasamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Massachusetts, 1996.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

[3] Gonzaga, Solange M. e Menezes, Ana Cristina, A Methodology for Project Prioritization in CENPES, PETROBRAS/CENPES, 1997.
[4] Han, Jiawei e Kamber, Micheline, Data Mining - Concepts and Techniques, Academic Press, San Diego CA, 2000.
[5] Lee, Mushin e OM, Kiyong, Different Factors Considered in Project Selection at Public and Private R&D Institutes, Technovation, 1996.
[6] Nagai, Walter Aoiama, Evaluating the Knowledge Discovered in Regression Problems, M.Sc thesis, São Paulo - USP - São Carlos, 2000.
[7] Peat, David, Artificial Intelligence: How Machines Think, Baen Enterprises, New York EUA, 1985.
[8] Pyle, Dorian, Data Preparation for Data Mining, Morgan Kaufmamm Publishers, San Francisco, 1999.
[9] Rodriguez, Martius, Enterprise Knowledge Management, QualityMark, Rio de Janeiro, 2003.