Data Mining as Support for Customer Relationship Management: A Case Study

José Gonçalo dos Santos

Instituto Federal de Brasília - IFB

QNM 40, Área Especial 01, às margens da BR 070.

CEP: 72146-000

Taguatinga/DF, Brazil (BR)

ABSTRACT

This article describes the data mining application to CRM - Customer Relationship Management. The article starts with an introduction showing the importance of the CRM strategy for the company, after it's introduced the theoretical about CRM, Knowledge Discovery Database and its stages, with emphasis to the mining stage and concludes with presentation of a case study and the conclusions. For the case study it was developed a prototype of an information system of a bookstore, it was implemented, beyond the conventional functions, the association rules discovery algorithm. The implementation of the data mining technique allowed to the system supply support so that the user knows better the client, becoming possible the application of the strategy of CRM.

General Terms

Knowledge Discovery Database

Indexing terms

Knowledge Discovery Database; Data Mining; CRM; Association Rules

Academic Discipline and Sub-Disciplines

The subject addressed by this article is the science of computing whose area is knowledge system.

SUBJECT CLASSIFICATION

Knowledge system

COVERAGE

This paper refers to the use of Data Mining as support for Customer Relationship Management.

TYPE (METHOD/APPROACH)

The methodology used to develop this study was to Literary Analysis and verification practice through a case study.

Council for Innovative Research

1. INTRODUCTION

In general, companies always developed their sales campaigns focusing on the product, using techniques known as mass marketing. In this scheme, the form of customer communication is unilateral. The company uses information gathered by marketing area through surveys, usually made by market segmentation. Based on these data campaigns are developed for a particular type of public. It is hoped that all people of a particular segment have the same needs and behavior, always discounting a margin of error. But that's not what happens because each individual is unique and therefore their tastes are differentiated. Therefore, it is necessary to adopt another type of strategy, shifting the focus from product to customer.

Shift in focus from product to customer requires a new attitude of companies. But that does not mean you can leave aside surveys to measure market share and so little set actions of communication through the mass media. The issue is that it just is not enough. It is also necessary to adopt initiatives to retaining traditional customers and attracting new, being closely monitored by management and enterprise-wide. To do this we must identify those customers and specifically ones most important to the company because generate profit, and also ones least important, because your service generates high cost and results in little profit for the company. In fact, what companies need is a way to create relationship with customers and manage this relationship.

To anticipate situations that may improve the relationship with customers it is necessary to know them. And the best way to know them is to discover their habits and liking. This is possible with the use of KDD (Knowledge Discovery in Database), which is implemented through stages with highlights to data mining which is where the knowledge is actually discovered.

KDD is an interdisciplinary research field that merges concepts of statistics, artificial intelligence and database [15]. Their study is motivated by the growing complexity and amount of data from all spheres of human domain and the need to extract useful information from data collected [24].

The discovery of association rules is an area of KDD which aims to find sets of frequent items in a transaction database and infer rules able to show how a set of items influences the presence of other sets of items [24]. It is noteworthy that the association may be between products or between products and customers.

The use of the techniques of discovering association rules in the KDD process was first introduced by [2]. Since then, many researchers as [1], [3], [7], [14], [17], [19], [20] e [22], have used this technique successfully in discovering customer buying habits. All models used by these researchers were based on the measure of support and confidence.

Thus, this study aimed to use the support and confidence to assist in managing customer relationship.

2. CRM – Customer Relationship Management

CRM is a strategy of business management through the relationship with the client, to achieve greater profitability and gain competitive advantage, pointing to the participation of this technology as a way to automate various business processes such as sales, marketing, services consumer and support field. CRM integrates people, processes and technology to optimize management of all relationships, including customers, business partners and distribution channels.

CRM is a business strategy focused on serving and anticipating the needs of current and potential customers of a company. From a technological standpoint, CRM involves capturing customer data throughout the enterprise, consolidate all internal and externally captured data into a central database, analyze the consolidated data, distribute the results of this analysis to the various points of contact with the client and use this information to interact with the customer through any point of contact with the company [9].

According to [9] CRM consists in:

➢ Helping the company, allowing for its departments identify and target their best customers, manage marketing campaigns with clear goals and objectives, and generate quality reports for the sales team, as well as improving transformation strategies of smaller customers in best customers in addition to recover lost customers and increase profitability by customer.
➢ Helping the organization to improve account management and sales management by optimizing information shared by multiple areas, and streamlining existing processes and reducing costs. Allowing the formation of individualized relationships with customers, with the goal of improving customer satisfaction and maximizing profits; identifying the most profitable customers and providing the most appropriate level of service.

The CRM strategy began to be studied and applied since 1999, since then she has been disclosed by many professionals in the field of marketing and technology, each giving their view of what is CRM. For more details on this topic, see [10], [12], [23] e [25].

3. Knowledge Discovery in Database (KDD)

When starting the topic of Knowledge Discovery in Databases (KDD), it arise the following question: KDD and Data Mining have the same meaning? In reality, there is no consensus about it because, for some authors, the two terms have the same meaning. According to [13], these issues are opened to debate and the definition of each can vary, depending on the author chosen to reading.

To [4], the search for useful patterns in databases has received various names such as Knowledge Discovery in Databases, data mining, information discovery, archeology of data or data normalization process. While the data mining term is used by statisticians and data analysts, researchers in Artificial Intelligence (AI) use the term KDD.

KDD (Knowledge Discovery in Databases), according to [4], is the discovery of new knowledge, which can be patterns, trends, associations, odds or facts that are not obvious or easy to identify. Already [13] defines data mining as the search for trends and patterns in the database. The definitions of [13] and [4] suggest that both terms have the same meaning. But to [16], the term KDD refers to the entire process of discovering knowledge while data mining is seen as a key step in this process, where applies algorithms to extract and verify hypotheses.

3.1 Steps of the KDD

The KDD process starts with defining the scope of the problem in question. Some authors define this step as belonging to the phase of data mining. Already [13] states that the KDD process begins with the preparation of data. To [4], the definition of the objective is the definition of the knowledge that the user wants to obtain about the data. It is at this stage that which is defined the type of pattern that you want to find in the database. In this research, the KDD process starts with the definition of the objectives, as illustrated in Figure 1.

The second step of the process is to acquire data that can be made with the aid of a data warehouse (DW). The use of Data Warehouse is advocated by some authors [15] and denied by others [13].

The figure 1 shows the remaining steps of the KDD, starting with defining the goals and following through data selection, pre-processing, processing, mining and interpretation. Each step described below have an important role in the KDD process.

3.1.1 Definition of objectives

This stage is where the goals are traced, because, for a job of discovery of knowledge have successful, is need to be clear in what is being sought. Normally this phase is done with the help of an expert in the field of application.
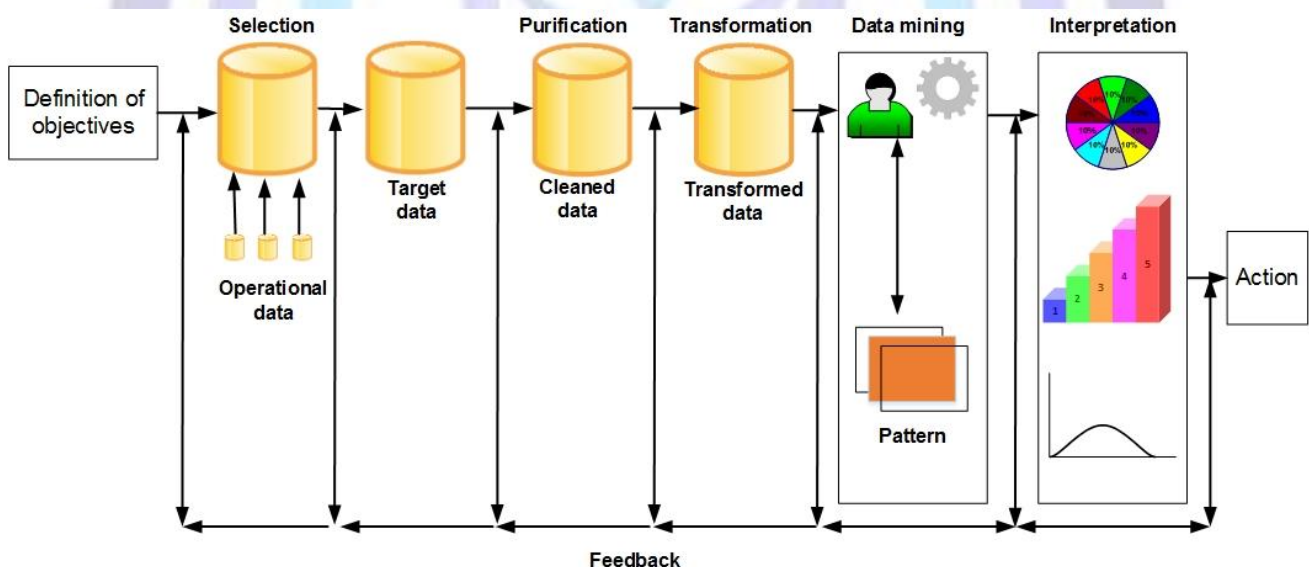

Fig 1: Steps of KDD, adapted from [11].

3.1.2 Selection

At this stage, it selects a set of data or focuses on a subset of attributes or data instances, in order to create a set of target data, in which the discovery is affected. To accomplish this step, it is necessary to have an understanding of the domain and of the goals of the task, according [4] e [21].

3.1.3 Purification

According to [15], at this stage the data cleaning is performed, which involves:

- The treatment of absent data - which can be done by eliminating the tuple, or using often mean of existing values to fill the fields;
- Reduction or elimination of noise - which can be done by binning (replace noisy values by drawing lots of values belonging to the neighborhood), grouping or human inspection with the aid of computational tools;
- Correction of data inconsistencies - which deals with the correction or elimination of inconsistent data.

3.1.4 Transformation

According to [15], in this step the data are processed so that they become suitable for the task mining, which it will be subjected. May involve, among others:

- Aggregation- there is often no need to represent all ranges of values of a given variable. It can group them into broader bands, thereby reducing the number of ranges of values and the complexity of the problem;
- creating attributes – the attributes are created and added to the dataset to assist in the mining process;
- Generalization data – the initial values (low level) of attribute are exchanged by concept high-level hierarchical. For example, the values of age attribute may be replaced by young, adult or old.

3.1.5 Data Mining

It is this stage that is made the discovery of knowledge or standards, per se. At this time, the techniques are chosen according to the kind of problem to be solved. Due to the fact it is a step of fundamental importance, it will be treated in more detail in 3.2 Section.

3.1.6 Interpretation

At this step is done the interpretation of discovered knowledge, and the return to the previous steps, if necessary. Redundant or irrelevant patterns are removed and the useful patterns are translated into terms understandable to users. Furthermore, one should incorporate the knowledge gained to improve system performance by taking actions based on knowledge or simply documenting and reporting this knowledge to stakeholders.

3.2 Data Mining

Data Mining, according to [16], is considered the central phase of the KDD process. This phase is solely responsible for the mining algorithm, i.e., is responsible for implementing the algorithm that seeks to extract the implicit knowledge and potentially useful data.

Phases of Data Mining:

a) choice of algorithms - At this stage, the algorithms are chosen according to the objectives defined in the initial phase of KDD, taking into consideration the type of data you have for the next phase is successfully completed;

b) Discovery of new relations - At this stage is that new relationships that are not easily identifiable are discovered, but that can be visualized with the help of some techniques, through a systematic and comprehensive analysis of a large database;

c) Analysis of the relationships discovered - At this stage, the findings are analyzed by a domain expert to check for an informational value and are consistent. One should also check whether objectives were fully achieved, otherwise you should return to the previous stage;

d) Uses of discovered relations - At this stage, decisions are made in order to use the best possible way, the discoveries relations. The use of these relationships must be made in a rational way in order to obtain the best possible result;

e) Evaluation of results - This is the final phase, in this moment we can to check if the problem was resolved or whether objectives were achieved. Therefore, to start a KDD job you should be aware of what problem is trying to solve for the results obtained can be validated.

There are several techniques of data mining, including: classification [6] e [13]; grouping [13], [6] e [15]; time series analysis [5] e [18] and discovery of association rules, shown below.

3.2.1 Association Rules

Association rules are simple classes of sentences that can be discovered in large data sets whose values are zeros and ones (zero for no particular event and one for presence). Its usefulness lies in the ability of the algorithm to find all rules that satisfy certain conditions set by the user [16].

To [13], association refers to useful business information that can be extracted from aggregate associations between the different items sold in catalogs or in store (physical or virtual). The entries for the association analysis are transactional data coming of sales points, and the outputs are information and recommendations on associations between products and buying behavior of customers.

According to [8], the discovery of association rules is a task of data mining that aims to find relationships or frequent patterns between data sets. Simply put, association rule is a relationship of the type "If (x) then (y)" where x and y are sets of items with empty intersection.

Basically, two factors are assigned to each rule: support - number of records found containing 'X' and / or 'Y' on the total number of records, according to (see eq. 1); confidence - number of records found containing 'X' and 'Y' on the number of records containing 'Y', according to (see eq. 2). Finally, consists in comparing two or more products, relating them in such a way that, based on previous cases, calculates the probability of these products be purchased together and the affinity between them.

$$\sup(A \to B) = prob(A^{\wedge}B) = \frac{n(A^{\wedge}B)}{N} \qquad \text{(eq. 1)}$$

being: $\sup(A \to B)$= support "if A, then B";

$prob(A^{\wedge}B)$ = probability of A and B occur;

$n(A^{\wedge}B)$ = number of occurrences of A and B;

N = number of cases in the database.

$$conf(A \to B) = \frac{freq\_rel(A^{\wedge}B)}{freq\_rel(A)} \qquad \text{(eq. 2)}$$

being: $conf(A \to B)$= confidence "if A then B",

f$req\_rel(A^{\wedge}B)$ = relative frequency of A and B,

f$req\_rel(A)$ = relative frequency of A.

4. Case Study

As a case study was developed for a system to support the decisions of the sales area as a whole from the array of products on the sales floor to support the sale to the final consumer indicating products most likely to be acquired by him. The system works by comparing patterns and by checking on previous purchases which products are likely to be purchased together, given a profile of the potential customer. Thus, the system can support CRM strategy implemented in the company. The details of this process are presented following.

The system maintains a dynamic table and each time a new product is registered a new column is generated and each time a sale is done is generated a new line. This table is referred as co-occurrence matrix containing zero (0) and one (1), where 0 represents no occurrence and 1 represents each occurrence of sale of the product. The table 1 shows the fragment of co-occurrence matrix contained in the system.

The user interaction occurs as follows: two columns are presented, both representing the store's products in question, where the user selects in the first column one or more products and in the second column, one or more products which he wants to test the affinity with the first selected, see Figure 2.
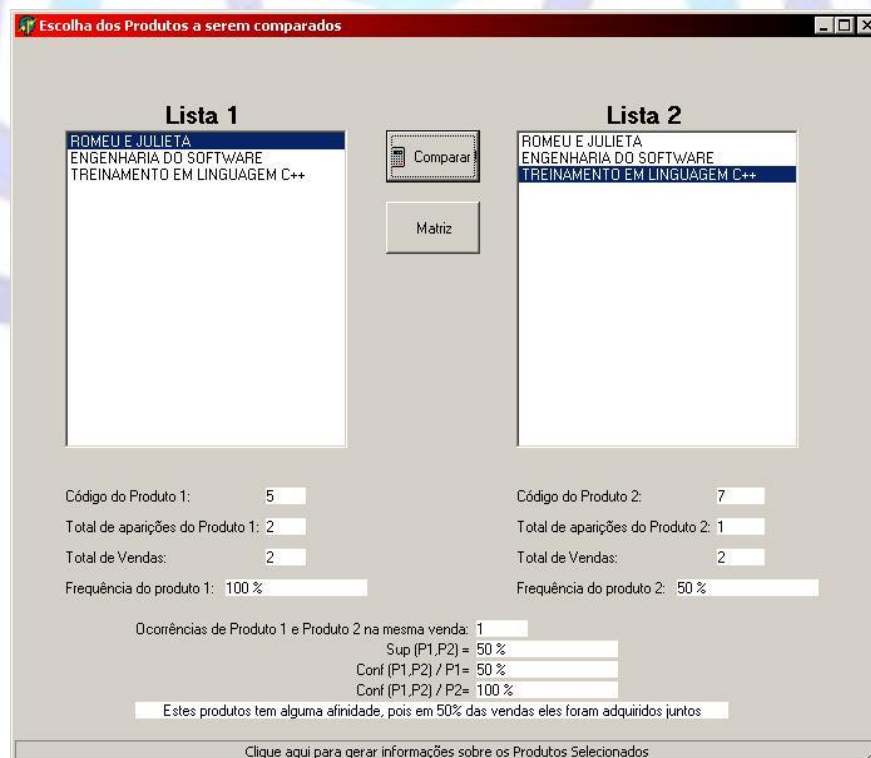

Fig 2: Interaction between the user and the system.

4.1 Support

Suppose the user has selected police genre and genre adventure and wants to know the affinity between these two categories. The rule would be triggered as follows: "if customer buys police genre book then buys adventure genre book ".

Support for this rule would be 20%, i.e. at least 20% of transactions in the database contain the cops books and adventure, together. This analysis can also be done as follows: the percentage of cases in which the occurrence of cop book "foresees" correctly the occurrence of adventure books.

This information is important to organize the arrangement of books on the store shelves, it allows to determine which genera should be placed close to each other, usually those who have greater support (most likely to be purchased together). Also, shows which the best options for possible promotional packages to be offered by the bookstore in order to improve the sales of a particular genre. That is possible also define which products should be prioritized in the commercial area or in marketing area, by defining different strategies for selling and bringing consequent improvement in profits.

Based in this choice, the process of searching for association rules is fired, generating, thus, interesting information with regard to the sale of products. Assuming Table 1 as the matrix of co-occurrence complete for system, the support and confidence values are presented and discussed following.

Table 1: Example of co-occurrence matrix of the system

CUSTOMER	PRODUTO						
	Romance	Fiction	Self Help	Police	Drama	Suspense	Adventure
1	0	1	0	0	0	0	1
2	1	0	1	0	0	0	0
3	1	0	0	0	0	0	0
4	0	0	1	0	1	0	0
5	0	0	0	1	0	1	1
6	0	1	0	1	0	0	1
7	0	0	0	0	1	0	0
8	1	0	1	0	0	0	0
9	0	1	0	1	1	0	0
10	0	0	0	0	0	1	0
11	0	0	0	0	1	1	0
12	1	0	1	0	0	0	0
13	0	0	0	1	0	0	0
14	0	1	0	1	1	1	1
15	1	0	0	0	1	0	0
16	0	0	0	1	0	0	0
17	1	0	0	0	0	1	0
18	0	0	0	0	1	1	0
19	0	0	0	0	0	0	1
20	0	0	0	1	0	0	1

4.2 Confidence

The information obtained in the previous example does not give a real idea of the relationship between the selected products. To eliminate this deficiency more information is needed. The most common information used in conjunction with support for discovery of association rules is confidence, which is the minimum percentage of transactions containing together books the police genre and adventure genre, for example. Confidence can also be interpreted as the percentage of transactions in which the co-occurrence is observed. In our example, the confidence is 57%, i.e., indicates that in 57% of the time it was bought a police book, the customer took a book of adventure. The confidence plus the support gives better information about existing relationship between the two genres. With the information from the previous example, one can reach the following conclusion: "if the customer purchases a police book, then he buys a book of adventure in 57% of cases and this rule applies to 20% of cases".

5.Conclusions

It is possible create algorithms that inform, based on previous cases, what a client can potentially buy, based on the product you have just purchased. It is possible also perform search the database with an algorithm that compares and generates association rules with the support and confidence maximum possible, demonstrating all the major relationships between products.

The use of data mining as a means of knowledge discovery brings as a benefits solutions that can be implemented computationally so that they can assist the user in the decision-making process.

Association rules is a technique for data mining that has great advantage when used to support the management of relationships with customers, because besides being easy to implement, provides important information about the habits of customers. Thus, causing the company can offer only those products that are of interest to certain customers, or can foresee a need or desire of its customers.

REFERENCES

[1] AGGARWAL, C. and YU, P. Online generation of association rules. ICDE-98, 1998, pp. 402-411.

[2] AGRAWAL, R., IMIELINSKI, T., SWAMI, A. Mining association rules between sets of items in large databases. SIGMOD-1993, 1993, pp. 207-216.

[3] AGRAWAL, R. and SRIKANT, R. Fast algorithms for mining association rules. VLDB-94, 1994.

[4] AMARAL, Fernanda C. N. Mineração de dados: Técnicas e Aplicações para o Marketing Direto. São Paulo: Berkeley, 2001.

[5] BERK, Kenneth N. Data Analysis with Systat. USA: SYSTAT, 1994.

[6] BERRY, Michael, J. A. & LINOFF, Gordon. Data Mining Techniques: for Marketing, Sales, and Customers. USA: John Wiley & Sons, 1997.

[7] BRIN, S. MOTWANI, R. ULLMAN, J. and TSUR, S. Dynamic Itemset counting and implication rules for market basket data. SIGMOD-97, 1997, pp. 255-264.

[8] BRUSSO, Marcos José – Um modelo para a mineração de regras de associação aplicado ao uso da Web. In: SEMANA ACADÊMICA DO PPGC. 4.: 1999, AGO.16-20. Porto Alegre. Anais... Porto Alegre: PPGC da UFRGS, 1999.

[9] BUCCI, Antonio G. CRM: Conceitos e Aplicabilidade. Available in: <http://www.bucci.com.br/artigos/CRMart080801.pdf >. Acesso: 16/03/2005.

[10] CARDOSO, Mário Sérgio; FILHO, Cid Gonçalves. CRM em ambiente e-business. São Paulo: Atlas, 2001.

[11] FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory e SMYTH Padhraic. The KDD Process for Extracting Useful Knowledge. Communications of the ACM Digital Library, Nov., 1996, Vol. 39, p 27-34.

[12] GREENBERG, Paul. CRM - Na Velocidade da Luz. Rio de Janeiro: Campus, 2001.

[13] GROTH, Robert. Data Mining: Building Competitive Advantage. New Jersey – USA: Prentice Hall, 2000.

[14] HAN, J. and FU, Y. Discovery of multiple-level association rules from large databases. VLDB-95.

[15] HAN, Jiawei& KAMBER, Micheline. Data Mining: Concepts and Techniques. USA: Morgan Kaufmann, 2001.

[16] KLÖSGEN, Willi& ZYTKOW, Jan M. Handbook of Data Mining and Knowledge Discovery. USA: Oxford University Press, 2002.

[17] LAKSHMANAN, L.; NG. R. T.; HAN, J. Exploratory mining and pruning optimizations of constrained association rules. SIGMOD-98, 1998.

[18] NOTARI, Daniel L. Aplicação de Redes Neurais Artificiais à Mineração de Dados. Available in :<http://www.inf.ufrgs.br/~dlnotari/trabalhos/ucs/arnmd/index.html>. Accessin 15/02/2000.

[19] PARK, J. S. CHEN, M. S. and YU, P. S. An effective hash based algorithm for mining association rules. SIGMOD-95, 1995, pp. 175-186.

[20] RASTOGI, R. and SHIM, K. Mining optimized association rules with categorical and numeric attributes. ICDE –98.

[21] ROBIN, Jacques; BEZERRA, Ricardo. Descoberta de Conhecimento em BD. Available in: <www.di.ufpe.br/~compint/aulas-IAS/KDD-991/KDD.ppt >. Access in: 13/08/2003.

[22] SRIKANT, R., VU, Q. and AGRAWAL, R. Mining association rules with item constraints. KDD-97, 1997, pp. 67-73.

[23] SWIFT, Ronald. CRM – Customer Relationship Management. O Revolucionário Marketing de Relacionamento com o Cliente. Rio de Janeiro: Editora Campus, 2001.

[24] VELOSO, A. et. al. Mineração Incremental de Regras de Associação. In: XVI Simpósio Brasileiro de Banco de Dados. IME – Rio de Janeiro, p. 80-94, 2001.

[25] ZENONE, Luiz; LOPES, Adalberto P.; BRAGA, Ana Cláudia L.S.; POSER, Denise Von; FEYO, Miguel; MANZIONE Jr., Sydney. CustomerRelationship Management (CRM) Conceitos e Estratégias. São Paulo: Atlas, 2001.

Author's Biography with Photo

José Gonçalo dos Santos was born in Cuiabá-MT, Brazil. He has graduated in Mathematics from the Federal University of MatoGrosso (1994) Technology and IT Management from the Catholic University of Brasilia, postgraduate in computational mathematics from the Federal University of MatoGrosso (1996) and requirements engineering and business process Federal University of Rio Grande do Sul (2012), Masters in Computer Science from the Federal University of Santa Catarina (2001) and Ph.D. in Computer Science from the Federal University of Santa Catarina (2004). Has experience in the area of Computer Science, with emphasis in Computer Systems Architecture, mainly in the following themes: expert systems, artificial intelligence, statistical analysis, data mining, database, ITIL, Object Oriented Analysis, Oriented Programming objects, programming languages (Java, C / C + +, Delphi, PHP), and SQL Development JAVA WEB.