

# SINGULAR<sup>®</sup>

ENGENHARIA, TECNOLOGIA E GESTÃO

ANO 1, V.1, Nº 1 - MARÇO DE 2019

 **CEULP/ULBRA**  
CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS

# **Singular Engenharia, Tecnologia e Gestão**

**Vol. 1, N. 1, Fevereiro, 2019**

**eISSN: 2596-2604**

**<https://doi.org/10.33911/singular-etg.v1i1>**

**EXPEDIENTE**

**Centro Universitário Luterano de Palmas**

**Reitor**

Adriano Chiarani da Silva

**Direção Acadêmica**

Parcilene Fernandes Brito

**Singular Engenharia, Tecnologia e Gestão**

**Editora Chefe**

Dra. Parcilene Fernandes Brito

**Editora Assistente**

Me. Heloise Acco Tives Leão

**Leitura de Prova**

Me. Fabiano Fagundes

**Normalização**

Dr. Pierre Soares Brandão

**Comunicação**

Dra. Irenides Teixeira

Me. Luiz Gustavo Santana

Me. Sonielson Luciano de Souza

**Comissão Editorial**

Me. Fabiano Fagundes

Me. Jackson Gomes de Souza

Me. Madianita Bogo Marioti

**Conselho Editorial**

Dr. Alberto Cesar Cavalcanti Franca, UFRP, Brasil

Dr. Daniel Costa de Paiva, UFF, Brasil

Dr. Edeilson Milhomem, UFT, Brasil

Dra. Edna Dias Canedo, UNB, Brasil

Dr. Flavius da Luz e Gorgônio, UFRN, Brasil

Dr. Leandro Maciel Almeida, UFPE, Brasil

Dr. Leandro Pasa, UTFPR, Brasil

Dra. de Oliveira Rech, UFSC, Brasil

Dra. Luciano de Souza Cabral, UFPE, Brasil

**Projeto Gráfico**

Adriano Marinho Ribeiro

**Diagramação**

Fernanda Gomes

Me. Heloise Acco Tives Leão

**Imagem da Capa**

Leonardo Moreira

**Apoio Técnico**

Gabriel Borges

Murillo Roseno Feitoza Lima

**Avaliadores deste número**

Bruno S. C. M. Vilar - Instituto de Pesquisas Eldorado. Analista de Software com experiência em Ciência de Dados, Pesquisa e Desenvolvimento de Software. Tem experiência de pesquisa nos temas: Web, Web Semântica, Ontologias, Adaptação Dirigida por Contexto e Mineração de Dados. Atualmente trabalha no Instituto de Pesquisas Eldorado, atuando em projeto utilizando técnicas de Mineração de Dados e Aprendizagem de Máquina.

Edna Dias Canedo - Universidade de Brasília. Doutora em Engenharia Elétrica pela Universidade de Brasília (UNB). Desde 2000 dedica-se a atividades de docência universitária nos cursos de graduação e pós-graduação.

Leandro Pasa - Universidade Tecnológica Federal do Paraná. Doutor em Engenharia Elétrica e da Computação pela Universidade Federal do Rio Grande do Norte (2016) na área de Processamento Inteligente da Informação. Atualmente é professor da UTFPR.

Luciana Rech - Universidade Federal de Santa Catarina. Atualmente é pesquisadora e professora adjunto do Departamento de Informática e Estatística (INE) na UFSC. Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Computação.

Luciano de Souza Cabral - Instituto Federal de Pernambuco. Atualmente é Professor Adjunto do Instituto Federal de Pernambuco e Sócio-fundador da startup DAX Tecnologia. Tem experiência na área de Ciência da Computação, com ênfase em Inteligência Artificial, Engenharia de Software, Redes, Segurança e Tecnologias emergentes (RF-ID, IoT, etc)

Editada em fevereiro de 2019.

Última edição em abril de 2019.

Publicada em fevereiro de 2019.

Nota da Editora: Os autores são responsáveis pela apresentação dos fatos contidos e opiniões expressas nesta obra.

**Dados Internacionais de Catalogação na Publicação na (CIP)**

R454 Singular Engenharia, Tecnologia e Gestão [recurso eletrônico] / Centro Universitário Luterano de Palmas. – Volume 1, n. 1 (fev./2019). Dados eletrônicos. – Palmas: Ceulp, 2019-

Semestral.

Modo de Acesso: World Wide Web:

<<http://ulbra-to.br/singular/index.php/SingularETG/>>

Título varia: Revista Singular ETG.

Descrição baseada em: Volume 1, n. 1 (fev.2019).

eISSN: 2596-2604

<https://doi.org/10.33911/singular-etg.v1i1>

I. Interdisciplinar. II. Computação. III. Engenharia. IV. Gestão. V. Título: Singular Engenharia, Tecnologia e Gestão.

CDU: 001

Ficha catalográfica elaborada pela bibliotecária Thais Fernandes, CRB-2/1680

**SINGULAR ENGENHARIA, TECNOLOGIA E GESTÃO**  
**CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS**  
Avenida Teotônio Segurado 1501 Sul  
Palmas - TO CEP 77.019-900 Caixa Postal nº 85  
Fone: (63) 3219 8125 email: [revista.singular@ceulp.edu.br](mailto:revista.singular@ceulp.edu.br)

## EDITORIAL

A Singular Engenharia, Tecnologia e Gestão apresenta à comunidade científica sua primeira edição. Fruto de um trabalho incansável de sua editoria, esta revista busca oferecer um espaço a mais para apresentação de artigos que contribuam significativamente para estudos nestes campos.

A Singular Engenharia, Tecnologia e Gestão é uma publicação semestral, avaliada por pares que se propõem a contribuir na difusão de debates e ideias no espaço acadêmico e, com isso, criar uma rede de pesquisadores com perspectivas teóricas e metodológicas diversas, propiciando a troca de informações e o debate sobre as principais questões nesses campos.

A partir desta primeira edição nos colocamos à disposição da comunidade científica no objetivo de buscar atender aos seus anseios e necessidades no que diz respeito à um espaço de publicação e apresentação de resultados das pesquisas realizadas. Estamos abertos à todo tipo de contribuição neste sentido.

Agradecemos aos pesquisadores que submeteram seus trabalhos para esta edição bem como ao conjunto de avaliadores que se dispôs a contribuir com a análise dos artigos, oferecendo suas considerações para que atingíssemos o resultado esperado.

Boa leitura.

Palmas-TO, fevereiro de 2019.

## SUMÁRIO

### **1 - Market Basket Analysis in a Financial Institution.**

Felipe de Moura Rezende dos Santos and Marcelo Ladeira. Páginas (6 - 12)

### **2 - Nomeando a dor da engenharia de requisitos no domínio de sistemas robóticos.**

Tayse Virgulino Ribeiro, Laura Maria Palomino Marino e Fabio Queda Bueno da Silva. Páginas (13 - 20)

### **3 - It change management process improvement: the case study of a brazilian court.**

Eduardo da Silva Sousa, Carlos Eduardo Machado Pires, Simone Borges Simao Monteir and Ana Carla

Bittencourt Reis. Páginas (21 - 26)

### **4 - Desenvolvimento do Módulo de Pre-processamento da ferramenta SentimentALL**

Parcilene Fernandes Brito e Luan Gomes de Almeida Araújo. Páginas (27 - 35)

### **5 - Apoio à decisão na prevenção de acidentes de trânsito: uma aplicação dos métodos AHP e PROMETHEE**

João Vicente Pereira, Rômulo de Medeiros Palmeira e Ana Carla B. Reis. Páginas (36 - 47)

### **6 - Modelando um esconde-esconde no *Minecraft* utilizando *Hidden Markov Model***

Ronaldo e Silva Vieira, Matheus Rodrigues Leal e Luiz Chaimowicz. Páginas (48 - 52)

### **7 - Classifying Descriptions of Goods with Artificial Neural Networks.**

Vinicius Di Oliveira and Marcelo Ladeira. Páginas (53 - 57)

# Market Basket Analysis in a Financial Institution

**FELIPE DE MOURA REZENDE DOS SANTOS,  
MARCELO LADEIRA.**

Computer Science Department, University of Brasília-(UnB), P.O. Box 4466, 70910-900, Brasília-DF, Brazil.  
(e-mail: feliperezende86@gmail.com, mladeira@unb.br)

Autor Correspondente: Felipe Rezende (e-mail: feliperezende86@gmail.com).

• **ABSTRACT** - This article demonstrates a study on Market Basket Analysis of a financial institution, showing rules of personal consumer association of the state of São Paulo. A concept about three association algorithms is presented, but a study with only one is performed. The paper is divided into an introduction, describing a brief account of the reason for choosing the subject. Understanding the business, where it is explained about the financial institution and the importance of the study to the institution. The way the data are handled is demonstrated in Understanding the Data, just as the Data Preparation is described in the sequence, putting all the filters and treatments that were done on the data. In the following, it is described the Modeling, which reports on algorithms of association rules and on examples of these algorithms, as well as which algorithm was chosen to be treated in the paper. Evaluation explains on the results obtained with the study and the Implementation as it was done all the analysis of the data and the results obtained. Finally, we have the Conclusion about the learning obtained with the article and what future work to do.

• **KEYWORDS** - Data mining, Data, Mining, Association, Apriori

## I. INTRODUCTION

What does beer have to do with diapers?

It was a common question until one of the largest US retailers conducted a study and found that diaper sales were associated with beer. In general, buyers were men who went out at night to buy diapers and ended up buying some beer. The products were placed side by side and the result was increased sales of the two products. It is unknown whether this study actually occurred or whether it is only a "legend" about data mining study.

The competitiveness of companies is becoming increasingly dependent on the quality of their decision-making. Therefore, it is no wonder that companies often try to learn from past transactions and decisions in order to improve the quality of decisions made in the present or future. To support this process, large amounts of data are collected and stored during business operations. Later, this data is analyzed for relevant information. This process is called data mining or knowledge discovery in databases. Data mining is relevant to many different types of companies. As examples, retail stores obtain customer profiles and their purchasing patterns and supermarkets analyze their sales and the effect of advertising on sales. This "targeted marketing" is becoming increasingly important.

Thus, this case study tends to propose a specific group of products purchased by customers of a financial institution and predispose a tying or otherwise known as cross-selling, ie, point where the product this customer is likely to purchase, by consumption profile.

## II. UNDERSTANDING THE BUSINESS

Every financial institution has various types of products to be offered to customers, credit products (loans), investment products (investment funds) and services products (insurance). In some cases they are offered by the attendants, products that are not useful for the client profile, generating a certain discomfort and a problem in the accomplishment of this purchase.

It will be used Data Mining concepts to have a predisposition of certain items to be offered in conjunction with products that the customer already has, increasing the probability of making this offer effective. Through association of the products already owned by him. For this, all the products contracted by the customer will be determined, so that a grouping of this information can be made and a profile of customers and products acquired be generated. This process would have the cost of being processed in a large database, which is not prepared for such a procedure, running the risk

of not being able to perform in a timely manner to carry out the analysis. Therefore a study on a sample of data was determined. Only one month, individual customers and products used.

As a criterion of success, one can place the possible creation of customer profile groups with the same interest of consumption of the products of the Financial Institution. With this having the possibility of a better offer of consumption of other products for clients who are doing business with the institution.

### III. UNDERSTANDING DATA

We have a lot of customer data. In order to start this project, it was defined that a study of only 1 month, in the case, July 2017, would be done because it is the most recent month already processed. Nevertheless, it remained with more than 40 million lines, with that restricted to only individuals that contain current account in the institution. For a more restricted analysis of data, the state of Sao Paulo was chosen, with a total~ of 9,343,551 lines.

The database used to conduct this study contains 160 variables and more than 1 billion lines, since it contains data referring to all operations of all clients of the financial institution since January 2011. This database is partitioned by year and is indexed by year and month.

### IV. DATA PREPARATION

The database used in the study contains no problems with the need to be corrected to be a basis of use for disclosure of results of products for the dependencies, so it was only required to make necessary filters and the reduction of variables needed to study.

Initially, a filter was performed to have only the data of July 2017, excluding operations that do not contain customers, bringing only the customer code, the unit of the federation of cadastre and the products contracted by him in the institution. Altogether, more than 236 million lines were generated, with 1 line per customer and product of this customer.

After it was decided that it would be restricted to a specific state for the purpose of studying and executing the algorithms in a home computer, the state of São Paulo was chosen. Included in the preparation of the data were the permanence of only the Individuals customers. Totaling approximately 9 million lines.

A pivoting of the data was performed, generating a column for each product of the bank, having for each customer all the existing products in their relation with the bank.

### V. MODELING

The analysis of purchases and product records typically uses association rules representing the relationship patterns found between the data items of the analyzed set. In databases where records are transactions, these rules are known as transactional association rules, whereas in the case of databases where records represent clients, accounts, products, services

and others, they involve multiple attributes and, therefore, are called multidimensional association rules.

In the context of marketed product sales, or cross-selling the membership rules allow a store to recommend product B for the customer who is buying product A, since it knows the rule, for example, that 30% of your customers who purchase A also buy B. Thus, the customer is encouraged to buy more products, which may possibly interest him, based on consumption characteristics of previous purchases in the system. In this way, not only the quantity of sales is maximized, but also the quantity of sales of a particular product.

Association rules were introduced [1] as follows: are  $I = \{i_1, i_2, \dots\}$ , in a set of  $m$  distinct items and  $D$  a database formed by a set of transactions, where each transaction  $T$  is composed of a set of items such that  $T \subseteq I$ . An association rule is an expression of type  $XY$  where  $X \subseteq I$ ,  $Y \subseteq I$ ,  $X \cap Y = \emptyset$ ,  $Y \neq \emptyset$  and  $X \cap Y = \emptyset$ , that is, both the antecedent ( $X$ ) and the consequent ( $Y$ ) of an association rule can be formed by sets containing one or more items, but can not be empty and are independent sets since they have no items in common. For example:

{Cerveja, Faldas} → {Leite}  
 {Cerveja, Leite} → {Faldas}  
 {Faldas, Leite} → {Cerveja}  
 {Cerveja} → {Faldas, Leite}  
 {Leite} → {Cerveja, Faldas}  
 {Faldas} → {Cerveja, Leite}

Figure 1. Example of Association Rules

The importance of an association rule can be measured in terms of support and confidence. The support of a rule determines how often a rule is applied to a dataset, that is, the probability that the first term of the implication is true. The confidence of a rule determines how often the items in  $Y$  appear in transactions that contain  $X$ . In the previous example, the likelihood of a customer buying beer and diapers (ie, the term  $X$  is true) is referred to as support, while the conditional probability of a customer buying milk, since he bought beer and diapers, is referred to as confidence.

The problem of association mining is to find all association rules that have support and trust greater or equal, respectively, to a minimum support (SupMin) and a minimum trust (ConfMin), specified by the user.

In fact, for a long time, the search for association rules was of exclusive interest to applications that deal with information from Market Baskets, which led to this technique being commonly called market basket analysis (MBA).

However, we can see that association rules can be extracted from any database where there are implicit relationships

between the different attributes. The usefulness of these rules is not only in the generation of new knowledge, but also in the confirmation of business rules that are used but never proven. Thus, association rules apply to several business areas such as studies of access to computers, search for new clients, population census and medical information analysis, among others.

The discovery of standards through analysis of association rules is an important one to support decision making. With this information, a manager can identify new business opportunities, better know their customers and their buying profiles, identify products that influence the sale of others, and various other information that can increase their competitiveness in the market.

The first algorithms to be used in the discovery of association rules were the artificial immune system (AIS) [2] and set-oriented mining (SETM) [3]. In most of the more recent studies the Apriori algorithm has been widely used, as well as its variations. In this work, due to being used in more recent studies and because of the shortage of time for further deepening of other algorithms, Apriori was chosen and used.

**A. ARTIFICIAL IMMUNE SYSTEM**

The technique used by the artificial immune system (AIS) generates and counts sets of items as you read the database transactions.

For each transaction, it is determined which of the largest sets found in the previous transaction are also in the current transaction and new sets of items are generated by extending these sets with items from this record.

The disadvantage is that this method generates and counts unnecessarily sets of items that are considered small. The following figure illustrates how it works.

The first pass of the algorithm in the database makes the search for baskets of only one item counting how many times these are repeated in the transactions, ignoring those that occur only once. The second pass in the base searches for baskets with two items, counting how many times they appear together in the scanned records.

In this example, the last pass contains baskets of three items, which are the largest basket that is often larger than one. Thus, the algorithm has as solution the set 2,3,5.

**B. SET-ORIENTED MINING**

Like AIS, SET-oriented Mining (SETM) generates the candidate item sets at the time it is going through the database transactions. However, counting is only performed at the end of the sweep of the base.

The transaction identifier is stored together with the candidate set in a sequential structure.

At the end of the reading of all records, the most frequent set of items is determined by the aggregation of the sequence structure.

In addition to having the same disadvantage as AIS, for each candidate set, there are several calculated size values.

The following figure illustrates the steps in the algorithm.



Figura 2. Example of Artificial Immune System

In SETM, the first pass of the database results in the frequency of occurrence of each item in the transactions, as in the previous example based on AIS. However, the second sweep in the base results in a list of combinations of product pairs where each combination receives the identifier of the transaction from which it was extracted.

Thus, we will have repeated combinations with different identifiers such as the basket 1,3 that in this example occurs in transaction 100 and in transaction 300.

The last pass at the bottom reveals the baskets of three items and the solution found is the one that appears in most transactions.

With this, we also have the set of items 2,3,5 as a result of applying SETM to this recordset.

**C. APRIORI ALGORITHM**

This algorithm was the first tool to discover association rules in databases with large volumes. Several modifications were

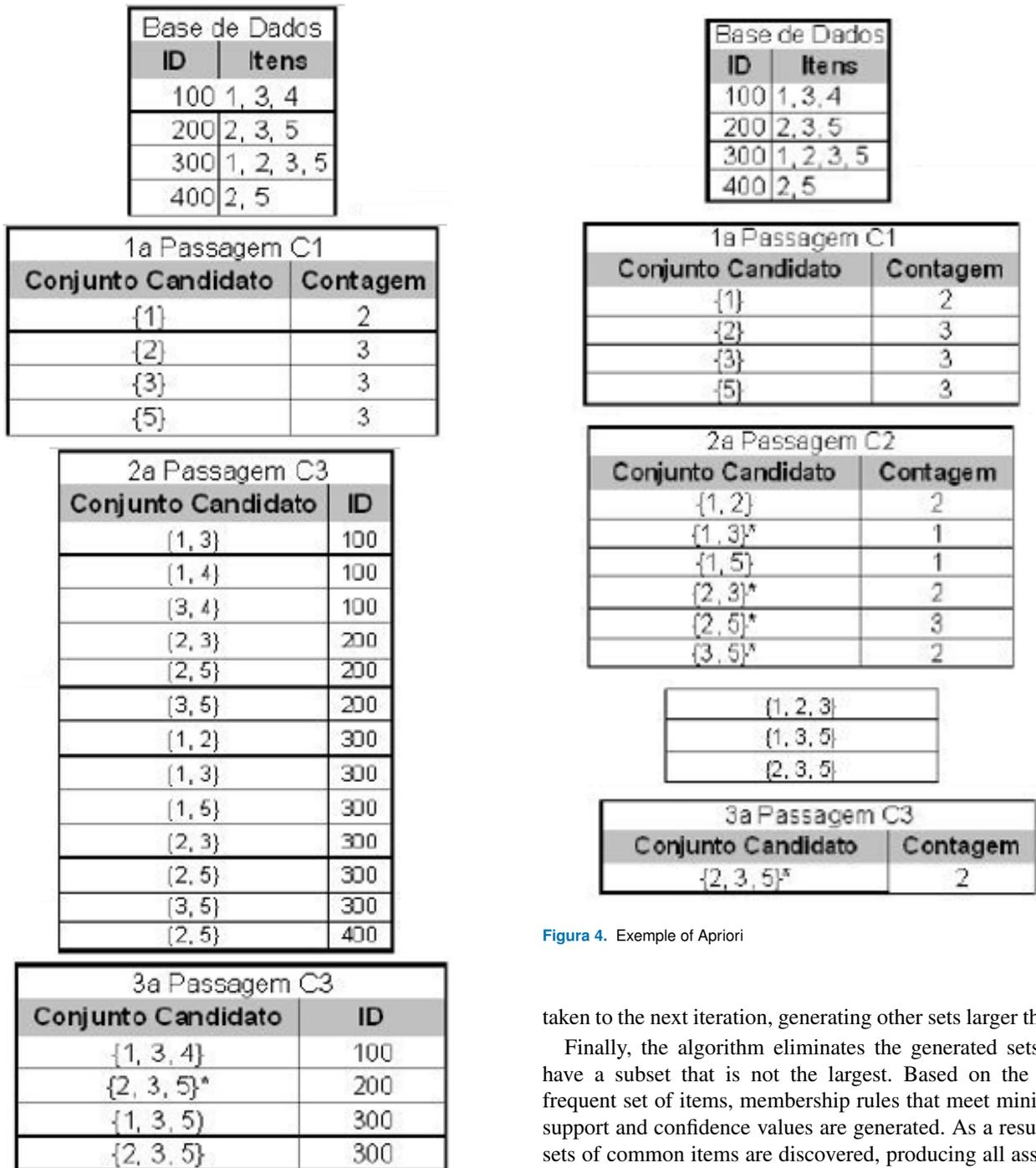


Figura 4. Exemple of Apriori

taken to the next iteration, generating other sets larger than 1.

Finally, the algorithm eliminates the generated sets that have a subset that is not the largest. Based on the most frequent set of items, membership rules that meet minimum support and confidence values are generated. As a result, all sets of common items are discovered, producing all association rules that respect those limits.

As in the previous examples, the first pass of the Apriori algorithm in the database results in the listing of the frequency of the items in the transactions (baskets of an item only). In the next step, the generated list contains the baskets of two items with their respective frequencies.

Of these baskets only the most frequent baskets will be considered and, thus, the baskets with three items are already prepared for the last step, which will only then count the baskets more frequent related in the previous step. With this, the result is the basket of items 2,3,5.

Figura 3. Exemple of SET-oriented Mining

proposed to improve their efficiency and similar algorithms were proposed [4] [5], introducing more expressive rules.

First, the algorithm identifies the items that are part of each transaction. The algorithm then determines the membership rules to enter these items by selecting the associations that occur most frequently (or larger) in the set of transactions in question.

The largest set of candidate items from the previous step is

## VI. EVALUATION

Based on the results obtained, we can successfully evaluate the study because we were able to respond to the initial questioning, determining a shopping basket of products of the financial institution under study.

The process has been reviewed and tested countless times, changing conditions and rules to produce the necessary information demonstrated in the next chapter.

As next steps we can list the study of other algorithms for the association of shopping baskets. As well as the inclusion of other ways to predict possible purchases from certain groups of customers.

## VII. IMPLEMENTATION

All customer information base was obtained through SAS at the financial institution, where the month of July and the year 2017 were filtered, as well as restricting only individuals from the state of São Paulo who have a bank account with the financial institution.

This data is confidential inside the financial institution, because of this, we can not disponibilize best explanations or the real data with this study.

In view of the large volume of data without these filters, a total of more than 35 million lines were still present. After pivoting, considering each customer's product as a specific column, it dropped to 9,343,551 million lines. Where it was divided into females and males with a total of 3,879,539 and 5,464,012 lines respectively. The data was exported in CSV format for use and study of the mined information.

Analyzing women's data, we have the most common items in addition to the account (which was the filter criterion), 100% have a savings account and 82.21% have a credit card in their product basket, 47% service via cellular SMS and 36.6% has service package. We can consider these 4 items as the basic basket of products for women and from there check the items to be offered for new hires.

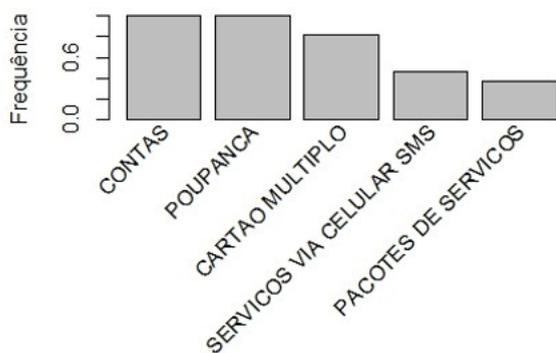


Figura 5. Frequency of women's products

By performing the Apriori algorithm by RStudio, with a support of 5% and 90% confidence we found 580 rules where ordered by lift and presenting only the first 5 rules, would be as follows: In this case, we can verify that, with a high

	lhs	rhs	support	confidence	lift	count
[1]	{CARTAO MULTIPLO, PREVIDENCIA ABERTA}	=> {FUNDOS DE INVESTIMENTOS}	0.053	0.97	10	204116
[2]	{CARTAO MULTIPLO, POUPANCA, PREVIDENCIA ABERTA}	=> {FUNDOS DE INVESTIMENTOS}	0.053	0.97	10	204116
[3]	{CARTAO MULTIPLO, CONTAS, PREVIDENCIA ABERTA}	=> {FUNDOS DE INVESTIMENTOS}	0.053	0.97	10	204116
[4]	{CARTAO MULTIPLO, CONTAS, POUPANCA, PREVIDENCIA ABERTA}	=> {FUNDOS DE INVESTIMENTOS}	0.053	0.97	10	204116
[5]	{PREVIDENCIA ABERTA}	=> {FUNDOS DE INVESTIMENTOS}	0.055	0.96	10	212600

Figura 6. Random Rules for Female Customers

confidence and lift, we have, in all cases, when a person has Open Retirement, they tends to have Investment Funds.

Another form of research would be to assist in the goals set for agencies. Where we can check people who own a particular product, which products already exist in your bank basket. Determining the audience to be offered this product. As an example, we use the Capitalization Title with a single payment, in this case, we use as the sorting criterion the confidence that indicates the probability that a transaction that has an item lhs also contains the item rhs. And we kept showing the first 5 items, in order to find 160 rules.

	lhs	rhs	support	confidence	lift	count
[1]	{CARTAO MULTIPLO, CHEQUE ESPECIAL, EMISSAO DE TITULOS, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[2]	{CARTAO MULTIPLO, CHEQUE ESPECIAL, EMISSAO DE TITULOS, POUPANCA, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[3]	{CARTAO MULTIPLO, CHEQUE ESPECIAL, CONTAS, EMISSAO DE TITULOS, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[4]	{CARTAO MULTIPLO, CHEQUE ESPECIAL, CONTAS, EMISSAO DE TITULOS, POUPANCA, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[5]	{CHEQUE ESPECIAL, EMISSAO DE TITULOS, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4348

Figura 7. Target specific for women clients

When the agency receives a task of executing Capitalization Bonds with a single payment, we can consider, with confidence of 45%, the following situations of customer baskets:

- 1) Multiple Card, Special Check, Issuance of Securities, and Capitalization Bond with monthly payment;
- 2) Multiple Card, Special Check, Issuance of Securities, Savings and Capitalization Bond with monthly payment;
- 3) Multiple Card, Special Check, Issuance of Securities, Accounts and Capitalization Bond with monthly payment;
- 4) Multiple Card, Special Check, Issuance of Securities, Savings, Accounts and Capitalization Bond with

monthly payment;

5) Multiple Card, Issue of Securities and Capitalization Bond with monthly payment;

In all cases, we had a support of 0.11% and a lift of 11. More than 4,300 rules were found in each of the baskets. When we analyzed male clients, we found the most common items were the multiple card with 79.9%, Cellular SMS Services with 44.3%, Savings with 36.2% and Services Package with 32.2. In the case of the algorithm Apriori, with a support of 5% and confidence of 60%, we found 580 rules where ordered by lift and presenting only the first 5 rules, would be as follows:

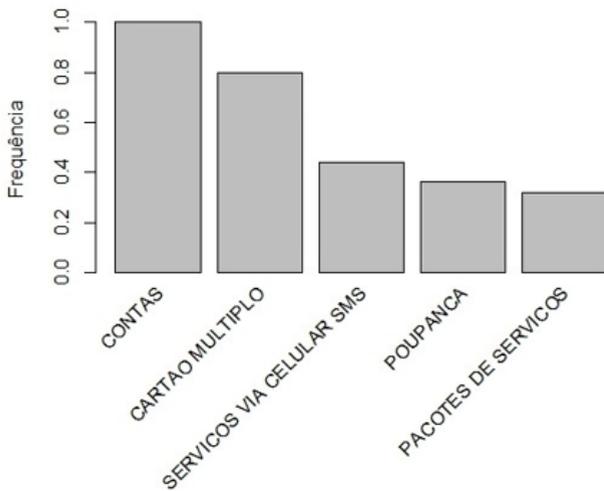


Figura 8. Frequency of men's products

lhs	rhs	support	confidence	lift	count
[1] {CARTAO MULTIPLO, PACOTES DE SERVICOS, SEGURO DE PESSOAS}	=> {CDC EMPRESTIMO}	0.054	0.66	5.0	295120
[2] {CARTAO MULTIPLO, CONTAS, PACOTES DE SERVICOS, SEGURO DE PESSOAS}	=> {CDC EMPRESTIMO}	0.054	0.66	5.0	295120
[3] {PACOTES DE SERVICOS, SEGURO DE PESSOAS}	=> {CDC EMPRESTIMO}	0.055	0.66	5.0	300510
[4] {CONTAS, PACOTES DE SERVICOS, SEGURO DE PESSOAS}	=> {CDC EMPRESTIMO}	0.055	0.66	5.0	300510
[5] {CARTAO MULTIPLO, SEGURO DE PESSOAS, SERVICOS VIA CELULAR SMS}	=> {CDC EMPRESTIMO}	0.065	0.60	4.6	357393

Figura 9. Random Rules for Male Customers

In this case, we can verify that we have, in all cases, when a person has People Insurance they tends to have CDC Loan. Differentiating from the goal set for women, let's put it that we want to hire Auto Insurance for men, what baskets of customers should we prioritize for contacts? The first 5 items were kept, sorted by confidence. Found 160 rules.

When receiving a task to perform Auto Insurance for men, we can consider, with confidence of 39% the following situations of customer baskets: 1 - Multiple Card, Special Check, Investment Fund, Receipt of Securities, Asset Insurance, Personal Insurance, Cellular SMS Service; 2 - Multiple Card, Special Check, Accounts, Investment Fund, Receipt of Securities, Asset Insurance, Personal Insurance, Cellular

lhs	rhs	support	confidence	lift	count
[1] {CARTAO MULTIPLO, CHEQUE ESPECIAL, EMISSAO DE TITULOS, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[2] {CARTAO MULTIPLO, CHEQUE ESPECIAL, EMISSAO DE TITULOS, POUPANCA, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[3] {CARTAO MULTIPLO, CHEQUE ESPECIAL, CONTAS, EMISSAO DE TITULOS, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[4] {CARTAO MULTIPLO, CHEQUE ESPECIAL, CONTAS, EMISSAO DE TITULOS, POUPANCA, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4322
[5] {CHEQUE ESPECIAL, EMISSAO DE TITULOS, TITULOS DE CAPITALIZACAO PAGAMENTO MENSAL}	=> {TITULOS DE CAPITALIZACAO PAGAMENTO UNICO}	0.0011	0.45	11	4348

Figura 10. Target specific for men clients

SMS Service; 3 - Special Check, Investment Fund, Securities Receipt, Asset Insurance, Personal Insurance, Cellular SMS Service; 4 - Special Check, Accounts, Investment Fund, Receipt of Securities, Asset Insurance, Personal Insurance, Cellular SMS Service; 5 - Multiple Card, Special Check, Investment Fund, Receipt of Securities, Asset Insurance, Personal Insurance; In all cases, we had a support of 0.11% and a lift of 18. More than 6,000 rules were found in each of the baskets.

### VIII. CONCLUSION AND FUTURE WORK

Three association rule algorithms were presented for Market Basket Analysis. It was conceived on the three algorithms, in the case AIS, SETM and Apriori, it was verified that both work in a very similar form and due to other recent studies prefer Apriori, being thus, the algorithm chosen for the study of the case.

Based on customers from the state of São Paulo, the most common products of each sex and which groups are most common in both cases. In this way it was verified that the interest between men and women differ in relation to the preference of purchase of products of the institution. It was also verified, given a product as "goal" of increment in the baskets of the customers, for which profile of customers should be offered these products. And we were able to achieve the goal proposed by the study, showing customers that have certain product groups, tend to acquire the product requested in the campaign, and thus, will have a greater probability of success in the offer to this profile of customers.

### Referências

- [1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules" between sets of items in large databases," in *Acm sigmod record*, vol. 22, pp. 207-216, ACM, 1993.
- [2] D. Dasgupta, Z. Ji, and F. Gonzalez, "Artificial immune system (ais) research in the last five years," in *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, vol. 1, pp. 123-130, IEEE, 2003.

- [3] M. Houtsma and A. Swami, "Set-oriented mining for association rules in relational databases," in Proceedings of the Eleventh International Conference on Data Engineering, pp. 25–33, Mar 1995.
- [4] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in ACM SIGMOD Record, vol. 26, pp. 255–264, ACM, 1997.
- [5] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in ACM sigmod record, vol. 29, pp. 1–12, ACM, 2000.



**FELIPE DE MOURA REZENDE DOS SANTOS** Possui graduação em Sistemas de Informação pela Faculdade de Pindamonhangaba (2008), MBA em Engenharia de Software (2014) e pós graduação em Banco de Dados e Business Intelligence (2016). Atualmente é assessor de unidade estratégica - Banco do Brasil - Direção Geral. Tem experiência na área de Ciência da Computação, com ênfase em Arquitetura de Sistemas de Computação.



**MARCELO LADEIRA** Possui graduação em Engenharia Mecânica pela Universidade de Brasília(1976), especialização em Curso Avançado de Operação de Sistema Hidrotérmico pela Universidade Federal do Rio de Janeiro(1981), mestrado em Análise de Sistemas e Aplicações pelo Instituto Nacional de Pesquisas Espaciais(1981), doutorado em Computação pela Universidade Federal do Rio Grande do Sul(2000) e pós-doutorado pela Universidade de Lisboa(2004). Atualmente é Professor Adjunto 2 da Universidade de Brasília, Membro de Comitê Consultivo do Ministério da Ciência, Tecnologia, Inovações e Comunicações, Revisor de periódico da Journal of Information and Data Management - JIDM e Membro de comitê assessor do Conselho Nacional de Desenvolvimento Científico e Tecnológico. Tem experiência na área de Ciência da Computação, com ênfase em Metodologia e Técnicas da Computação. Atuando principalmente nos seguintes temas: Inteligência Artificial, raciocínio probabilístico, diagramas de influências, redes bayesianas, Teoria da decisão e Representação de conhecimento incerto.

•••

•••

# Nomeando a dor da engenharia de requisitos no domínio de sistemas robóticos

TAYSE VIRGULINO RIBEIRO,  
LAURA MARIA PALOMINO MARINO e  
FABIO QUEDA BUENO DA SILVA.

Universidade Federal de Pernambuco, Centro de Informática (CIn) - UFPE, Recife-PE, Brasil  
(e-mail: {tvr, lmpm, fabio}@cin.ufpe.br)

Autor Correspondente: Tayse Virgulino Ribeiro (e-mail: tvr@cin.ufpe.br).

• **RESUMO** - A pesquisa no campo da engenharia de *software* tem sido explorada com maior ênfase nos últimos anos, dado o crescente grau de complexidade e sofisticação no avanço da tecnologia no mercado. A fim de contribuir para este último aspecto, é tido como conceito essencial a aplicação de práticas de engenharia de requisitos. No âmbito da Robótica, torna-se relevante uma vez que existem muitos tipos de robô que intervêm em atividades vitais ou que podem ter alguma repercussão para o ser humano. Nesse sentido, é imprescindível estudar como potenciais desenvolvedores nessa área, tendem a desenvolver seus projetos. Assim, o estudo visa explorar como ocorre este processo no grupo de pesquisa RobôCIn - UFPE, Brasil, enfocada especificamente em aspectos que são levados em consideração para a definição de requisitos. Isso, com a intencionalidade de avaliar o grau de correspondência entre o modelo proposto por [1] e a dinâmica de desenvolvimento no grupo. Para a coleta de dados, foram realizadas entrevistas semi-estruturadas durante dois meses (outubro e novembro de 2018) e aplicação de questionários. Depois de analisar as entrevistas e os questionários, apresentamos resultados que avaliam características do modelo, descrevendo em que medida eles corroboram ou não o mesmo.

• **PALAVRAS-CHAVE** - Sistemas Robóticos, Definição de Requisitos, Engenharia de Requisitos.

## I. INTRODUÇÃO

Um sistema robótico é uma combinação de algumas peças - *hardware* para montagem de sistemas e *software* para operações do sistema - que deve ser integrado para permitir que um sistema robótico funcione como esperado [1]. No dicionário de Oxford [2], um robô é definido como uma máquina que pode executar uma série complicada de tarefas automaticamente, especialmente uma programável por um computador. Espera-se que os sistemas robóticos auxiliem ou substituam suas contrapartes humanas para um desempenho eficiente e eficaz de todos os tipos de tarefas [3].

Existem muitos tipos de robôs e eles são usados em diferentes ambientes, embora sejam muito diversos em aplicação e forma. Além disso, a tecnologia robótica está se expandindo das aplicações industriais para as domésticas, criando assim novos tipos de robôs, tais como companheiros robóticos, robôs de logística, robôs para vigilância e intervenção, e robôs educacionais. Os robôs mencionados acima devem ser capazes de fornecer serviços para tornar a vida diária dos seres humanos muito mais fácil; com tarefas típicas, incluindo tratamento médico de pacientes na privacidade

de suas próprias casas, transporte de pacientes ou pessoas com deficiência para hospitais e locais públicos, proteção de moradores e pertences de pessoas, educação de crianças, companheirismo social para idosos, etc [4].

No entanto, essa presença onipresente carrega um alto nível de dependência, o que inevitavelmente exige a necessidade de sistemas cada vez mais disponíveis, confiáveis e seguros. A literatura sobre Sistemas Críticos de Segurança (SCS) relatou muitos casos em que sistemas falharam devido à falta de especificações de requisitos ou mal-entendidos relacionados a problemas na engenharia de requisitos, contribuindo para acidentes que causam danos ao meio ambiente, ferimentos em pessoas e até mesmo a perda de vidas [5].

De acordo com isto, o desenvolvimento com qualidade destes sistemas adquire uma importância relevante, o que implica explorar metodologias de engenharia de *software* para abstrair complexidades e aumentar a eficiência para modelar, desenvolver, manter e evoluir sistemas robóticos de forma rentável [1].

Portanto, este trabalho teve por intenção avaliar o grau de correspondência entre o modelo proposto por [1] e a dinâ-

mica de desenvolvimento do grupo de pesquisa RobôCIn, a fim de verificar quais aspectos são levados em consideração na definição de requisitos para desenvolver projetos no RobôCIn atualmente.

O artigo está organizado da seguinte maneira. Na seção II, apresentamos um breve contexto sobre as teorias de sistemas robóticos, engenharia de requisitos e trabalhos relacionados. Na seção III, descrevemos o método do estudo realizados. Na seção IV, os resultados do estudo de caso apresentado. Finalmente, na seção V, são apresentadas conclusões e orientações para trabalhos futuros.

## II. REFERENCIAL TEÓRICO E TRABALHOS RELACIONADOS

Esta seção apresenta os principais conceitos, teorias e desafios de pesquisa direcionadas ao estudo da robótica relacionado com definição de requisitos, sendo que o primeiro passo para o desenvolvimento de um sistema robótico é determinar seus requisitos, que são descrições de como o sistema deve se comportar e as restrições associadas à sua operação [3].

O *Institute of Electrical and Electronics Engineers* (IEEE) apresenta o padrão 830 que define a Engenharia de Requisitos (RE) como uma técnica de medição do contato entre comprador e fornecedor para estabelecer e manter os requisitos necessários para o sistema, *software* ou serviço de interesse, concentrando-se na descoberta, elicitação, desenvolvimento, análise, validação, comunicação, documentação e gerenciamento de requisitos. O objetivo da Engenharia de requisitos é então determinar as propriedades que um sistema deve ter para obter sucesso, sendo algumas vezes relativamente simples e outras não [6].

Nos últimos anos, tem havido um fluxo contínuo de pesquisas relatadas sobre técnicas de definição de requisitos para sistemas robóticos. A partir da complexidade de tais sistemas, requer a coleta sistemática de requisitos em níveis não considerados em sistemas gerais. No entanto, a maioria das técnicas de definição de requisitos para sistemas robóticos vem das que normalmente são usadas em sistemas gerais. Assim, acreditamos que várias características técnicas relacionadas aos sistemas robóticos não são adequadamente abordadas pelas técnicas de definição de requisitos no estado da arte [1].

Na área da robótica alguns dos trabalhos relacionados e que vale fazer referência são: [7] relataram as soluções que suportam projeto, desenvolvimento e operação de sistemas robóticos baseados em serviços de *software* em uma revisão sistemática do Desenvolvimento Orientado a Serviços de Sistemas Robóticos com base em 39 estudos primários publicados de 2005 a 2011. [8] conduziram outra revisão sistemática de abordagens de engenharia de *software* para robótica com base em 67 estudos publicados de 1999 a 2011. A revisão destaca a aplicação de desenvolvimento de robótica baseado em componentes, orientado a serviços e orientado a modelos, conforme as tendências de pesquisa emergentes. [9] realizaram uma revisão sistemática da literatura buscando identificar abordagens em técnicas de desenvolvimento ori-

entado pelo modelo (MDD), e suas principais características técnicas, bem como os tipos de requisitos específicos, questões comportamentais e sociais.

Finalmente, [10] também usaram o método SMS para identificar e analisar 56 artigos publicados de 1991 a 2015. Esta revisão taxonomicamente classificou as pesquisas existentes e sistematicamente mapeou as soluções, estruturas, notações e métodos de avaliação para destacar o papel da arquitetura de *software* em sistemas robóticos.

No entanto, para nosso conhecimento, não há estrutura para suportar adequadamente RE para sistemas robóticos. Os esforços de pesquisa relacionados a sistemas robóticos utilizam uma área de atuação específica, algumas delas são as associadas a Sistemas de Segurança Crítica (SCS), em especial na melhoria do processo de engenharia de requisitos da SCS [5], também neste assunto podemos citar a pesquisa relacionada com requisitos de segurança em sistemas robóticos teleoperados e, por extrapolação, em outros domínios críticos [11]. Outro exemplo é a revisão feita dos robôs especificamente selecionados ou configurados para atender aos requisitos da indústria de processamento de alimentos [12] ou aqueles direcionados à área de requisitos para robôs domésticos [4], etc.

De maneira mais geral encontramos PDF, um *framework* de design unificado para sistemas de robôs móveis que pode ser usado para orientar a análise de requisitos, o *design* e a implementação de novos sistemas de robôs ou a classificação de sistemas de robôs existentes [13]. Mais recentemente destaca-se o trabalho de [1], onde propõe identificar algumas características técnicas que afetam diretamente a tarefa de definição de requisitos para sistemas robóticos. Esta pesquisa teve por objetivo recuperar o maior número possível de estudos relevantes para evitar possível viés de seleção da literatura.

## III. MÉTODO

Nossa pesquisa visa explorar definição de requisitos em especificidades de contextos. A partir da análise do estado da arte sobre o assunto, surgiu a necessidade de realizar um estudo exploratório para avaliar como potenciais desenvolvedores nessa área tendem a desenvolver seus projetos. Assim, a pesquisa tem como contexto o grupo RobôCIn - UFPE, Brasil, e é voltada especificamente para as características da definição de requisitos; com a intenção de avaliar o grau de correspondência entre o modelo proposto por [1] e a dinâmica de desenvolvimento do grupo. Nossa principal pergunta de pesquisa é formulada da seguinte forma:

- *Quais os aspectos são levados em consideração na definição de requisitos para desenvolver projetos no RobôCIn?*

*Estratégia de seleção dos participantes:* todo o procedimento de seleção foi intencional. Primeiro, este caso considera grupos de pesquisa no domínio da robótica, que de acordo com o contexto apresentado, especificamente o grupo de pesquisa RobôCIn. Que trabalha com intuito de solucionar problemas na área da robótica. Este é um estudo exploratório

e a população-alvo são os engenheiros de *software*, especificamente com especialização e/ou interesse em áreas da robótica e afins. A unidade de análise são engenheiros de *software*, com especialidade e/ou interesse na área de sistemas robóticos que integram o RobôCIn. Portanto, os dados foram coletados de fontes com perfis característicos, a fim de possibilitar triangulação dos dados de acordo com o objetivo da pesquisa. Baseado nisso, os participantes foram escolhidos para garantir uma boa cobertura no nível de adequação ao estudo, no que diz respeito à formação, qualificação e atividades relacionadas ao contexto estudado. Além disso, levou-se em consideração a função dos participantes no grupo, entre outros fatores. Também foram coletados dados relacionados à forma organizacional de trabalho da equipe, tanto com o grupo como com o gerente de projetos do grupo de pesquisa.

**Coleta de dados:** para alcançar o objetivo, foram feitas entrevistas semiestruturadas durante dois meses (outubro e novembro de 2018) em conjunto com a aplicação de questionários. Foram realizadas entrevistas com 5 engenheiros de *software*, destes, 1 gerente de projeto, 2 pesquisadores da pós-graduação integrados no grupo do RobôCIn com diversas atividades já realizadas, 1 engenheiro de *software* na área de 2D e 1 engenheiro de *software* na área lógica.

A entrevista era semi-estruturada, ou seja, utilizamos um *script* já projetado, com algumas perguntas abertas, como também um questionário que foi aplicado ao fim da entrevista. A entrevista incluiu vários tipos de perguntas, destinadas a explorar o nível de adequação do participante ao estudo, características do modo de desenvolvimento de *software/hardware* da equipe, aplicações de técnicas e práticas de Engenharia de Requisitos, opinião e valores sobre a definição de requisitos no processo de desenvolvimento de *software/hardware* para o domínio da robótica. O refinamento das entrevistas foi realizado baseado na validação com o participante, numa entrevista realizada após o primeiro contato (veja a Tabela 1).

**Preocupações éticas:** os participantes do grupo de pesquisa RobôCIn assinaram o TCLE - Termo de Consentimento Livre e Esclarecido com a pesquisa. Cada participante foi contatado pela primeira vez com antecedência, e cada entrevista ocorreu em uma sala de reunião privada nas próprias instalações do Centro de Informática - CIn da UFPE, durante os meses de outubro e novembro. Ao assinarem o TCLE, os participantes garantiram a confidencialidade dos dados fornecidos, anonimato da participação e o direito de poderem desistir de participar da pesquisa a qualquer momento. As entrevistas gravadas totalizaram juntas um somatório de 3 horas e 24 minutos de tempo de áudio.

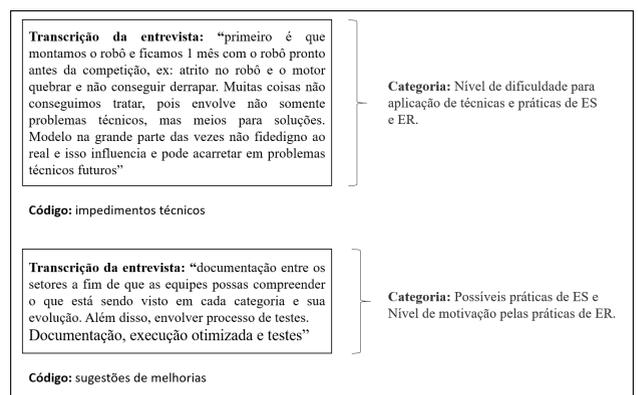
Após as entrevistas os participantes foram contatados no intuito de refinar informações características relevantes sobre o ambiente e forma de desenvolvimento. As informações serviram para complementar relatos incompletos que foram registrados durante as entrevistas registradas.

**Análise dos dados:** de acordo com [14], a construção de categoria diz respeito a análise de dados. Com isso, o processo é baseado na compreensão do objetivo da análise, construção

**Tabela 1.** Apresentação minimizada do guia da entrevista

<p>Q1. Qual seu histórico de formação? - Técnico, Graduação, Especialização, Certificação, Mestrado, Doutorado - Poderia informar também em qual instituição realizou sua formação, como também a localização, duração. - Quais foram as atividades acadêmicas exercidas durante este período? ...</p> <p>Q3. Tem alguma experiência curricular em desenvolvimento para sistemas robóticos? - Quais principais atividades já realizadas?</p> <p>Q4. Como é que você desenvolve? - Em qual etapa de desenvolvimento você está envolvido(a)? Implementação/Lógica; Arquitetural; Hardware - Você utiliza alguma metodologia ou processo para auxiliar no desenvolvimento? - Existem situações em que você se desvia desse padrão ou modo de fazer? ...</p> <p>Q8. De acordo com sua consideração, as práticas de engenharia de requisitos são abordadas no desenvolvimento de sistemas robóticos?</p> <p>Q9. Você acha que existem características técnicas que impedem a utilização desta prática?</p> <p>Q10. Você tem alguma sugestão de melhoria para esse processo de desenvolvimento? ...</p> <p>Q12. Como você modela os requisitos internos do seu sistema robótico? Descreva as entidades envolvidas</p> <p>Q13. Como você modela os requisitos externos do seu sistema robótico? Descreva as entidades envolvidas</p>
---

da categoria e análise qualitativa dos dados. O áudio das entrevistas foi transcrito e para obtenção da codificação foi realizada em forma *Ad-hoc*, neste processo dois integrantes codificaram enquanto outro categorizou para apoiar o processo da análise. Após isso, foram rotulados determinados trechos das entrevistas, isso devido a conformidade que os conceitos se tornavam mais claros na pesquisa, conforme é apresentado na Figura 1.



**Figura 1.** Representação da codificação de uma categoria

De acordo com o que é apresentado na Figura 1, os códigos foram rotulados de forma a identificar as respectivas categorias, levando a causas explicativas destes conceitos, no intuito de sustentar e responder o questionamento da pesquisa.

#### IV. RESULTADOS E DISCUSSÕES

## A. DESCRIÇÃO DO CONTEXTO: GRUPO DE PESQUISA ROBÔCIN

O estudo de caso foi realizado no grupo RobôCIn (<https://robocin.github.io/>). É um grupo de pesquisa do Centro de Informática da UFPE, que resolve problemas voltados para a área da robótica. O grupo foi criado em 2015 por 12 alunos e pelos professores Edna Barros e Hansenclever Bassani. Hoje contam com 35 membros e 4 professores para coordenar a equipe para alcançar soluções inovadoras usando A.I., visão computacional, mecânica e eletrônica combinadas em robôs. O grupo atualmente tem pesquisas focadas em quatro categorias, que são: SSL (*Small Size League*), VSSS (*Very Small Size Soccer*), 2D (Simulação 2D) e @Home.

A estrutura organizacional do grupo trabalha de forma horizontal. Dos 35 membros do grupo de pesquisa, 1 concentra suas atividades na área de gerenciamento de projetos e algumas atividades de implementação/lógica. Enquanto isso, os demais são caracterizados como engenheiros/pesquisadores de sistemas robóticos alocados nas categorias apresentadas acima. O grupo é focado também em apresentar seus resultados em competições de robóticas, tanto que durante o período de competições o fluxo de atividades se torna bastante intenso, tanto em nível organizacional, quanto operacional. Após o período de competições o grupo se estende em atividades práticas de pesquisa e correções de eventuais problemas ocorridos durante as competições.

## B. VALIDAÇÃO DO MODELO

### 1) Descrição do Framework

Neste estudo qualitativo foi utilizado como base um *framework* para definição de requisitos para o domínio de sistemas robóticos. O contexto de estudo sobre o *framework* consiste em proporcionar benefícios aos pesquisadores interessados em conhecer o estado da arte da definição de requisitos para os sistemas robóticos; e praticantes que possam estar interessados em compreender as técnicas adequadas para a definição de requisitos para este domínio [1]. Neste sentido foi proposto uma estrutura para identificar requisitos robóticos, com o objetivo de lidar com os desafios e problemas descritos em sua pesquisa. Na Figura 2 é apresentada uma abstração dos requisitos. Durante a tarefa de definição de requisitos, várias características técnicas internas e externas específicas para sistemas robóticos devem ser consideradas [1]. Esses recursos técnicos serão descritos brevemente a seguir:

*Nível de inteligência:* este conceito está relacionado a capacidade em que o robô tem em expressar um comportamento útil. Isso associado a conceitos como, como: autonomia, deliberatividade e adaptabilidade.

*Tecnologia escolhida:* devido a variedade de aplicações na área de *software* e *hardware* na composição de sistemas robótico, leva-se em consideração a escolha de aplicações específicas ao se desenvolver para sistemas robóticos, segundo [1].

*Avaliação de Requisitos de Tempo de Execução:* segundo [1], os sistemas robóticos geralmente requerem um meca-

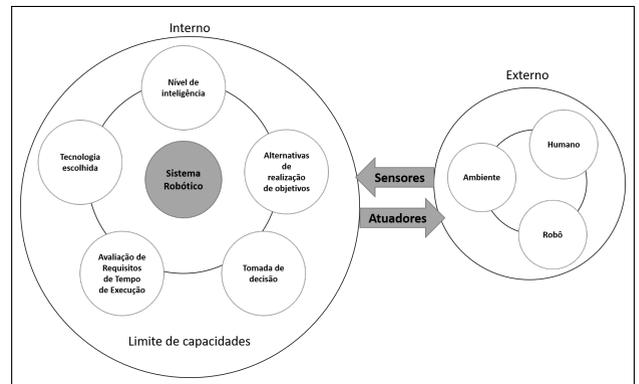


Figura 2. Abstração de requisitos para sistema robótico

nismo para monitorar e avaliar o sucesso ou a falha de alcançar os objetivos do sistema em tempo de execução. E esse mecanismo ocorre durante o processo de avaliação de requisitos em tempo de execução.

*Tomada de decisão:* o conceito de tomada de decisão está associado ao requisito no que diz respeito a ajudar a decidir sobre a ação a ser tomada pelo sistema robótico.

*Alternativas de realização de objetivos:* o requisito de alternativas de realização de objetivos tem o propósito de apresentar alternativas de atingir metas para objetivos ao se desenvolver para sistemas robóticos.

*Interações do Sistema:* o conceito de interação de sistemas aborda a utilização de sensores e atuadores, e essas interações se dão por meio destes. É relatado ainda que, os sensores permitam que o robô perceba as mudanças do ambiente e reaja a eles alterando seu próprio comportamento. Enquanto os atuadores são as ferramentas com as quais o robô influencia no ambiente [1].

*Recursos externos:* os requisitos externos estão associados a interação de ambiente, robô e humano. Isso, devido a ambientes estarem vinculados a meios controlados e não controlados, que afeta por consequência o comportamento do robô. Além disso, a colaboração de robôs para robôs requer um envolvimento de nível de capacidade computacional no que diz respeito a troca de informações e execução de atividades, como também o fator do humano como colaborador do ambiente em atividades robóticas. Levando em consideração sua interação e exploração de conceitos no que diz respeito a tecnologias estabelecidas no contexto utilizado.

### 2) Recursos aplicados pelo RobôCIn

No RobôCIn atualmente, o planejamento de trabalho é feito a partir do pressuposto para cada competência ou robô a ser desenvolvido, eles tem um objetivo e iniciam com o levantamento de requisitos dos itens desejados para o robô e *software*; nesse sentido os diferentes fatores para o desenvolvimento de Robô são atendidos, tais como *Nível de Inteligência*, *Tecnologia*, *Avaliação de Requisitos de Tempo de Execução*, *Tomada de decisão*, *Alternativas de realização de objetivos*, *Interações do sistema*.

As entrevistas realizadas em conjunto com a aplicação dos questionários evidenciaram o grau no qual esses aspectos são levados a consideração para modelar os requisitos e a importância que os entrevistados outorgam a cada um particularmente.

Como é possível observar na Tabela 2, os integrantes do projeto atualmente trabalham, como parte da dinâmica de desenvolvimento deles, os elementos do *framework* proposto, em maior ou menor grau. Assim o *Nível de inteligência*, é levado em consideração na maioria das atividades que eles desempenham de forma implícita, o que significa que é um elemento básico, ou seja, indispensável mas é contemplado como parte da autonomia do robô.

A escolha da *Tecnologia* é aplicado também e é indispensável para o desenvolvimento do robô. Eles otorgam uma importância entre moderada e muito alta, com tendência a prevalecer o último critério de avaliação.

Do mesmo jeito, nas entrevistas se evidencia que os desenvolvedores também aplicam e atribuem uma importância geralmente muito alta, a *Tomada de decisão*, que é referido quando o robô realiza uma tomada de decisão, se tem alternativas; por exemplo quando eles vão fazer um passe para outro robô ou um gol no time adversário, no contexto de jogo de futebol.

As *Interações do sistema* é um ponto sempre pensando antes de desenvolver o robô, pois é referida a comunicação entre o computador e o robô que geralmente não é modificada, porém alguns deles consideram ter pouca relevância na prática.

Finalmente, os aspectos *Alternativas de realização de objetivos* e *Avaliação de Requisitos de Tempo de Execução* são atualmente aplicados com menos ênfase, pois eles ainda estão sendo avaliados. A importância que eles atribuem a esses fatores varia de baixa a muito alta, ou seja, não há consenso entre eles.

Ao explorar se eles levam a consideração características externas, como as propostas por [1] no seu modelo, as respostas tendem a apontar que atualmente a equipe considera fundamentalmente *Ambiente e Robô*. O ambiente, onde eles relatam que é pensado sempre antes de construir um robô, pois o mesmo pode interferir se existisse algum nível de obstáculo ou se é dinâmico ou não. Referente a interação com robô eles referem que consideram a comunicação, dados que são transmitidos e alternativas que podem ser compartilhadas a fim de atingir o mesmo objetivo; assim como também a influência de outros robôs. O fator *Humano* já é menos referido, o que indica que atualmente eles desconsideram o mesmo. Particularmente na categoria de 2d a modelagem de requisitos externos não atende.

3) Análise qualitativa comparativa entre o framework e o contexto estudado

Após a realização das entrevistas, o próximo passo foi realizar a análise dos dados, apresentando a relação entre o conceito de definição de requisitos e a forma de desenvolvimento do estudo realizado com o grupo de pesquisa

RobôCIn. Foi aplicado o uso de proposições para apresentar os elementos que se conectam a teoria e aos dados coletados nas entrevistas. Além disso, foi utilizada a relação de fatores influenciadores que contribuem e/ou são relevantes ao se desenvolver com qualidade sistemas robóticos, ou sistemas críticos.

Primeiro, a definição de requisitos é um fator importante para o desenvolvimento com qualidade de sistemas críticos, o que implica explorar metodologias de engenharia de *software*. Com isso, abstrair complexidades e aumentar a eficiência para modelar, desenvolver, manter e evoluir sistemas robóticos [1], como a área da engenharia de requisitos. Além disso, resultando em experiências no ambiente de trabalho, em níveis organizacionais e forma de trabalho dos membros da equipe.

Conforme relatado pelos participantes, o ambiente de trabalho do grupo de pesquisas é altamente sociável e leva em consideração princípios de uma gestão horizontal no intuito de facilitar e manter uma integração da equipe. Além disso, o ambiente conta com membros de diversos backgrounds, em nível de formação e experiência curricular/extracurricular. Para tanto, observa-se os primeiros fatores influenciadores no que diz respeito a aplicação de conceito das práticas de engenharia de requisitos.

- **Fator 01:** *nível organizacional de trabalho do grupo e equipe*, são condições para verificar a aplicação dos conceitos de sistemas robóticos com as práticas de engenharia de requisitos.

Vale ressaltar que no grupo não contém nenhum especialista e/ou pesquisador do domínio na área da engenharia de *software*.

Os entrevistados possuem um perfil associado ao desenvolvimento na área de robótica, tendo em vista que todos pertencem ao grupo de pesquisa RobôCIn, da UFPE. Dos entrevistados, 3 possuem grau de formação de estudantes de graduação em Engenharia da Computação e 2 são profissionais formados que cursam mestrado e doutorado associados a linha de pesquisa a robótica. Destes, 2 referem-se ao trabalho em implementação e lógica, e um em um nível mais organizacional, como gerente de projeto, embora ele também se dedique à implementação e à lógica.

Nesse caso, o ambiente e os membros a quais estão integrados estão condicionados a fatores organizacionais de funcionamento da empresa. Como também, são caracterizados como fatores influenciadores os processos de desenvolvimento e a forma de trabalho. Assim sendo:

- **Fator 02:** *nível processos de desenvolvimento no setor de trabalho do grupo e nível de dificuldade para aplicação de técnicas de engenharia de requisitos*, são condições para verificar a aplicação dos conceitos de sistemas robóticos com as práticas de engenharia de requisitos

Em relação ao planejamento do trabalho, os entrevistados referem que eles têm um padrão de desenvolvimento, ainda quando não seguem uma metodologia em específica. Eles

**Tabela 2.** Avaliação dos fatores internos de modelagem por cada entrevistado e valorização da importância dos fatores no trabalho

	Entrevista 1	Entrevista 2	Entrevista 3	Entrevista 4	Entrevista 5
<b>Aspectos internos</b>					
<b>Nível de inteligência</b>	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Nunca apliquei
<b>Tecnologia</b>	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero
<b>Avaliação de Requisitos de Tempo de Execução</b>	Já apliquei e considero	Considero aplicar, avaliando	Considero aplicar, avaliando	Já apliquei e considero	Considero aplicar, avaliando
<b>Tomada de decisão</b>	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero
<b>Alternativas de realização de objetivos</b>	Já apliquei e considero	Considero aplicar, avaliando	Considero aplicar, avaliando	Já apliquei e considero	Já apliquei e considero
<b>Interações do Sistema</b>	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Já apliquei e considero	Nunca apliquei
<b>Nível médio de importância dos recursos</b>	Muito alto	Alto	Moderado	Muito alto	Baixo
<b>Legenda</b>	Muito alto	Alto	Moderado	Baixo	Muito Baixo

partem do pressuposto que tem um objetivo. Então começam a pensar na linguagem, arquitetura do *software* (de acordo com a categoria do robô), o que será feito primeiro, ordem de implementação das classes. Ainda assim, levando em consideração possíveis problemas que possam ser encontrados no caminho para o desenvolvimento do objetivo. As equipes se baseiam em pesquisas, projetos e/ou *Team Description Paper* (TDP) para aprimorarem e/ou desenvolverem o seu próprio robô.

No que tange a organização do trabalho, o RobôCIn se estrutura em categorias, que são: SSL, VSSS, 2D, @Home. Com isso, cada equipe é uma área e cada área contém um coordenador. As equipes têm suas subdivisões onde fazem *deadlines*, buscando trabalhar sempre com reuniões de alinhamento, com periodicidades semanais, entre as equipes. Periodicamente as diferentes áreas realizam uma *sprint review* do que está acontecendo nas categorias. Nestas categorias, têm pessoas responsáveis para fazer o robô e lógica do robô (estratégia). Os grupos compartilham em algum momento também atividades, no que tange a funções de suporte e aprendizado. Para gerenciar melhor o trabalho na equipe, são empregadas algumas ferramentas como Git, Trello. Porém, os entrevistados referem que por vezes fica difícil manter um fluxo organizacional; devido a área ser voltada à experimentação e estarem bastante envolvidos com competições.

Outro fator que os entrevistados descreveram foram os impedimentos técnicos, a fim de influenciar na definição de requisitos. Isso a nível de *hardware*, pois a maior parte dos requisitos surgem conforme o desenvolvimento é realizado, e a recorrência de mudança é maior e é difícil de mitigar. Mas, acham esse fator normal em correspondência com a área da robótica. Relataram que tentam tratar, mitigar, mas nem sem-

pre é possível. Então sim, características técnicas dificultam a utilização de prática de requisitos, mas não impedem, tudo é questão de planejamento e visão de possíveis cenários.

Portanto, os entrevistados manifestaram certo nível de conscientização das práticas de ER, eles utilizam da prática de engenharia de requisitos quando vão começar a desenvolver, sem burocracia, “o mais horizontal possível”; ainda eles reconheceram que é possível construir um robô completo com um planejamento inicial, visando chegar num produto final. Pois leva-se em consideração os itens desejados e a partir disto modelar o objetivo do negócio. Porém, segundo os entrevistados, muito se perde em nível de hardware. Para os integrantes entrevistados, o processo ideal deve levar em consideração os seguintes itens:

- 1) O que preciso atingir em nível de autonomia no robô (nível de inteligência);
- 2) O tipo de tecnologia e modelos, ou seja, como o robô será avaliado/testado;
- 3) Como o robô será mapeado no mundo simulado do real (treinar);
- 4) Ambiente (identificar como e onde vai funcionar).

Ademais atender ao desenvolvimento (estratégia, arquitetura, práticas) como a aplicação de metodologias, correspondência entre o protótipo e produto final, o grau de manutenção de *software*, a documentação de código e o processo de testes

## V. CONCLUSÃO

Existem algumas características de práticas na etapa de desenvolvimento, que indicam aplicação de técnicas de Engenharia de Requisitos. Baseado nisso, os entrevistados referem que eles pensam nas alternativas de objetivos que o robô precisa realizar, fazendo uma análise dos requisitos, mas de

forma implícita, ou seja, é realizada uma discussão entre a equipe para alinhar os itens a serem desenvolvidos. Com isso, eles iniciam com a etapa de levantamento de requisitos dos itens desejados para o robô, a nível de *software* e *hardware*, mas sem conseguir ter um planejamento completo de sua arquitetura.

De acordo com isso, existem vantagens e desvantagens a serem debatidas. A ideia de fazer planejamento, mesmo sem muita experiência, com diagramas e requisitos, a nível básico, legendado resulta em código organizado e fácil manutenção, segundo entrevistados. Como as categorias são bem divididas, as atividades são bem definidas. Portanto, esses fatores são caracterizados como pontos positivos, conforme é apresentado pelos entrevistados.

Porém, como o desenvolvimento não segue uma metodologia em específico, existe falta de documentação, tanto em nível de arquitetura que não segue nenhum padrão, tanto de *software* e pode haver a ocorrência de erros neste processo: códigos mal estruturados, sem definição de esqueleto nem estratégia; ou podem acontecer dificuldades quando é preciso fazer alterações pontuais, pois as mesmas acabam ganhando uma complexidade indesejada.

Na concepção do *hardware*, não seguir o processo comum de planejamento, prototipação e teste do produto pode levar a sérios problemas. Quando acontece erro no planejamento após *hardware* concebido, se precisa de soluções paliativas, esses são alguns dos pontos negativos, segundo a experiência de desenvolvimento relatada pelos entrevistados.

Conforme apresentado nesta pesquisa, na intenção de avaliar o grau de correspondência entre o modelo proposto por [1] e a dinâmica de desenvolvimento do grupo de pesquisa RobôCIn. Com isso, verificar quais os aspectos são levados em consideração na definição de requisitos para desenvolver projetos no RobôCIn.

Conforme é apresentado também na Tabela 2, foram detectadas correspondências entre o *framework* e em como os desenvolvedores fazem a modelagem dos requisitos para sistemas robóticos. No caso dos requisitos internos, levando em consideração *Nível de Inteligência (Aplicação de nível autônomo), Tecnologia, Avaliação de Requisitos de Tempo de Execução, Tomada de decisão, Alternativas de realização de objetivos, Interações do Sistema (sensores e atuadores)*.

De igual modo a modelagem de requisitos externos em sistemas robóticos: *Ambiente* (ambiente dinâmico ou estático, controlado ou não controlado) *Robô, Humano*. No caso do RobôCIn, na categoria de 2d a Modelagem de requisitos externos não atende completamente, pois os mesmos não levam em consideração interações a nível humano.

A fim de abstrair complexidades e aumentar a eficiência para modelagem, desenvolvimento, manutenção e evolução de sistemas robóticos com boa relação custo-benefício [1]. A engenharia de software, com linha de pesquisa centralizada em RE vem explorando sobre técnicas de definição de requisitos para sistemas robóticos.

Fundamentado nisso, este artigo traz como abordagem inicial a avaliação do *framework* para abstração de requisitos

para sistemas robóticos. No intuito de explorar conceitos já analisados e elaborados, com base na literatura, como também no intuito de investigar na prática o estudo realizado. Com isso, foi conferido que o modelo proposto por [1], contemplou fatores bem básicos na definição de RE no desenvolvimento para os sistemas robóticos em um contexto como o Robô Cin; recomendamos que o mesmo seja avaliado em um estudo mais aprofundado com a finalidade de ajustar o *framework* ou enriquecer com o foco de conseguir a robustez precisa.

## Referências

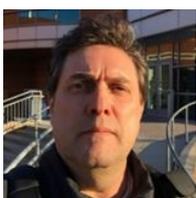
- [1] Albuquerque, D., Castro, J., & Sousa, A. A Requirements Definition Framework for the Robotic Systems Domain-An Exploratory Study. (2018, September). In 21st Workshop on Requirements Engineering (WER). CibSE.
- [2] Oxfordlearnersdictionaries.com. (2018). robot noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com. [online] Available at: <https://www.oxfordlearnersdictionaries.com/us/definition/english/robot?q=robot> [Accessed 5 Dec. 2018].
- [3] Albuquerque, D., Castro, J., Ribeiro, S., & Heineck, T. Requirements Engineering for Robotic System: A Systematic Mapping Study. (2017, maio). In 20th Workshop on Requirements Engineering (WER). CibSE.
- [4] Mitka, E., Gasteratos, A., Kyriakoulis, N., & Mouroutsos, S. G. (2012). Safety certification requirements for domestic robots. *Safety science*, 50(9), 1888-1897.
- [5] Martins, L. E. G., & Gorschek, T. (2016). Requirements engineering for safety-critical systems: A systematic literature review. *Information and software technology*, 75, 71-89.
- [6] Heineck, T. (2016). Desenvolvimento orientado a modelos no domínio de robótica: uma revisão sistemática da literatura.
- [7] Oliveira, L. B. R., Osório, F. S., & Nakagawa, E. Y. (2013, March). An investigation into the development of service-oriented robotic systems. In Proceedings of the 28th annual ACM symposium on applied computing (pp. 223-228). ACM.
- [8] Pons, C., Giandini, R., & Arévalo, G. (2012). A systematic review of applying modern software engineering techniques to developing robotic systems. *Ingeniería e Investigación*, 32(1), 58-63.
- [9] Heineck, T., Gonçalves, E., Sousa, A., Oliveira, M., & Castro, J. (2016, September). Model-Driven Development in Robotics Domain: A Systematic Literature Review. In Software Components, Architectures and reuse (SBCARS), 2016 X Brazilian Symposium on (pp. 151-160). IEEE.
- [10] Ahmad, A., & Babar, M. A. (2016). Software architectures for robotic systems: A systematic mapping study. *Journal of Systems and Software*, 122, 16-39.
- [11] Navarro Martínez, E. M., Sánchez Palma, P., Letelier Torres, P., Pastor Franco, J. Á., & Ramos Salavert, I. (2006). A goal-oriented approach for safety requirements specification.
- [12] Iqbal, J., Khan, Z. H., & Khalid, A. (2017). Prospects of robotics in food industry. *Food Science and Technology*, 37(2), 159-165.
- [13] Goodwin, J. R., & Winfield, A. (2008). A unified design framework for mobile robot systems (Doctoral dissertation, University of the West of England, Bristol).
- [14] Merriam, S. B., & Tisdell, E. J. (2015). Qualitative research: A guide to design and implementation. John Wiley & Sons.



**TAYSE VIRGULINO RIBEIRO** Mestranda em Ciência da Computação no CIn-UFPE, na Área de Engenharia de Requisitos e pesquisadora no Requirements Engineering Laboratory - LER (<http://www.cin.ufpe.br/ler/>). Graduada em Sistemas de Informação pelo Centro Universitário Luterano de Palmas - TO. Com mais de 4 anos de experiência no mercado, centraliza suas atividades na área de Engenharia de Software e afins (Líder de projetos, Scrum Master, Desenvolvedora Web).



**LAURA MARÍA PALOMINO MARIÑO** Mestranda em Ciência da Computação no CIn-UFPE. Possui graduação em Engenharia da Informática pela Universidad de las Ciencias Informáticas in La Habana.



**FABIO QUEDA BUENO DA SILVA** Possui graduação em Ciência da Computação pela Universidade de Campinas (1987). É PhD em Ciência da Computação pelo Laboratory for Foundations of Computer Science, University of Edinburgh, Escócia (1992). Possui pós-doutorado na Universidade de Toronto, Canadá, na área de engenharia de software (2011-2012). Entre 2016 e 2017, trabalhou como Senior Visiting Research Scientist na University of Maryland Baltimore County, Estados Unidos.

É Professor Titular do Centro de Informática da UFPE, onde trabalha desde 1993. Participa como revisor de diversas revistas internacionais, incluindo IEEE Software, IEEE Transactions on Software Engineering, Empirical Software Engineering Journal, Information and Software Technology Journal, Journal of Systems and Software, e Journal of Computer Standards & Interface.

...

# It change management process improvement: the case study of a brazilian court

**EDUARDO DA SILVA SOUSA,  
CARLOS EDUARDO MACHADO PIRES,  
SIMONE BORGES SIMAO MONTEIRO and  
ANA CARLA BITTENCOURT REIS**

Computer Science Department, University of Brasília-(UnB), P.O. Box 4466, 70910-900, Brasília-DF, Brazil  
(e-mail:{edussousa, kadumpires}@gmail.com, {simoneborges, anacarlabr}@unb.br)  
Corresponding author: Eduardo da Silva Sousa(e-mail: edussousa@gmail.com).

• **ABSTRACT** - IT became an indispensable structure for organizations, whether public or private. Although it is considered a support area, IT plays a strategic role in organizations, its procedures must be systematically conducted. This article addresses best practices in IT Service Change Management and compares these practices with current implementation in a court of law. The purpose of this comparison is to verify the adherence of the implemented process and to suggest adjustments that can guarantee quality of services. For that, a review of the pertinent bibliography and a case study of the processes implemented and executed by the court was carried out. The preliminary comparative analysis demonstrated that the change management activities of the court's IT services have opportunities for improvement both in the process and for implementation of control mechanisms, which were suggested by the researchers.

• **KEYWORDS** - ITSM, ITIL, service management, change management, process improvement, quality of service.

## I. INTRODUCTION

When observing the development of companies, it is noticed that Information Technology (IT) systems have become increasingly complex, heterogeneous and dynamic, so it becomes imperative to efficiently manage this environment without neglecting customer satisfaction [1].

To meet this challenge and ensure full service to the needs of the business, companies begin to introduce the concepts of IT Service Management (ITSM). Thus, by proposing the idea of process-based operation, the ITSM has become one of the main drivers for business success [2]. The use of standards in IT infrastructure allows this alignment to occur providing less complexity and higher quality to the services [13].

In this context, the research question is defined: "is the IT change management process of the Brazilian Court of Justice proper with the best practices recommended by the literature?"

This research aims to compare the current process of IT change management of the Court with the best practices suggested by the reference authors consulted in the literature review, proposing improvements based on identified opportunities and threats.

To enable this study, this research was structured in five sections. Section II presents the theoretical framework used for service management and change management. Section III will present the research methodology applied to the research project. In section IV the data will be analyzed and the result presented. And, finally, the final considerations will be presented in section V.

## II. THEORETICAL REFERENCE

### A. IT SERVICE MANAGEMENT

IT Service Management (ITSM) enables the extraction of valuable business perspectives that lead to greater efficiency and quality of services delivered by Information Technology to the organization [3].

The literature review by [1] and [4] identified several frameworks that address standards for IT Service Management, including: Information Technology Infrastructure Library (ITIL), Control Objectives for Information and Related Technology (COBIT), ISO/IEC 20000, Microsoft Operations Framework (MOF) and IBM Tivoli Unified Process (ITUP). Among them, it is possible to note the prominence of the ITIL

framework as a reference for the internalization of ITSM best practices in companies [1] [4] [5] [6].

In the study conducted by [5], the authors conceptualize the Information Technology Infrastructure Library (ITIL) as a process-based methodology that provides best practices for IT Service Management (ITSM) and helps organizations align IT to the needs of the business, promoting quality of service and reducing, in the long term, the costs to provision of IT services.

The quality of service delivery is directly linked to the practices of IT Service Management. Considering the need to create and monitor key performance indicators (KPIs) were studied by [12].

Between the years of 2007 and 2008, ITIL version 3 was launched, which is composed of five books and organized in the following service life cycles: Service Strategy, Service Design, Service Transition, Service Operations and Continual Service Improvement. Each of the five major publications covers a life-cycle stage of the service [6]. Figure 1 presents the organization of knowledge regarding the new version.



Figura 1. ITIL Service Life Cycle [7]

In the book Service Transition is, among others, the process of Change Management, whose implementation will be object of this study case [7].

## B. CHANGE MANAGEMENT

Changes are inherent to business and necessary for the survival of organizations in general. In a competitive business environment, changes occur rapidly, in different ways, scales and quantities [8]. Organizations change to cope with increasing competitiveness, comply with laws or regulations and introduce new technologies [9].

In fact, the triggering triggers of the transformation need are of the most varied types, which can, however, be grouped as: external causes, organizational causes or causes internal to the project [14]. As an example, the fragile definition of needs and specifications at the beginning of the project is motivated by: (i) the reduction of costs in the specification stage (internal cause to the project); (ii) by changes in the business need (organizational cause); or (iii) by the legal or regulatory imposition of new rules (external cause).

Because of the devastating potential that a negative impact of a change can impose on organizational goals or the success of a project, effective change management is a premise for companies to continue to exist [8]. It is unquestionable, therefore, the importance of this theme, which has awakened in the researchers interest in deepening in the subject and to create models that help the organizations to identify and treat the changes in a controlled way, minimizing risks and optimizing resources aiming to reach the objective of the organization or project.

The success of an organizational transformation or implementation of changes is not, however, only based on a systematized process. Critical factors for successful change management include: (i) leadership, (ii) teamwork and collaboration, (iii) sharing of vision and responsibility, (iv) communication, (v) sense of urgency [7] [11].

According [11], which was highlighted in the literature review because it has been cited in several recent studies analyzed, the following are success factors of the transformation process:

- i - The implementation of change implies altering procedures, modifying culture and transforming the previous status so that it becomes necessary to be led by those who clearly see the need for such changes. Therefore, the change should be sponsored by the responsible authority.
- ii - To ensure the team's commitment to change, they must be clearly informed about business directions and needs and be involved in decision-making processes.
- iii - The existence of a vision helps to define the direction in which the organization needs to move. Without it, transformation efforts can dissolve into a range of projects that take the organization to the wrong place or nowhere.
- iv - Employees will not strive to promote the necessary change in the organization unless they believe it is useful and possible. This belief will only be achieved through a massive communication process. Nothing weakens the change more than the contradictory behavior to the words of important individuals in the process of transformation.
- v - The sense of urgency may be motivated by an impending crisis or a new opportunity and should awaken in those involved the feeling that the current status is more dangerous than throwing itself into the unknown. When the sense of urgency for transformation is not high enough, the transformation process will not be successful and the long-term future of the organization is compromised.

The concepts and success factors presented apply to the management of change in the broad sense, without focusing on a particular scenario or branch of action. In spite of having been found in the review of the literature publications that have deepened in the theme oriented to the constructive industry, with regard to IT, academic studies are scarce that approach the management of changes as a central focus [14].

## C. IT SERVICES CHANGE MANAGEMENT

Setting standards brings great benefits to organizations. However, as we have seen, a effective management of changes demands that there be commitment from both the com-

pany’s employees and the leaders of the affected areas. The challenge of motivating employees was studied by [10] who proposed the adoption of dynamics in an IT Service Desk capable of increasing the quality of the operator’s performance.

The present work studied the adequacy of the process of change management in IT implanted in a Court of Justice comparing with the recommendations of success factors presented. For this purpose, the methodology used was described in section III.

**III. METHODOLOGY**

This applied research was developed through a bibliographical research related to ITIL best practices, especially with regard to Change Management. For data collection, documents and observation were used in the organization under study. The data collected were analyzed through a qualitative approach in order to compare them with the related literature review.

After a bibliographic review, a comparative analysis was performed between the best practices indicated by ITIL and related literatures and the current status of the organization, in order to identify if the implementation of change management activities are in line with the best practices proposed by the model. Figure 3 shows the structuring of the research carried out in this study.

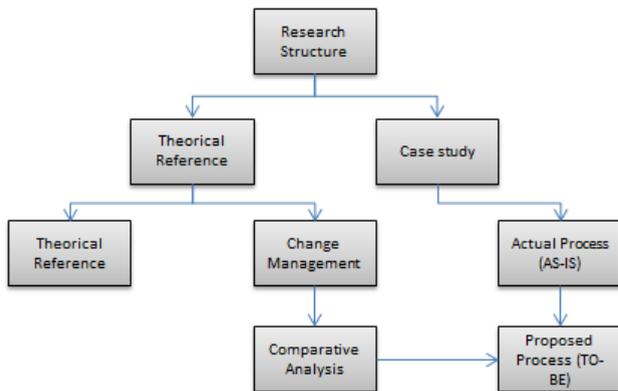


Figura 2. Research Structure

Finally, adjustments were suggested in the current process in order to increase the adherence to the best practices studied.

**IV. DATA PRESENTATION**

Currently, the court is in the process of internalizing the good practices of IT Service Management. And one of the processes which needs to mature is Change Management. Currently this management process has been carried out in an incipient way, through a process that lacks documentation and maturation. Figure 4 shows the current Change Management process in use at the institution (AS-IS process).

The current service flow begins with the Request of Change (RFC) record in the call control system.

If the request is for out-of-hours service, the Change Committee is advised and RFC is scheduled. If the request

Service Available Time (SAT)	
Goal:	Minimize downtime
Quality criteria:	Credibility and Competence
Service level:	SAT >99.5%
Method:	SAT = (Minutes in which service became available / Minutes total) * 100

Tabela 1. SAT Indicator

is for attendance during office hours, the Change Committee must evaluate / authorize it.

In the case of emergency changes, the decision of the Committee is simplified to the approval of the managing member of the applicant unit. Not being a emergency change, all the members of the Committee must carry out the approval. Once approved the Request of Change (RFC), the service team verifies if the request involves the activation of the contracted company - whose performance requires authorization through the system - or can be served by employees of the Court, directing the service to the executor.

The executor performs the service, updates the system with information collected during the performance and closes the Request of Change (RFC). Understanding the current process, the strengths, weaknesses, opportunities, and threats of the current process were evaluated using the SWOT matrix. Figure 5 shows the result of this analysis.

The scenario analysis pointed to weaknesses in communication during the process and publicity of the results since the current process does not provide the use of an institutional tool. In addition, it was identified the lack of "Authorize change build and test" and "Coordinate change build and teste" stages of the change since there is currently no prior consultation for approval the change planning (the RFC is made with the job ready to be applied). Finally, it has been identified that the service of Requests for Change (RFC) outside office hours is not subject to a formal authorization process and that there is no default planning for a recovery flow in the event of an error.

After evaluating the activities of the ITIL change management process and the success factors for change management, the proposed future process is presented in Figure 6.

The proposal presents improvements mainly with respect to communications, review flow, change planning and testing, approval of off-hours changes and the recovery flow (in case of failures).

However, in addition to improving the process for adopting best practices, it is observed the need for effective monitoring of the process to verify that it is being executed with quality.

In this way, it is proposed the gradual adoption of quality and performance indicators related to the Requests for Change (RFC), with monthly evaluation of the results, as presented in tables 1, 2, 3 and 4.

With the monthly evaluation of the “Service Available Time” (SAT), table 1, it will be possible to define minimum quality metrics to be sought in order to make the services available to the target audience as long as possible.

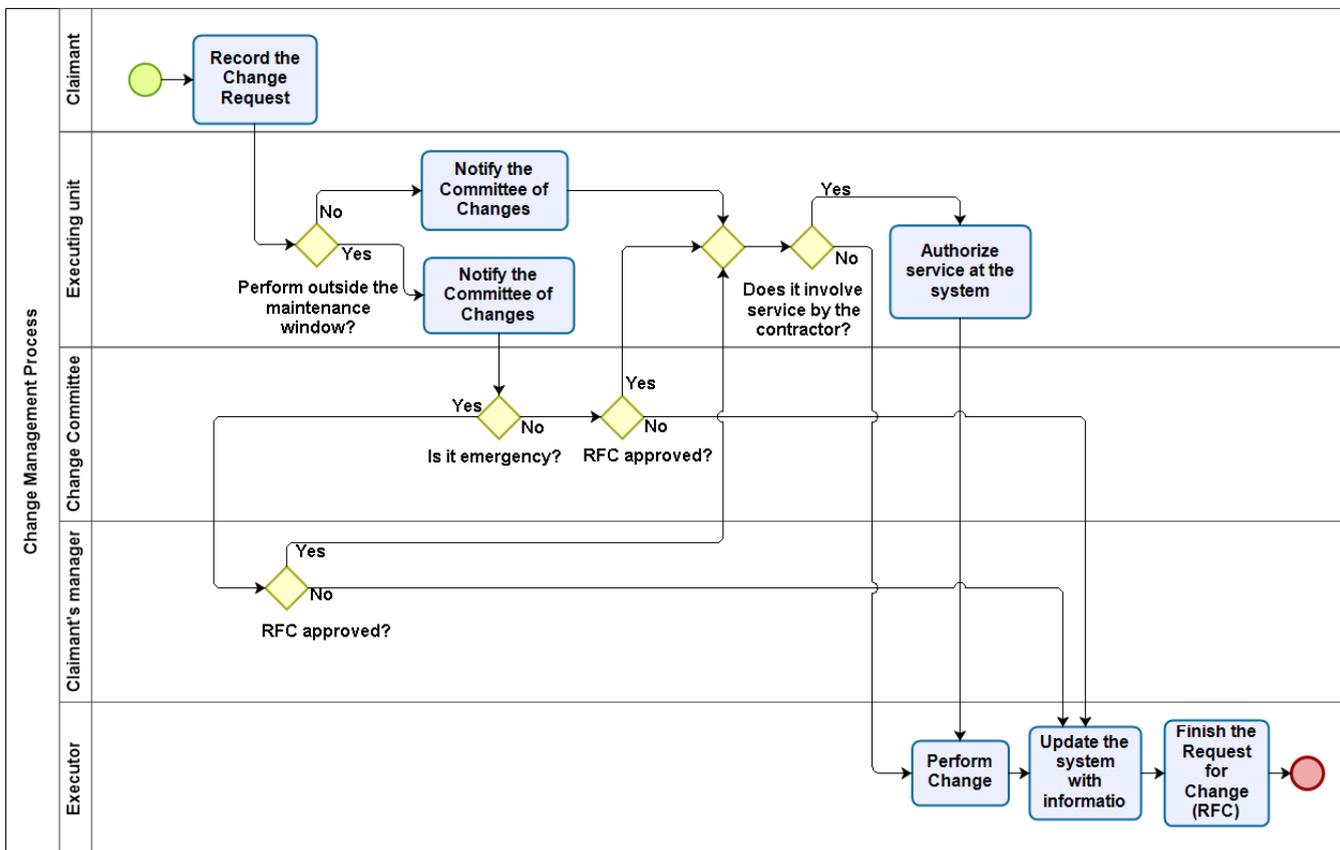


Figura 3. Current status of the process ("AS-IS")

SWOT ANALYSIS		
INTERNAL ENVIRONMENT	<b>FORCES</b>	<b>OPPORTUNITIES</b>
	Team Commitment	RFC Review
	Sense of urgency	RFC Planning
	Proper RFC Record	
	<b>WEAKNESSES</b>	<b>THREATS</b>
	Communication during the process	Sharing of vision
	Advertising of results	Leadership involved
	Construction and testing approval	
	Construction and testing perform	
	RFC Approval	
Recovery in case of error		

Figura 4. SWOT Analysis of the actual process

For "Percentage of Appointments within the Term", table 2, are considered to be of low complexity those Requests for Change (RFC) that refer to documented procedures, which are commonly repeated and which require the performance of only one executor (such as requests for release in production of update in the codes of the web systems - deploy - without updating in the database).

The RFC classified with medium complexity refers to the requests that, although are well documented, require the performance of more than one executor with different knowledge, as an instance, a deploy requests that, in addition to updating the application source code, require updating the

Percentage of Appointments within the Term (PAT)	
Goal:	Minimize time to meet Request of Change (RFC)
Quality criteria:	Service speed
Service level:	PAT >95%, considering: <ul style="list-style-type: none"> <li>• Low Complexity: &lt;6h</li> <li>• Average Complexity: &lt;12h</li> <li>• High Complexity: &lt;24h</li> </ul>
Method:	$PAT = (\text{Number of RFCs met on time} / \text{Total of RFCs performed}) * 100$

Tabela 2. PAT Indicator

database used by it. Finally, a highly complex RFC is one in which there is a need to create a multidisciplinary team involving more than one sector, including the demanding sector and the potential affected, in order to ensure that the change has the desired effect with no consequences (such as updating the version of packages used by a Linux server that hosts one or more critical systems in the organization).

The monitoring of this indicator assists in the definition of action plans and decision making that make possible to assure the fulfillment in the term.

The accompanying "Percentage of Rejected RFCs (PRR)", table 3 facilitates the maturation of the demand proposition by allowing the analysis of the quality of RFCs received, avoiding the rework and minimizing risks that arise from

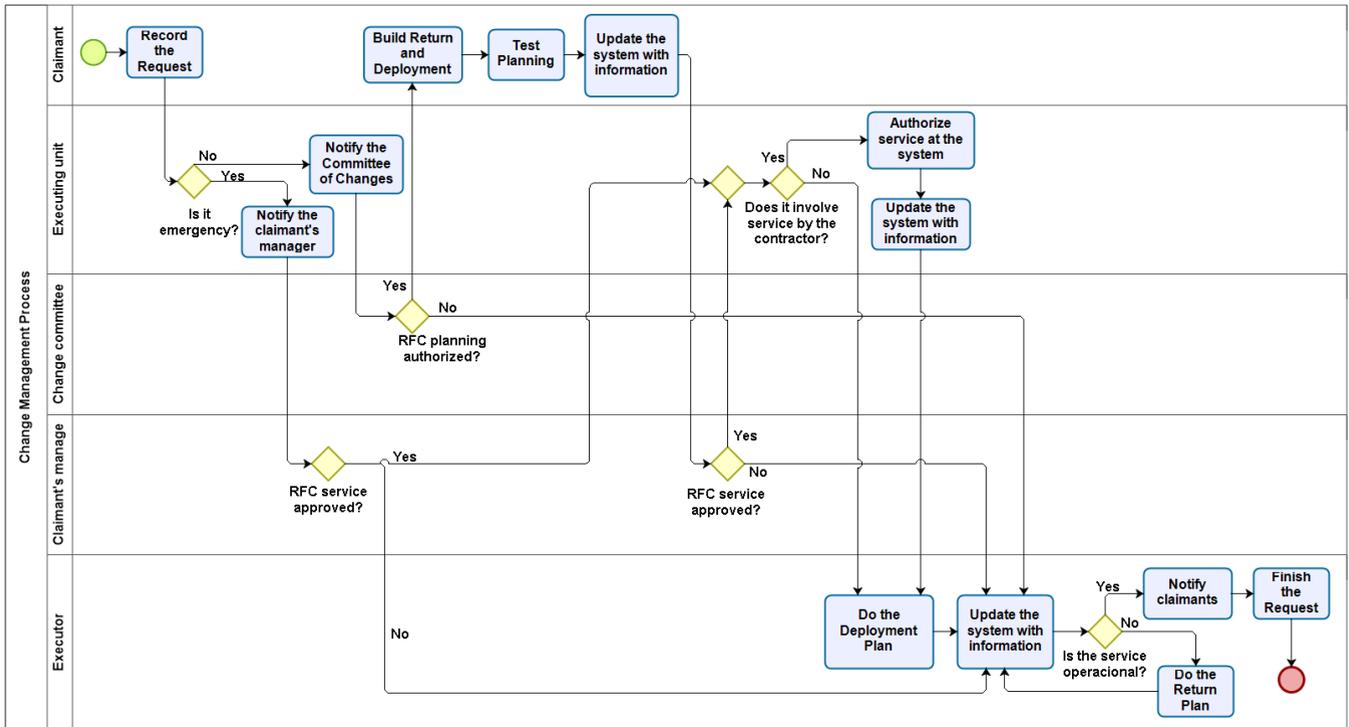


Figura 5. Future state of the process ("TO-BE")

Percentage of Rejected RFCs (PRR)	
Goal:	Increase order quality by minimizing rejection of Request for Changes (RFCs)
Quality criteria:	Flexibility and Efficiency
Service level:	PRR <5%
Method:	$PRR = (\text{Number of RFCs rejected} / \text{Total RFCs requested}) * 100$

Tabela 3. PRR Indicator

poorly formulated requests.

Percentage of Documented RFCs (PDR)	
Goal:	Increase the knowledge base
Quality criteria:	Service speed and Efficiency
Service level:	PDR >99.5%
Method:	$PDR = (\text{Number of RFCs documented} / \text{Total RFCs performed}) * 100$

Tabela 4. PDR Indicator

Finally, the monitoring of the "Percentage of Documented RFCs (PDR)", table 4, will enable the organization to disseminate knowledge and give speed and quality to the future attendance of similar problems.

Being operational, those indicators may contribute to the maturity of the change management process and the availability of the services offered to the users of the court.

## V. CONSIDERATIONS

As result of the research, it was possible to review the bibliography referring to best practices in Change Management

in IT Services and, based on the in-depth study of ITIL best practices and success factors for change, it was possible to analyze critically the current process of change management of the court studied.

As a consequence of strengths, weaknesses, opportunities and threats analysis of the current process it was possible to propose improvements in the way of: (i) propose a improvement in the current process; and (ii) suggestions for the application of quality and performance indicators for IT Service Change Management focused on availability of services.

The implementation of these improvements, the systematic monitoring of the changes, the data collection to feed performance indicators based on the proposed metrics are objects of a future work that will, in practice, compare the productivity gain and quality in the management of the changes of the systems, whether software or infrastructure, in the IT of the court, object of study of this work.

## Referências

- [1] N. Jamous, S. Bosse, C. Görling, J. Hintsch, A. Khan, F. Kramer, H. Müller and K. Turowski, "Towards an IT Service Lifecycle Management (ITSLM) Concept", in 4th International Conference on Enterprise Systems, 2016, p. 29-38.
- [2] S. Guo, F. Qi, Y. Yuan, X. Gong and Z. Zhan, "The design and implementation of process engine supporting absolute customization based on ITSM", in International Conference on Advanced Intelligence and Awareness Internet (AIAI), 2010, p. 420-423.
- [3] Y. Diao, E. Jan, Y. Li, D. Rosu and A. Sailer, "Service analytics for IT service management", in IBM J. Res. & Dev., v. 60, n. 23, marmay 2016, pp. 13:1-13:17.
- [4] F. B. e Abreu, J. C. Costa, J. M. Freitas and R. B. V. da Porciúncula, "Definition and Validation of Metrics for ITSM Process Models", in 7th

International Conference on the Quality of Information and Communications Technology, 2010, p. 79-88, 2010.

- [5] Z. Yao and X. Wang, "An ITIL Based ITSM Practice: A Case Study of Steel Manufacturing Enterprise", in 7th International Conference on Service Systems and Service Management (ICSSSM), 2010, p. 1-5.
- [6] S. L. Corrêa, M. P. Méxas, G. M. Drumond and M. J. Meiriño, "Cost Elements Identification for Maintenance and Support of ERP Systems in Brazilian IFES: An Approach Based on TCO and ITIL", in IEEE Latin America Transactions, v. 14, n. 5, mai / 2016.
- [7] Office of Government Commerce (OGC), "ITIL Service Transition", in The Stationery Office (TSO), London, 2011. Z. Wang, B. T. H. Lim and I. Karmadeen, "Change Management Research in Construction: A Critical Review", in Proceedings of the 19th International CIB World Building Congress: Construction and Society, 2013, pp. 1-14.
- [8] J. M. C. Hernandez and M. P. Caldas, "Resistência à mudança: uma revisão crítica", in Revista de Administração de Empresas, v. 41, n. 2, 2001, pp. 31-45.
- [9] Da Conceicao, F.S., da Silva, A.P., de Oliveira Filho, A.Q. and Silva Filho, R.C., 2014, September. Toward a gamification model to improve IT service management quality on service desk. In Quality of Information and Communications Technology (QUATIC), 2014 9th International Conference on the (pp. 255-260). IEEE.
- [10] J. P. Kotter. "Leading change: why transformation efforts fail", in Harvard Business Review, Boston, v. 73, n. 2, mar/1995, p. 59-67.
- [11] Loumos, V., Christonakis, G., Mpardis, G. and Tziouva, P., 2010, April. Change Management and Quality of Service through Business Process Modeling: The N-VIS, a Public Sector Project. In Information Technology: New Generations (ITNG), 2010 Seventh International Conference on (pp. 1300-1303). IEEE.
- [12] Da Silva, L.F. and e Abreu, F.B., 2010, September. An IT infrastructure patterns approach to improve IT service management quality. In Quality of Information and Communications Technology (QUATIC), 2010 Seventh International Conference on the (pp. 171-176). IEEE.
- [13] J. Bröchner and U. Badenfelt, "Changes and change management in construction and IT projects", in Automation in Construction, v. 20, n. 7, 2011, p. 767-775.
- [14] Panjwani, M., Jäntti, M. and Sormunen, J., 2016, September. IT Service Management from a Perspective of Small and Medium Sized Companies. In Quality of Information and Communications Technology (QUATIC), 2016 10th International Conference on the (pp. 210-215). IEEE.



**CARLOS EDUARDO MACHADO PIRES** He holds a degree in Computer Science from Universidade Católica de Brasília (2006), holds a law degree from Universidade Católica de Brasília (2013), and a postgraduate degree in IT Governance from the University Center of Maranhão (2008). He is currently a Master's student in Applied Computing at the University of Brasília (UnB) and Geoprocessing Manager at the Environmental Company of the Federal District - CAESB. Has an interest in Data Science and Geographic Information Systems - GIS. He has experience in the area of Computer Science, with emphasis on Information Systems, working mainly in the following subjects: Project Management, Database and Corporate GIS.



**SIMONE BORGES SIMÃO MONTEIRO** Has a degree in Chemical Engineering from the Federal University of Uberlândia (1995), a Master's Degree in Production Engineering (Management of Agroindustrial Systems) from the Federal University of São Carlos (1998), Specialization in Lean Production Systems at USP (2004) and a PhD in Engineering (Quality Management) by the Federal University of São Carlos (2006). She was a professor at the University of Franca (UNIFRAN)

from 2000 to 2010. Now, she is Adjunct Professor III of the Department of Production Engineering at the University of Brasília (UnB) and Coordinator of Projects at CEFTRU - Interdisciplinary Center for Transportation Studies. She is a Professor and Researcher in Master of Applied Computing program at the Risk Management Research Line and has experience in the area of Production Engineering, with emphasis on Quality Control Assurance, Quality Management, Project Management and Process Improvement and Analysis.



**ANA CARLA B. REIS** Has a degree in Electrical Engineering from the Federal University of Pernambuco (2005), a master's degree (2007) and a PhD (2011) in Production Engineering from the Federal University of Pernambuco. Currently an Adjunct Professor in the Department of Production Engineering at the University of Brasília (UnB) and in the Master of Applied Computation program at the Risk Management Research Line. Has experience in the area of Production Engineering, with emphasis on Operational Research, working mainly on the following topics: Decision Models, Multicriteria Decision, Production Control and Production and Process Improvement.

...



**EDUARDO DA SILVA SOUSA** He holds a degree in Computer Science from Universidade Católica de Brasília - UCB (2005), an MBA specialist in IT Governance from Universidade Católica de Brasília - UCB (2007), a specialist in IT Management at Belo Horizonte Metropolitan College - FMBH (2008), a specialist in Higher Education Teaching by Unyleya (2017). He is currently a student of a Master's Program in Applied Computing at the University of Brasília (UnB) and Security

Manager/System Analyst at the Federal District and Territories Court - TJDF, currently acting as Information Security Manager.

# Desenvolvimento do Módulo de Pre-processamento da ferramenta SentimentALL

PARCILENE FERNANDES BRITO e  
LUAN GOMES DE ALMEIDA ARAÚJO.

Centro Universitário Luterano de Palmas, Teotônio Segurado 1501 Sul Palmas - TO CEP 77.019-900 Caixa Postal nº 85

(e-mail: {parcilene, luangomes2121}@gmail.com)

Autor Correspondente: Parcilene Fernandes Brito (e-mail: parcilene@gmail.com).

• **RESUMO** - Os sistemas que utilizam em seu processo de análise a linguagem natural (por exemplo, sistemas que analisam comentários de sites, reviews de filmes, posts em redes sociais etc.) geralmente usam técnicas relacionadas ao Processamento de Linguagem Natural (PLN). Neste artigo, serão apresentadas as etapas de PLN utilizadas no desenvolvimento do Módulo de Pré-Processamento da Ferramenta SentimentALL, que tem como propósito realizar análises de textos que emitem opiniões sobre produtos. O módulo foi desenvolvido para ser usado em qualquer contexto, mas neste trabalho foi utilizado na análise dos comentários extraídos do site de turismo TripAdvisor. O módulo é composto das etapas de normalização, correção ortográfica, pos-tagging e identificação de expressões compostas. Para a sua implementação, foi utilizada a biblioteca NLTK, da linguagem de Programação Python.

• **PALAVRAS-CHAVE** - Processamento de Linguagem Natural, SentimentALL, Pos-Tagging, NLTK.

## I. INTRODUÇÃO

Processamento de Linguagem Natural (PLN), ou em inglês Natural Language Processing (NLP), é uma subárea da linguística computacional. Segundo [8] a linguística computacional é uma “área de pesquisa que se originou da interseção entre a linguística e a uma subárea da ciência da computação, a inteligência artificial”. A PLN trata do uso de técnicas computacionais para que computadores possam “entender” e gerar linguagem natural, a fim de que haja uma melhor interação entre homem e computador, ou para que o computador possa adquirir informações que estão em linguagem natural ([8]; [17]).

A maioria das informações presentes na internet está na forma não estruturada e em linguagem natural. Portanto, para a resolução de determinados problemas, é necessária a utilização de técnicas de PLN para extrair tais informações, transformando-as para a forma estruturada. Dados estruturados são aqueles que possuem uma identificação do que eles representam e que podem ser organizados em linhas e colunas, tornando fácil a sua recuperação. Geralmente os dados estruturados são armazenados em banco de dados relacionais. Já os não estruturados, são dados que não possuem uma estrutura definida (como documentos textuais, e-mails e vídeos), são mais difíceis de serem trabalhados, logo a

tarefa para recuperá-los não é feita de forma trivial como, por exemplo, em uma consulta em SQL, necessitando do uso, por exemplo, de PLN.

Linguagens naturais, que estão no formato não estruturado, são linguagens usadas por seres humanos para se comunicarem, como, por exemplo, inglês, português, chinês e francês. No entanto, essas linguagens naturais, considerando o seu tamanho e que estão em constante mudança, podem se tornar ambíguas. A ambiguidade é evidenciada quando um enunciado possui dois ou mais sentidos, o que torna difícil sua compreensão, especialmente para a máquina, pois o computador diante de enunciados ambíguos tem que decidir qual o sentido correto, e essa decisão nem sempre produz bons resultados. Por exemplo: “Pedro entregou a Maria o seu lápis”. De quem é o lápis? De Pedro ou de Maria? Essa frase já exemplifica que mesmo em uma análise manual a identificação correta dos elementos de uma frase é uma tarefa complexa, às vezes até impossível dada a sua estrutura. Então, um sistema que trabalhe com PLN tem que estar preparado para tentar resolver esse tipo de problema.

O módulo apresentado neste artigo é parte do projeto SentimentALL de [20] e está vinculado diretamente à implementação da segunda versão da ferramenta SentimentALL, desenvolvida em [2] e [19]. Neste projeto, foram desenvol-

vidos três módulos: coleta de dados, pré-processamento e análise de sentimentos. Na coleta de dados foram extraídos mais de seis milhões de comentários das páginas do site de turismo TripAdvisor. Esses dados foram pré-processados utilizando as técnicas de PLN, foco deste trabalho, que serão apresentadas nas próximas seções. O pré-processamento consistiu em realizar a normalização dos dados, correção ortográfica, identificação da classe morfológica das palavras e identificação de expressões multipalavras.

## II. ETAPAS DE ANÁLISE EM PLN

As análises tradicionais (sintaxe, semântica e pragmática) servem, no máximo, como um ponto de partida para o processamento de textos reais em linguagem natural ([7]). Então, quando se lida com dados reais, os estágios são decompostos como na Figura 1.

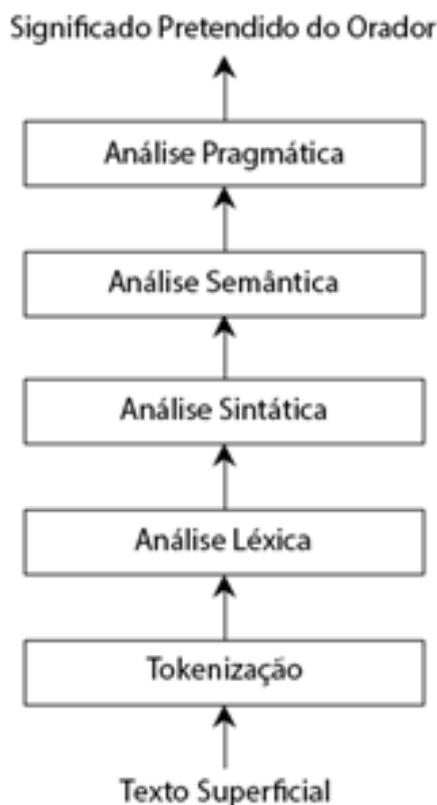


Figura 1. Os estágios de análise em PLN. [7] (traduzido)

Como pode ser visto na Figura 1, o processo inicia recebendo como entrada um texto, passa pelo estágio de tokenização, depois por diferentes granularidades de análise e, por fim, tem como saída o significado pretendido do orador. A seguir, esses estágios serão apresentados com mais detalhes.

### 1) Tokenização

Segundo [16], a tokenização quebra uma sequência de caracteres de um texto em palavras (tokens), localizando os limites das palavras, os pontos onde uma palavra termina e outra

começa. Há duas abordagens de tokenização: para idiomas delimitados por espaços e para línguas não-segmentadas. As linguagens delimitadas por espaços são aquelas em que alguns limites entre as palavras são indicados pela inserção de espaços em branco, como por exemplo, inglês, português e espanhol. Nas linguagens não-segmentadas, as palavras são escritas em sucessão, sem indicação de limites entre as palavras, por exemplo, tailandês e chinês.

[16] explica que podem existir ambiguidades no uso de sinais de pontuação que afetam a tokenização, uma vez que os sinais de pontuação podem ter muitas funções diferentes. O Exemplo 1 apresenta uma frase que contém ambiguidade.

Exemplo 1: “João, pai de Maria, comprou um pacote de bolacha de R\$1,25 para ela.”

No Exemplo 1, tem-se então uma ambiguidade no uso da vírgula. As duas primeiras vírgulas são usadas para delimitar o aposto e a terceira vírgula para separar a parte inteira da parte decimal de um número decimal. Na tokenização deve-se conseguir determinar quando uma pontuação faz parte ou não de um token.

### 2) Análise Léxica

A análise léxica ou morfológica é uma análise realizada a nível de palavra. As palavras que foram separadas na etapa anterior, agora são decompostas em partes (morfemas). Segundo [5], a análise léxica “examina os modos pelos quais palavras se desmembram em componentes e como isso afeta o status gramatical delas”. Como por exemplo, quando uma palavra tem “s” no final, costuma estar no plural. Nessa etapa é indicado em qual classe uma palavra se encaixa, se é um adjetivo, advérbio, verbo, entre outras.

Para exemplificar a análise léxica, será utilizada a palavra “infelizmente”. Essa palavra pode ser decomposta em 3 partes: prefixo “in”, radical “feliz” e sufixo “mente”. Essas são as menores partes (morfemas) que se pode extrair dessa palavra, não podendo ser mais fragmentada. Cada um dos morfemas tem um significado gramatical, podendo ou não modificar a classe gramatical da palavra. O radical “feliz” é um adjetivo que expressa alegria. O prefixo “in” faz a negação do radical, mas não modifica a classe gramatical da palavra. O sufixo “mente” modifica o radical, transformando o adjetivo em advérbio. Com isso, a palavra “infelizmente” é da classe gramatical advérbio.

### 3) Análise Sintática

A análise sintática verifica se a estrutura gramatical de uma frase está correta. Segundo [11], para realizar a análise sintática pode-se utilizar a gramática livre de contexto, que é representada pelo seguinte formalismo:  $G = \{\Sigma, N, S, R\}$ . Onde,  $\Sigma$  é um conjunto de símbolos terminais, palavras (Ex.: {pessoa,cachorro,ganhar}).  $N$  é um conjunto de símbolos não terminais, podendo ser categorias de palavras, frases ou segmentos de frases (Ex.: {frases nominais,verbos,artigos}).  $S$  é um símbolo inicial não terminal.  $R$  é um conjunto de regras de produção. Uma regra de produção possui a seguinte forma:  $A \rightarrow \alpha$ , sendo  $A$  um símbolo não terminal e  $\alpha$



trabalhar com PLN. Apesar de que às vezes não é necessário realizar todas as análises para resolver um problema.

### A. POS TAG

PoS tagging é uma tarefa que ocorre na etapa de análise léxica em PLN. Segundo [10], as palavras são agrupadas em classes chamadas de Part-of-Speech (PoS), por exemplo: verbos, substantivos, adjetivos, pronomes, e assim por diante. Um conjunto de PoS tag se chama tagset. PoS Tagging é um importante passo que ocorre no pré-processamento em algumas aplicações de PLN. PoS Tagging envolve, dada uma sentença, a seleção da sequência mais provável de PoS tags de um tagset, para cada palavra da sentença ([1]; [9]).

[9] diz que há duas dificuldades básicas em PoS tagging: palavras ambíguas e palavras desconhecidas. A ambiguidade ocorre quando uma palavra possui mais de um PoS tag possível, como por exemplo, “para”, que pode ser um verbo (parar) ou uma preposição. Para tentar resolver esse problema deve-se olhar o contexto, a estrutura sintática na qual a palavra está inserida. No caso do problema relacionado a palavras desconhecidas, [9] cita duas abordagens em que esse problema pode ocorrer: quando se usam regras escritas manualmente, com isso, algumas palavras podem não ter sido consideradas, ou quando se usa sistemas estatísticos e a palavra não aparece no conjunto de treinamento.

A seguir um exemplo para ilustrar como as palavras seriam marcadas com as POS tags.

Exemplo 3: “O menino não para de jogar vídeo game.”

A Tabela 1 representa um tagset simplificado disponível para identificar em qual classe uma palavra na frase se encaixa.

Tabela 1. Exemplo de Tagset

Tag	Descrição
AT	Artigo
ADJ	Adjetivo
ADV	Advérbio
PP	Preposição
PR	Pronome
SB	Substativo

Usando o tagset (Tabela 1), pode-se atribuir as tags em cada palavra, analisando a classe gramatical de cada palavra e a estrutura sintática da frase. O resultado fica como na Figura 6:

AT SB ADV VB PP VB SB SB  
*O menino não para de jogar vídeo game*

Figura 6. Exemplo de PoS tagging

O ideal é que se tenha mais granularidade nas tags, criando subclasses de palavras, por exemplo, os verbos poderiam ser separados em presente, passado e futuro.

A próxima seção apresenta a tarefa de identificação de expressões multipalavras, que pode ser realizada com a ajuda das palavras etiquetadas com PoS tags.

### B. EXPRESSÕES MULTIPALAVRAS (MWE)

Expressões Multipalavras, ou em inglês Multiwords Expression (MWE), são, segundo [18], palavras combinadas em que o significado da expressão não pode ser obtido a partir de suas partes. Ou seja, expressões multipalavras são conjuntos de palavras que se comportam como uma unidade, tendo um valor sintático e semântico diferente daquele caso sejam consideradas as palavras individualmente. A seguir um exemplo de expressão multipalavra.

Exemplo 4: “João comeu um cachorro quente.”

No Exemplo 4, “cachorro quente” é uma expressão multipalavra, pois ela se comporta como se fosse uma palavra. Se consideradas as palavras separadamente, então, seria um “cachorro” que está “quente”. Nenhuma das duas alternativas está errada, considerando o aspecto sintático e semântico, embora a probabilidade de ser uma expressão multipalavra é maior. Para ter conhecimento se um conjunto de palavras tem maior probabilidade de serem expressões ou não, pode-se utilizar a medida de associação entre palavras Pointwise Mutual Information (PMI). O cálculo de PMI, proposto por [4], é o seguinte.

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

Onde x e y são palavras. P(x,y) é a probabilidade de ocorrência de x e y ao mesmo tempo. P(x) e P(y) é a probabilidade de ocorrência independentemente de x e y, respectivamente. Com isso, tem-se o grau de dependência entre duas palavras. Quanto maior o resultado, maior a dependência.

Exemplificando o uso dessa equação, tem-se na Tabela 2 dados fictícios de um determinado contexto, as palavras e sua probabilidade de ocorrência, ou seja, qual a frequência em que ela pode aparecer em um documento.

Tabela 2. Exemplo de probabilidade de ocorrência de palavras

Palavra	Probabilidade de Ocorrência
Cachorro	23%
Raça	20%
Quente	11%
Ração	12%
Casinha	15%
Coleira	19%

Na Tabela 3, tem-se a probabilidade de ocorrência de possíveis expressões, ou seja, foram calculadas as probabilidades da ocorrência de duas palavras juntas iniciada com a palavra “Cachorro”.

Tabela 3. Exemplo de probabilidade de ocorrência de expressões

Possíveis Expressões	Probabilidade de Ocorrência
Cachorro Raça	3%
Cachorro Quente	41%
Cachorro Ração	2%
Cachorro Casinha	7%
Cachorro Coleira	5%

Utilizando os dados da Tabela 2 e da Tabela 3, a expressão “cachorro quente”, tem a probabilidade de ocorrência de

41%, a palavra “cachorro” de 23% e “quente” de 11%. Fazendo as substituições dos valores na equação:

$$\begin{aligned} I(\text{cachorro}, \text{quente}) &= \log_2 \frac{P(\text{cachorro}, \text{quente})}{P(\text{cachorro})P(\text{quente})} \\ &= \log_2 \frac{41\%}{23\% * 11\%} \\ &= 4,02 \end{aligned}$$

O problema dessa equação está no fato de poder calcular a dependência de apenas duas variáveis. Essa equação pode ser estendida para realizar o cálculo com mais variáveis através da equação proposta por [6].

$$SI(x_1, x_2, \dots, x_n) = \log_2 \frac{P(x_1, x_2, \dots, x_n)}{\prod_{i=1}^n P(x_i)} \quad (2)$$

A ideia dessa equação continua a mesma da anterior. A mudança está no cálculo da probabilidade de ocorrência de palavras em conjunto, em que se considera agora mais de duas palavras ( $P(x_1, x_2, \dots, x_n)$ ) e é realizado o produto da probabilidade individual de todas as palavras ( $\prod_{i=1}^n P(x_i)$ ).

O cálculo de PMI é realizado sobre um corpus. Nesse corpus é realizado o PoS tagging e depois, são identificados padrões morfológicos. [3], apresenta seis padrões morfológicos que podem ser utilizados para identificar possíveis expressões multipalavras do tipo composto nominal (que será utilizada nesse trabalho). A Tabela 4 apresenta esses padrões, onde as letras N (Substantivo), A (Adjetivo) e P (Preposição) são PoS tags.

Tabela 4. Padrões Morfológicos. [3]

Padrão	Exemplo
NN	Nações Unidas
NA	Governo Federal
NNN	Supremo Tribunal Federal
NNA	Fundo Monetário Internacional
NAA	Produto Interno Bruto
NP N	Casa de praia, bolsa de valores

Identificada as possíveis expressões multipalavras utilizando esses padrões morfológicos, em seguida é aplicada o PMI. Por fim, um conjunto de palavras é considerado uma expressão multipalavra quando o valor do PMI for maior ou igual a um valor definido pelo usuário.

As etapas de análise e as técnicas de PoS tagging, Expressões Multipalavras ou outras, podem ser aplicadas em algumas aplicações que realizam PLN, tais como, recuperação de informação, extração de informação e análise de sentimentos.

### III. METODOLOGIA

O módulo desenvolvido neste trabalho utilizou os dados obtidos na ferramenta SentimentALL, extraídos do site TripAdvisor. Especificamente, foram usados comentários feitos pelos usuários do site, em um total de 6.438.497 comentários dos 100 destinos turísticos brasileiros mais avaliados no site entre início de fevereiro e final de março de 2017.

Para o desenvolvimento do módulo foi utilizada a linguagem Python, codificada por meio da Integrated Development Environment (IDE) PyCharm da Jet Brains. Foi utilizada a biblioteca NLTK para processamento de linguagem natural. Essa biblioteca contém por volta de 35 módulos, cada um com diversos submódulos, que realizam múltiplas tarefas de PLN, como tokenização, PoS tagging etc. Para realização do PLN, desenvolveu-se o módulo de pré-processamento. Nele foi feita a normalização do texto de entrada (usando expressões regulares) e a correção ortográfica, utilizando o módulo PoS tagging e implementado o algoritmo de PMI para a definição de expressões compostas. Esse módulo teve como entrada os dados coletados (comentários do TripAdvisor) e como saída os dados normalizados, corrigidos ortograficamente, etiquetados e as possíveis expressões multipalavras identificadas.

## IV. RESULTADOS

### A. NORMALIZAÇÃO

A normalização textual é uma etapa do pré-processamento que realiza a padronização dos dados, removendo ou substituindo caracteres ou palavras. A identificação dos itens a serem normalizados é realizada através da utilização de expressões regulares.

Tabela 5. Exemplo de normalização textual

Texto de Entrada	Texto Normalizado
A entrada estava R\$ 20 por cabeça quando fui em junho/12. O preço é caro mas vale a pena. As cachoeiras são d+, deliciosas p/ se curtir. Muito legal, adooooorei !!! :)	A entrada estava valor por cabeça quando fui em data. O preço é caro mas gostei. As cachoeiras são demais, deliciosas para se curtir. Muito legal, adorei! Feliz

A Tabela 5 apresenta um exemplo de entrada dessa etapa e sua saída, o texto normalizado. As expressões, termos ou símbolos que estiverem destacados no texto de entrada são normalizados e suas respectivas saídas também aparecem em destaque no texto normalizado. A seguir alguns dados que foram substituídos ou removidos no processo de normalização:

- Remoção de links e e-mails. Ex: remove “www.google.com”.
- Remoção de itens duplicados (letras, emoticons e pontuação). Ex: remove a letra “e” repetida de “ameeeei”, resultando em “amei”.
- Substituição de datas e horas. Ex: substitui “01/04/2017” pela palavra “data”.
- Substituição de valores monetários. Ex: substitui “R\$ 300,00” pela palavra “valor”.
- Correção ortográfica de algumas palavras mais comuns (ótimo, excelente, bom, incrível, e entre outras). Ex: substitui “encrevel” por “incrível”.
- Correção de algumas abreviações (para, com, não, também, demais, e entre outras). Ex: substitui “p/” por “para”.
- Substituição de risos. Ex: substitui “rsrsrs” por “sorri-dente”.

- Substituição de emoticons (feliz, tristeza e neutro). Ex: substitui “:)” por “feliz”.

A lista acima apresenta apenas alguns tipos de itens. A lista completa, com sua respectiva expressão regular, pode ser encontrada em [2]. A maioria das expressões regulares foi criada por [14] para a versão 1 da SentimentALL. Porém, para a versão 2 algumas dessas expressões foram alteradas e, também, novas expressões foram adicionadas. As expressões adicionadas para remover datas com meses por extenso, para tratar algumas abreviações (de com, para, também, etc.), remoção de caracteres etc. As alterações nas expressões foram sutis, como, por exemplo, no grupo de expressões que deveria adicionar espaços antes e depois de delimitadores “[ ]()”, duas expressões não estavam funcionando como deveria, pois em uma estava faltando o delimitador “(“ e na outra o “)”. [2] apresenta todas as expressões alteradas, adicionadas e que permaneceram iguais.

## B. CORREÇÃO ORTOGRÁFICA

Essa etapa tem a função de corrigir a grafia das palavras. O algoritmo utilizado foi baseado em [13], com algumas adequações necessárias para o funcionamento dele na língua portuguesa.

O algoritmo precisa de uma lista de palavras conhecidas e a frequência dessas palavras com base em um *corpus*. Mais adiante isso será explicado com mais detalhes. Em síntese, o algoritmo funciona da seguinte forma: ele recebe a palavra; se ela estiver na lista de palavras conhecidas, não precisará ser corrigida; caso contrário, diversas transformações na palavra são realizadas para gerar palavras candidatas; a palavra candidata conhecida que tiver a maior frequência será a palavra corrigida; se o algoritmo não gerou nenhuma palavra candidata conhecida, a palavra continua com sua forma original.

A primeira transformação realizada foi adicionada ao algoritmo de [13]. Ela está relacionada aos erros ortográficos mais comuns, em que algumas letras têm mais chance de serem trocadas por outras semelhantes (foneticamente) ou pela falta de acentuação. Por exemplo, a palavra “cafe”, em que o autor deveria ter substituído “e” por “é”. Com isso, deve-se agrupar letras que têm mais probabilidade de serem trocadas. Os agrupamentos considerados nesse trabalho foram os seguintes.

- cç
- aáãã
- eêê
- íí
- oóôô
- uúü

Caso uma palavra tenha alguma letra que está em algum dos agrupamentos, cada letra do grupo substitui a letra original, gerando uma palavra candidata. No exemplo anterior, a palavra “cafe” gera as seguintes palavras candidatas (considerando apenas o grupo ‘eêê’): “cafe”, “café” e “cafê”. Quando mais de um grupo é encontrado, são realizadas

combinações, como apresentadas a seguir (considerando o grupo “eêê” e “aáãã”): “cafe”, “café”, “cafê”, “cáfe”, “cáfê” e “cáfê”.

Outros agrupamentos podem ser feitos para aumentar a precisão, por exemplo: ‘mn’, ‘sz’, ‘ie’ e ‘jg’. Porém, a quantidade de palavras candidatas geradas pode crescer exponencialmente quando um novo agrupamento for adicionado, e isso afeta o tempo de execução do algoritmo. Por isso, foram definidos apenas seis agrupamentos.

Na segunda transformação são realizadas algumas manipulações nas letras. As manipulações são de eliminação, adição, transposição e substituição. As letras utilizadas na adição e substituição estão no seguinte conjunto de letras: “aáããbcçdeêêfghijklmnoôôpqrstuúüvwxyz”. A seguir é apresentado um exemplo para cada uma dessas manipulações.

- Eliminação – Ex.: “cafré”, removendo o “r”, fica “café”
- Adição – Ex.: “fejão”, adicionando o “i”, fica “feijão”
- Transposição – Ex.: “arrzo”, trocando as “z” e “o” de lugar, fica “arroz”
- Substituição – Ex.: “carne”, substituindo “m” por “n”, fica “carne”

Na terceira transformação é realizado o procedimento anterior novamente e, para cada palavra gerada, são realizadas novas manipulações. Ou seja, é realizada a manipulação das palavras manipuladas. Por exemplo, a primeira manipulação da palavra “café” geraria (dentre outras) a palavra “chafé” (Adição). Na segunda manipulação geraria (dentre outras) a palavra “chalé” (Substituição).

Após cada transformação é gerado um conjunto de palavras candidatas. Antes de passar para a próxima transformação, primeiro é verificado se foi gerada uma palavra conhecida. Se foi, não serão necessárias as outras transformações.

O algoritmo necessita da lista de palavras conhecidas praticamente o tempo todo e a precisão do algoritmo depende da quantidade de palavras que ela tem e que estejam corretas. [13] utilizou apenas um *corpus* e, a partir dele, criou uma lista de palavras e calculou suas frequências. Porém, nesse trabalho, além do uso do *corpus*, para aumentar a quantidade de palavras conhecidas, foi utilizada uma lista com quase um milhão de palavras encontradas no site do Project Natura no link a seguir: <http://natura.di.uminho.pt/download/sources/Dictionaries/wordlists/LATEST/>.

Como o *corpus* de [13] está em inglês, foi necessário criar um novo. Para isso, foram coletadas notícias nos sites da BBC Brasil (<http://www.bbc.com/portuguese>) e TNH1 (<http://www.tnh1.com.br>). Foram escolhidos sites de notícias, pois eles têm muitos conteúdos em texto e têm uma menor probabilidade de ter erros ortográficos do que uma rede social, por exemplo. Ao final, foram adicionadas à lista de palavras, novas palavras dos sites de notícias e adicionada a frequência de cada palavra no *corpus*.

Como a conclusão do desenvolvimento da lista de palavras, o algoritmo foi testado para avaliar a sua eficiência. Primeiro foram informadas para o algoritmo cem palavras corretas e depois cem palavras incorretas. Com as palavras

corretas, o algoritmo teve 100% de acerto, ou seja, ele não tentou corrigir algo que já estava correto. Com as palavras incorretas, o algoritmo teve 88% de acerto, porém em muitas destas palavras observou-se uma complexidade na realização da correção, devido à semelhança apresentada com outras palavras, por exemplo, “fazrem” foi corrigida para “fazem”, mas a palavra correta é “fazerem”.

Ao final dessa etapa, tem-se os comentários tokenizados em sentenças normalizadas e corrigidas ortograficamente. Na seção a seguir será apresentada a etapa de a *PoS Tagging*, em que as sentenças serão etiquetadas morfológicamente.

### C. POS TAGGING

Para realizar o *PoS tagging* foi necessário utilizar o *corpus* Mac-Morpho. Esse *corpus*, disponibilizado na NLTK, possui mais de um milhão de palavras etiquetadas com 26 tipos de *PoS tags*.

A NLTK possui um módulo que realiza o *PoS Tagging*. Com ele é possível treinar classificadores, chamados de *taggers*. Para isso, foi utilizado o Mac-Morpho para treinar oito *taggers*. Os *taggers* tem a seguinte nomenclatura: *n-gram tagger*. Onde *n* representa a quantidade de *tokens* que formam o contexto de um *token* que será etiquetado. O contexto de um *token* é ele próprio mais as  $n - 1$  *tags* de *tokens* anteriores. A Figura 7 apresenta um exemplo do contexto de um *3-gram tagger*, também chamado de *trigram tagger*.

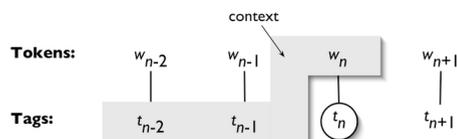


Figura 7. Contexto do Tagger. Bird, Klein e Loper (2009)

A área sombreada em cinza na Figura 7 representa o contexto do *token*  $w_n$ . Para o *3-gram tagger* determinar a *tag* de  $w_n$ , considera-se o *token*  $w_n$  e as duas tags ( $t_{n-2}$  e  $t_{n-1}$ ) dos *tokens* anteriores. O *3-gram tagger* busca pelo padrão, caso encontre determina a *tag*  $t_n$  de  $w_n$ .

Além dos *oitos taggers* (*8-gram tagger*, *7-gram tagger*, ...) treinados, há também um *tagger* padrão. A classificação de um *token* utilizando esses *taggers* funciona basicamente como explicado a seguir. Inicia-se utilizando o *8-gram tagger*, se no *corpus* possuir o mesmo padrão apresentado no contexto do *token*, ou seja, a mesma sequência de setes *tags* mais o *token*. Então, a mesma *tag* do *token* do *corpus* é atribuída ao *token* do contexto. Se não possuir o padrão, utiliza-se o *7-gram tagger* e o processo se repete. Se nenhum dos *n-gram taggers* conseguir classificar o *token*, então é usado o *tagger* padrão, etiquetando o *token* como a *tag* DEFAULT.

O procedimento de *PoS tagging* utilizando o *corpus* Mac-Morpho possui um problema ao etiquetar as contrações e combinações de preposições (de, a, em e por) com os artigos, alguns pronomes e alguns advérbios.

Esse problema se deve ao fato de que no *corpus* não existem essas contrações, elas são escritas na sua forma sem contrações, por exemplo, “da” é escrito como “de a”. Porém, esse problema foi contornado atualizando o Mac-Morpho para suportar as contrações e gerando um novo *corpus*. Para isso, foi feita uma varredura em todo o Mac-Morpho buscando pelas preposições (de, a, em e por), caso a próxima palavra for alguma palavra que pode ser contraída com a anterior, então é feita a contração. Quando uma contração é realizada, é feita a união das *pos tags* com um sinal de “+”. Por exemplo, a varredura encontra a palavra “de” com a *pos tag* “PREP” a próxima palavra é “a” com a *pos tag* “ART”. A contração de “de” com “a” é “da” e a nova *pos tag* será “PREP+ART”. Com isso, é gerado um novo *corpus* e o problema das contrações e combinações foi resolvido.

Um outro problema identificado foi o fato de que muitas palavras foram identificadas com a *tag* DEFAULT, pois não estavam presentes no MAC-MORPHO. Portanto, foi utilizada uma lista de adjetivos e uma de verbos, disponíveis no site Linguatca. Essas listas informam a palavra e uma frequência. As palavras com frequência menor ou igual a três foram removidas, pois não estavam classificadas de forma correta. Quando os *taggers* etiquetavam uma palavra como DEFAULT, era verificado se ela estava na lista de adjetivos, se estivesse, era definida a *tag* como “ADJ”. Se não fosse um adjetivo, verificava se ela estava na lista de verbos, se estivesse, definia a *tag* como “V”. Se não fosse um verbo, verificava se ela era um número usando a expressão regular “\d+([\.,]\d+)?” e etiquetava como “NUM”. Se depois de todas as verificações nenhuma condição fosse satisfeita, as palavras eram etiquetadas como substantivo (*tag* N).

A seguir, é apresentada a última etapa do pré-processamento, que utiliza a *PoS tag* para encontrar padrões para identificar possíveis expressões multipalavras.

### D. IDENTIFICAÇÃO DE EXPRESSÕES MULTIPALAVRAS

Essa etapa é composta de dois passos: encontrar expressões candidatas e filtrar as expressões. O primeiro passo faz a busca nos dados procurando por padrões morfológicas que identifica expressões candidatas. Os padrões utilizados servem para identificar expressões do tipo composto nominal. A Tabela 4 apresenta esses padrões.

O algoritmo utilizado para realizar o primeiro passo é baseado no pseudocódigo usado em [15]. A mudança está apenas no fato de que em [15] foram consideradas expressões com até quatro *tokens* e nesse trabalho foram consideradas apenas três, pois as contrações (por exemplo, “das” – “de” + “as”) não foram desmembradas. O algoritmo inicia procurando por um substantivo nos comentários, quando encontra, verifica se ele e mais dois *tokens* a frente formam alguns dos padrões morfológicos. Caso formem um padrão, os três *tokens* passam a compor o conjunto de expressões candidatas. Se não formar, verifica se o próximo *token* é um substantivo ou adjetivo e, se for, esses *tokens* serão uma expressão candidata. Esse processo se repete até que todas expressões candidatas sejam encontradas. Quando uma expressão candi-

data é encontrada, calcula-se o seu PMI, usando a Equação 2, proposta por [6] (2011):

Foram encontradas por volta de três milhões de expressões candidatas. Utilizando essas expressões candidatas, inicia-se o segundo passo, que é a de filtragem das expressões. A filtragem é realizada tendo como base um intervalo de valor do PMI. Para determinar esse intervalo, foram feitos testes com diversos valores de intervalos em um conjunto de cinquenta frases. Nesse conjunto foram destacadas sessenta e quatro expressões multipalavras. Para definir o melhor intervalo foi utilizado a *Precision* e o *Recall*. Para ter uma única medida de desempenho foi usado o *F-Measure*, que faz uma média harmônica com os valores de *Precision* e *Recall*. Essa métrica foi utilizada, pois ela tem a característica de que, caso um dos valores esteja muito baixo, a média também será baixa (Sasaki, 2007). Como tanto a *Precision* como o *Recall* são relevantes, isso é uma característica desejável. A seguir são apresentadas as equações de *Precision*, *Recall* e *F-Measure*, respectivamente.

$$Precision = \frac{V_p}{V_p + F_p} \quad (3)$$

$$Recall = \frac{V_p}{V_p + F_n} \quad (4)$$

$$FMeasure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

Onde  $V_p$  estão erradas; e  $F_n$  é a quantidade de expressões que deveriam ter sido reconhecidas, mas não foram.

O teste foi feito variando o  $x$  de 0 a 9, com intervalo de 1, da seguinte equação:  $x < PMI$ . Com isso, são consideradas como expressões multipalavras as expressões que tiverem o PMI maior do que  $x$ . A Figura 8 apresenta o gráfico com os valores de *Precision*, *Recall* e *F-Measure*, para cada valor de  $x$ .

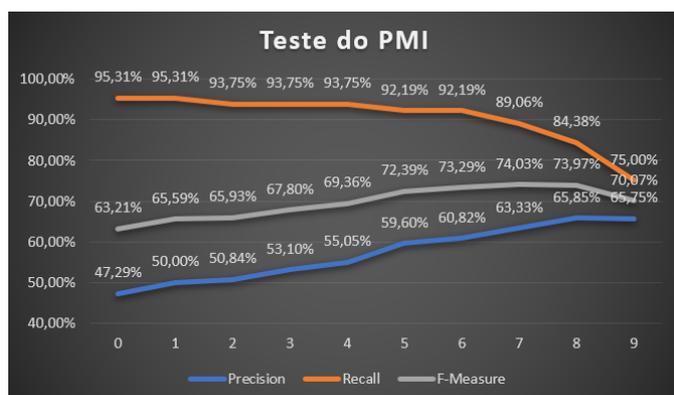


Figura 8. Resultado do teste do PMI

Como pode se observar na Figura 8, o valor de  $x$  em que houve a melhor *F-Measure* é 7. Ressalta-se ainda que, eliminando expressões com pouca frequência (menor do que 25), o resultado melhora. A *Precision* vai para 73,33%, o *Recall* para 85,94% e *F-Measure*, 79,14%.

## V. CONCLUSÕES

O módulo de pré-processamento foi implementado com o propósito de ser livre de contexto e de que sua utilização fosse relativamente simples, considerando que será usado por outras equipes do grupo de pesquisa. Para cada etapa do módulo há uma entrada e uma saída, sendo que nestas etapas é irrelevante a origem dos dados ou o que será feito com a saída. Com isso, pode se criar programas que carregam os dados de arquivos de texto, banco de dados, dentre outros, contanto que os dados de entrada estejam no formato requerido pelo módulo. O mesmo acontece com a saída, que pode ser salva em um arquivo de texto, banco de dados etc. Todas as etapas que compõem o módulo são independentes, portanto algumas delas podem ser ignoradas, por exemplo, a normalização ou a correção ortográfica. Contudo, nesse trabalho todas as etapas foram utilizadas e os dados tiveram como origem e destino o banco de dados.

O módulo de tokenização de sentenças e palavras estão bastante precisos para a língua portuguesa na biblioteca NLTK do Python, isso forneceu uma dinamicidade maior na implementação do módulo. O PoS tagging do NLTK se mostrou capaz de realizar a etiquetagem morfológica nos comentários, porém ocorreu um problema na classificação das palavras com contração que foi corrigido atualizando o Mac-Morpho. Houve, também, um problema em relação a palavras que não estão no corpus Mac-Morpho, mas foi contornado utilizando uma lista de adjetivos e verbos, disponíveis no site Linguatca e uma expressão regular para a identificação de números. Na correção ortográfica, o resultado do teste foi satisfatório, pois o algoritmo corrigiu 88% das palavras incorretas de forma correta, porém, uma análise do contexto em que a palavra está inserida poderá melhorar os resultados do algoritmo.

## Referências

- [1] ALLEN, James. Natural Language Understanding. 2. ed. S.I: Pearson, 1995. 654 p.
- [2] ARAUJO, L. G. A. SENTIMENTALL VERSÃO 2: Desenvolvimento de Análise de Sentimentos em Python, 103 f., TCC (Graduação) - Curso de Bacharelado em Ciências da Computação, Centro Universitário Luterano de Palmas, Palmas, Tocantins. 2017.
- [3] BOOS, Rodrigo Augusto Scheller; PRESTES, Kassius Vargas; VILLAVICENCIO, Aline. Identification of Multiword Expressions in the brWaC. In: Language Resources and Evaluation Conference (LREC), 9., 2014, Reykjavik (iceland).
- [4] Brito, Parcilene Fernandes de. RELATOS VERBAIS DE CONSUMIDORES EM AVALIAÇÕES ON-LINE: PROSPECÇÃO COMPUTACIONAL E INTERPRETAÇÕES COM BASE NO BEHAVIORAL PERSPECTIVE MODEL (BPM). 2018.182 fl. Tese(Programa de Pós-Graduação STRICTO SENSU em Psicologia) - Pontifícia Universidade Católica de Goiás, Goiânia-GO .
- [5] Brito, P. F., Oliveira, W. C. C., Souza, J. G. (2015). SentimentALL. Fábrica de Software: CEULP/ULBRA.
- [6] CHURCH, Kenneth Ward; HANKS, Patrick. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics. Cambridge, Ma, EUA, p. 22-29. mar. 1990.
- [7] COPPIN, Ben. Inteligência Artificial. Rio de Janeiro: Ltc, 2013. 636 p.
- [8] CRUYS, Tim van de. Two multivariate generalizations of pointwise mutual information. In: WORKSHOP ON DISTRIBUTIONAL SEMANTICS AND COMPOSITIONALITY DISCO, 11, 2011, Stroudsburg, Pa, EUA. Proceedings. Stroudsburg, Pa, EUA: Association for Computational Linguistics, 2011. p. 16 - 20.

- [9] DALE, Rober. Classical Approaches to Natural Language Processing. In: INDURKHYA, Nitin; DAMERAU, Fred J. (Ed.). Handbook of Natural Language Processing. 2. ed. Boca Raton, Fl: Chapman and Hall/crc, 2010. Cap. 1. p. 3-7.
- [10] DOMINGUES, Miriam Lúcia Campos Serra. ABORDAGEM PARA O DESENVOLVIMENTO DE UM ETIQUETADOR DE ALTA ACURÁCIA PARA O PORTUGUÊS DO BRASIL. 2011. 137 f. Tese (Doutorado) - Curso de Programa de Pós-graduação em Engenharia Elétrica, Instituto Tecnológico, Universidade Federal do Pará, Belém, 2011.
- [11] GÜNGÖR, Tunga. Part-of-Speech Tagging. In: INDURKHYA, Nitin; DAMERAU, Fred J. (Ed.). Handbook of Natural Language Processing. 2. ed. Boca Raton, Fl: Chapman and Hall/crc, 2010. Cap. 10. p. 205-235.
- [12] JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Englewood Cliffs, New Jersey: Prentice Hall, 2000. 934 p.
- [13] LJUNGLÖF, Peter; WIRÉN, Mats. Syntactic Parsing. In: INDURKHYA, Nitin; DAMERAU, Fred J. (Ed.). Handbook of Natural Language Processing. 2. ed. Boca Raton, Fl: Chapman and Hall/crc, 2010. Cap. 4. p. 59-91.
- [14] MENEZES, Paulo Blauth. Linguagens Formais e Autômatos. 6 ed. Porto Alegre: Bookman, 2011. 256 p.
- [15] NORVIG, P. "How to Write a Spelling Corrector". 2016. Disponível em: <<http://norvig.com/spell-correct.html>>, Acesso em: 12 jan. 2019.
- [16] OLIVEIRA, W. C. C. "SENTIMENTALL: Ferramenta para análise de sentimentos em português", 86 f., TCC (Graduação) - Curso de Bacharelado em Sistemas de Informação, Centro Universitário Luterano de Palmas, Palmas, Tocantins. 2018.
- [17] OLIVEIRA, W. C. C. e BRITO, P. F. "Utilização do Pointwise Mutual Information na Identificação de Expressões Multipalavras", In: XVII ENCOINFO – CONGRESSO DE COMPUTAÇÃO E SISTEMAS DE INFORMAÇÃO, Palmas. p. 25 - 34. 2015.
- [18] PALMER, David D. Text Preprocessing. In: INDURKHYA, Nitin; DAMERAU, Fred J. (Ed.). Handbook of Natural Language Processing. 2. ed. Boca Raton, Fl: Chapman and Hall/crc, 2010. Cap. 2. p. 9-30.
- [19] RUSSELL, Stuart; NORVIG, Peter. Inteligência Artificial. 3. ed. Rio de Janeiro: Elsevier, 2013. 1016 p. Tradução de: Regina Célia Simille.
- [20] SAG, Ivan A. et al. Multiword Expressions: A Pain in the Neck for NLP. In: CICLING INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 3., 2002, Mexico-city. Proceeding. London, Uk: Springer-verlag, 2002. p. 1 - 15.



**LUAN GOMES DE ALMEIDA ARAÚJO**  
Possui graduação em Ciências da Computação (CEULP/ULBRA), Técnico em Informática (IFTO) e Técnico em Mecatrônica (IFTO) e especialização em andamento em Prática de Metodologias Ágeis (UniCesumar). Atualmente trabalha no Tribunal Regional Federal da 1ª Região, desenvolvendo sistemas judiciais.

...



**PARCILENE FERNANDES BRITO** Doutora em Psicologia (PUC/GO), Mestre em Ciências da Computação (UFSC), Especialista em Ciências da Computação (UFSC), Especialista em Informática para Aplicações Empresariais (ULBRA), Graduada em Psicologia (CEULP/ULBRA), Graduada em Processamento de Dados (UNITINS). Pesquisadora do Grupo de Pesquisa Engenharia Inteligente de Dados (CEULP/ULBRA). Atualmente é Diretora Acadêmica do Ceulp/Ulbra, Coordenadora e Professora dos cursos de Sistemas de Informação, Ciência da Computação e Engenharia de Software na mesma IES. Tem experiência na área de Computação e Psicologia, com ênfase em Lógica Formal, Análise de Sentimentos, Ontologias, Psicologia do Consumidor e Informática na Educação.

# Apoio à decisão na prevenção de acidentes de trânsito: uma aplicação dos métodos AHP e PROMETHEE

**JOÃO VICENTE PEREIRA,  
RÔMULO DE MEDEIROS PALMEIRA e  
ANA CARLA B. REIS.**

Programa de Pós-Graduação em Computação Aplicada - UNB  
(e-mail: {joao.vicentep, romulopalmeira}@gmail.com, anacarlabr@unb.br)  
Autor Correspondente: João Vicente Pereira(e-mail: joao.vicentep@gmail.com).

• **RESUMO** - Uma das maiores causas de mortes no mundo é provocada por acidentes de trânsito envolvendo veículos e pedestres. Este artigo apresenta uma aplicação da metodologia multicritério para ordenação das Regiões Administrativas do Distrito Federal-DF, no Brasil, considerando os acidentes de trânsito com mortes e feridos. Identificou-se a variável natureza do acidente para criação dos critérios, tipo de acidente para subcritérios e a regiões administrativas (cidades) para alternativas. Aplicou-se os métodos AHP e PROMETHEE aos dados do Anuário Estatístico do DETRAN-DF de 2016. Os resultados obtidos foram comparados com a Unidade Padrão de Severidade (UPS), que é a medida utilizada pela maioria dos Detrans do país, indicando possibilidade de aperfeiçoamento nos critérios adotados pelos decisores para definição de estratégias, ações e medidas para melhoria da segurança no trânsito.

• **PALAVRAS-CHAVE** - Análise Multicritério, Apoio à Decisão, Acidentes de trânsito, AHP e PROMETHEE

## I. INTRODUÇÃO

Segundo dados da Organização Mundial de Saúde (OMS) [1], cerca de 1,25 milhão de pessoas morrem por ano vítimas de acidentes de trânsito nas estradas, sendo considerada a maior causa mundial de mortes. Grande parte dos óbitos vem acontecendo nos países emergentes, por causa do elevado crescimento econômico, do aumento nos níveis de urbanização e da elevação no número de veículos automotores. Estima-se uma perda de 3% do PIB das economias em desenvolvimento, provocada pelos acidentes rodoviários.

A OMS estima para o Brasil 47 mil mortes por ano provocadas por acidentes rodoviários, o que representa aproximadamente uma taxa de 23,4 mortes para cada 100 mil pessoas. Mais de 80% das vítimas são do sexo masculino e calcula-se uma perda de 1,2% do PIB brasileiro provocada pelas colisões nas estradas. Em relação às categorias de usuários rodoviários, o maior contingente de mortes ocorre com condutores de veículos com 2 rodas (28%), seguido pelos pedestres com 20% e passageiros com veículos de 4 rodas.

Dados preliminares do Sistema Único de Saúde (SUS) [2] mostram que em 2016 ocorreram 35,7 mil óbitos por acidentes de transporte no Brasil e as regiões com maiores

registros foram Nordeste (11,2 mil óbitos), Sudeste (10,9 mil óbitos) e Sul com 6 mil registros. A região Centro-Oeste registrou 3,9 mil óbitos, ficando em 4º lugar, com os maiores registros observados nos estados de Goiás e Mato Grosso.

O Distrito Federal (DF), capital do Brasil localizada na região centro-oeste, possuía em 2016 uma frota de veículos de 1,6 milhão e registrou cerca de 9 mil acidentes com vítimas de trânsito, com 363 mortes. A taxa de acidentes com vítima por 10 mil veículos foi de 53 e a média de acidentes com vítima foi de 24,1 por dia. Os principais tipos de acidentes com mortes ocorreram por colisão, atropelamento de pedestres, choque com objeto fixo e capotamento/tombamento. Esses dados e outras informações estão no Anuário Estatístico de Acidentes de Trânsito [3], elaborado pelo Detran com a participação de vários órgãos do Governo do DF como Secretarias de Segurança Pública e de Saúde, Polícias Civil e Militar e Corpo de Bombeiros, que alimentam o Sistema de Informações de Acidentes de Trânsito (SAT), cujo objetivo é subsidiar o planejamento, o desenvolvimento e a avaliação das ações voltadas à redução de acidentes de trânsito.

Um importante indicador de análise dos usuários do SAT é a Unidade Padrão de Severidade (UPS), definida pelo Departamento Nacional de Trânsito [4] com base no número de

acidentes, classificados em três níveis conforme a gravidade: sem vítima, com ferido e com morte. Em virtude da baixa qualidade dos registros de acidentes sem vítima, o DETRAN-DF adota apenas o número de acidentes com ferido e com morte como critérios para o cálculo do indicador, chamando de UPS modificada (UPSm). Ela é utilizada para ordenação das vias do DF, apontando os locais onde ocorrem os acidentes com maior severidade e servindo de informação relevante para apoiar e melhorar o processo de tomada de decisão relativas às políticas, estratégias, ações e medidas que proporcionem um trânsito mais seguro.

Diante deste contexto, o objetivo deste estudo é apresentar proposta de aperfeiçoamento do cálculo do indicador UPSm do DETRAN-DF, por meio da metodologia de Análise de Decisão Multicritério (MCDA) e da aplicação dos métodos *Analytic Hierarchy Process* (AHP) e *Preference Ranking Method for Enrichment Evaluation* (PROMETHEE), que levem em consideração outros critérios além do número de acidentes, visando assim melhorar a qualidade das informações utilizadas pelos órgãos de trânsito, como por exemplo DENATRAN, Secretarias de Saúde e de Segurança do DF, Polícia Civil e Militar, Corpo de Bombeiros e comunidade em geral.

Para cumprir o objetivo proposto, este artigo apresenta na seção 2 a revisão da literatura sobre o assunto, o processo MCDA e os métodos AHP e PROMETHEE aplicados a situações ou estudos de casos próximos ao tema de acidentes de trânsito. Na seção 3 é descrita a metodologia empregada para resolver o problema de decisão. Na seção 4 são apresentados os resultados da aplicação dos métodos AHP e PROMETHEE nos dados de acidentes de trânsito ocorridos nas regiões administrativas (cidades) do DF. Na seção 5 é realizada avaliação crítica dos métodos adotados e finalmente na seção 6 tem a conclusão com propostas de utilização dos métodos avaliados e sugestão de trabalhos futuros.

## II. REVISÃO DE LITERATURA

### A. ESTUDOS RELACIONADOS AO PROBLEMA DE ACIDENTES DE TRÂNSITO

Nesta parte da pesquisa bibliográfica, buscou-se identificar os principais estudos relacionados a trânsito de veículos, priorizando conteúdos publicados nos últimos dez anos envolvendo acidentes, segurança, estatísticas e informações correlatas ao problema estudado. A seguir são resumidos os resultados encontrados:

- Novoa et al. [5] realizaram estudo sobre a efetividade dos pacotes de medida de segurança rodoviária, tendo como população as pessoas feridas em acidentes rodoviários na Espanha, no período de 2000 a 2006. Foi utilizado modelo de regressão de séries temporais Quasi-Poisson com o número de pessoas feridas sendo a variável dependente e a intervenção política com medidas de segurança como variável explicativa. A conclusão foi de que a priorização da segurança rodoviária reduz o número de pessoas feridas em colisões de trânsito;

- Hu et al. [6] compararam duas bases de dados sobre taxas de mortalidade no trânsito da China, originadas de relatórios policiais e registros de óbitos, no período de 2002 a 2007. Para comparação, foi utilizada técnica estatística como o teste Qui-quadrado e o modelo de regressão linear para verificar a tendência da mortalidade ao longo do período. A taxa de mortalidade obtida dos registros de óbitos foi duas vezes maior que a taxa informada nos relatórios policiais, e a regressão baseada em fontes policiais mostrou uma redução de 27% na taxa de mortalidade e pouca mudança no modelo baseado nos registros de óbitos, indicando assim que a qualidade dos dados policiais precisa ser investigada, monitorada e aperfeiçoada;
- Chen et al. [7] apresentaram uma estrutura hierárquica do Índice de Risco de Segurança (RSRI), baseada em fatores humanos, do veículo, de estrada, do meio ambiente, de gerenciamento, risco pessoal e risco de tráfego, para a avaliação do desempenho da segurança rodoviária em um conjunto de províncias da China. A metodologia empregada foi a entropia melhorada TOPSIS, caracterizada por ser uma técnica indicada para problemas envolvendo preferência de ordens por semelhança e bastante conhecida como um dos métodos clássicos multicritérios de tomada de decisão (MCDM);
- Staton et al. [8] avaliaram dados sobre acidentes rodoviários em Ruanda e no Sri-Lanka, por meio de pesquisas com os usuários das estradas, como alternativa aos registros policiais, buscando um mapeamento mais preciso dos locais perigosos, isto é, com alta densidade de ocorrências de colisões. Os métodos empregados foram comparação entre os dados da polícia e dos usuários pesquisados e a análise espacial utilizando sistemas de informação geográfica para mapear os pontos críticos de tráfego. Os resultados apontaram para a possibilidade de uso das pesquisas, por ser uma alternativa viável, barata e rápida em relação aos dados policiais;
- Paixão et al. [9] analisaram a contribuição de três diferentes fontes de informações na descrição de acidentes de trânsito em Belo Horizonte, utilizando estatísticas como proporções, coeficientes e medidas de associação dos dados coletados nos Sistemas de Informação da Empresa de Transportes e Trânsito, de Internações Hospitalares e Mortalidade. Os autores verificaram problemas no preenchimento das informações, porém identificaram os principais fatores associados aos acidentes: os pedestres idosos, motociclistas, o uso de álcool e excesso de velocidade;
- Chen et al. [10] identificaram os principais fatores que afetam duas categorias de colisões ou batidas existentes na China: acidente de trânsito grave ou acidente de trânsito muito grave, de acordo o número de mortes, quantidade de pessoas feridas e o custo total, compreendendo dados do período de 2007 a 2014. Cinco fatores de risco foram identificados como significativos para avaliação de impacto das colisões: localização, alinhamento, velocidade, condição da estrada e condição do veículo.

mento vertical, classificação de segurança na estrada, distração do motorista e excesso de carga no veículo. Os autores sugerem algumas recomendações para diminuir os acidentes como a melhoria da manutenção de semáforos e da iluminação das ruas, maior fiscalização dos agentes de trânsito e aperfeiçoamento da sinalização e dos controles de tráfego;

- Hasan e Asad [11] avaliaram a literatura sobre os acidentes de trânsito no Iraque no período de 2005 a 2015, pesquisando estatísticas oficiais, estudos de acidentes de trânsito e políticas atuais de segurança rodoviária e leis de trânsito. O resultado revelou vários pontos importantes em aspectos como características de colisão, ocorrência, gravidade e medidas. Os resultados também podem contribuir nos estudos sobre segurança rodoviária em países em desenvolvimento e nas ações governamentais e não-governamentais para redução no número de vítimas de acidentes de trânsito. Outra conclusão importante são os principais fatores que contribuem para os acidentes, como os perfis dos usuários das estradas, a situação das estradas, as condições do tempo, o estado dos veículos e as características das colisões, como por exemplo localização, dia e horário dos acidentes;
- Boulieri et al. [12] analisaram as dependências entre os níveis de gravidade dos acidentes rodoviários na Inglaterra, no período de 2005 a 2013, com as correlações espaciais e temporais, utilizando o modelo de efeitos multivariados conforme formulação hierárquica bayesiana. Foram encontradas associações importantes entre os efeitos de severidade, os níveis de gravidade e os mapas de taxas de acidentes, que permitiram produzir mapas informativos com o comportamento dos acidentes ao longo do tempo e a identificação de áreas com maior risco, sendo tais informações relevantes, por exemplo, para as estratégias de políticas de saúde;
- Marques e Machado [13] verificaram alguns fatores que podem auxiliar na tomada de decisão sobre os recursos arrecadados com as multas de trânsito, através da técnica Delphi e aplicação de um questionário para entrevistar especialistas em 10 capitais brasileiras, tendo como resultado uma estrutura hierárquica dos fatores com as dimensões. Os fatores mais relevantes, após agrupamento em dimensões, foram o histórico dos acidentes de trânsito, o histórico das multas de trânsito, situação da sinalização da via, quantidade de vias municipalizadas, taxa de motorização da população, taxa de crescimento populacional, nível de cobrança da imprensa local e histórico da aplicação dos recursos.

## **B. APOIO À DECISÃO MULTICRITÉRIO MCDA**

Esta revisão da literatura busca referenciar os principais conceitos, características e possibilidades de utilização da metodologia de Análise de Decisão Multicritério (MCDA). Belton et al. [14] descrevem o processo MCDA como um conjunto de métodos formais para tornar explícitos critérios múltiplos que apoiem tomadores de decisão, individuais ou

em grupo, nos problemas de decisão importantes. O MCDA pode ser considerada uma ferramenta de tomada de decisão que procura integrar medidas objetivas com julgamento de valor, tendo como principal benefício aos decisores melhorar o entendimento do problema enfrentado e identificar possíveis ações ou providências a serem priorizadas.

A atividade de ajuda de decisão multicritério não pode ser restrita apenas à resolução de um problema em que é necessário agregar determinadas preferências em um determinado conjunto potenciais de ações. A identificação do conjunto de ações (e da confusão de sua fronteira), a construção dos critérios e a modelagem de preferências são aspectos fundamentais e, muitas vezes, difíceis da ajuda à decisão (COSTA, [15]).

Segundo Belton et al. [14], um dos principais objetivos da abordagem MCDA é ajudar o decisor a organizar e sintetizar as informações de forma que o leve a se sentir confortável e confiante em tomar uma decisão, minimizando o potencial de arrependimento pós-decisão ao estar satisfeito de que todos os critérios ou fatores foram adequadamente levados em consideração.

Costa [15] afirma que o MCDA permite melhorar o grau de conformidade e coerência entre a evolução de um processo de tomada de decisão, os sistemas de valores e os objetivos dos envolvidos nesse processo, sendo fundamental que os conceitos, ferramentas e procedimentos ajudem a ampliar a capacidade de solução de problemas na presença de ambiguidade, incerteza e abundância de alternativas.

Outro autor que analisa o conceito e o processo de decisão, levando-se em conta as interações entre os stakeholders como os agentes individuais, instituições e comunidades, é Roy [16] que destaca três importantes atores no processo de decisão: o tomador de decisão, o analista e o cliente.

### 1) AHP - Analytic Hierarchy Process

O AHP surgiu como um método útil para resolução de problemas relacionados à tomada de decisões complexas, envolvendo múltiplos critérios. Desenvolvido por Saaty [17], a técnica funciona pela decomposição de um problema em critérios hierarquizados. É baseado na avaliação das alternativas em termos de uma função preferência aditiva, em que são realizadas comparações par a par das alternativas com relação aos critérios. Conforme Ishizaka e Nemery [18], o processo abrange três etapas principais: definição do objetivo, dos critérios e das alternativas.

A comparação entre os critérios é feita com base nos julgamentos das partes interessadas, que atribuem valores à cada alternativa, com base em julgamentos objetivos ou subjetivos. Porém, devido à liberdade que o método fornece para a classificação dos critérios e alternativas, podem ocorrer inconsistências entre os julgamentos realizados. De forma a estabelecer um parâmetro que determinasse o nível aceitável de inconsistências de acordo com o número de comparações realizadas, Saaty [17] propôs uma “taxa de consistência” (CR), conforme a Equação (6):

$$CR = \frac{CI}{RI} \quad (6)$$

Onde:

- *CR* (*Consistency Ratio*): indicador que compara os julgamentos realizados pelos avaliadores;
- *RI* (*Random Index*): julgamentos aleatórios calculados pelo próprio autor para matrizes de diferentes tamanhos, conforme Tabela 1.
- *CI* (*Consistency Index*): índice de consistência calculado de acordo com a Equação (7):

$$CI = \frac{\lambda_{Max} - n}{n - 1} \quad (7)$$

Em que:

- $\lambda_{Max}$  é o resultado da multiplicação da matriz de comparação pelo seu vetor de eigen dividido pelo mesmo vetor de eigen, calculando-se, por fim, a média final dos valores obtidos;
- *n* é a quantidade de itens (critérios ou subcritérios) envolvidos no cálculo;

0.9

**Tabela 1.** Índices aleatórios calculados por Saaty [17]

n	2	3	4	5	6	7	8	9	10
RI	0,00	0,58	0,90	1,12	1,24	1,32	1,41	1,45	1,49

Saaty [17] sugere que um conjunto de decisões pode ser considerado “confiável” se a taxa de consistência permanecer abaixo de 0,1, ou 10%. Na visão do autor, taxas muito acima deste parâmetro aproximam-se demasiadamente de resultados aleatórios, o que indica que os julgamentos iniciais carecem de revisão.

## 2) PROMETHEE - Preference Ranking Method for Enrichment Evaluation

Semelhante ao AHP, o PROMETHEE também compara as alternativas par a par indicando o desempenho de cada uma para um determinado critério (BRANS e MARESCHAL) [19]. Para a realização da metodologia PROMETHEE se faz necessário calcular:

$$\Pi(a, b) = \sum_{i=1}^n w_i * P_i(a, b) \quad (8)$$

Onde:

- $\Pi(a, b)$  - Grau de preferência da alternativa a em relação à b, para todos os critérios;
- $w_i$  - Peso do critério  $i$  ( $i = 1, 2, \dots, n$ );
- $P_i(a, b)$  - Função de preferência que pode assumir valores entre 0 e 1.

A Função de Preferência se apresenta de seis formas diferentes, cabendo ao decisor escolher a mais adequada para o critério a representar, não necessariamente usando a mesma forma para todos os critérios.

A Equação (9) é usada para construção dos critérios das funções de preferências apresentados na Tabela 2.

$$\Delta(a, b) = g_i(a) - g_i(b) \quad (9)$$

Onde:

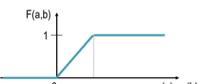
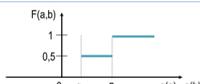
- $g_i(a)$  - representa o critério  $i$  da alternativa  $a$ ;
- $g_i(b)$  - representa o critério  $i$  da alternativa  $b$ ;
- $\Delta(a, b)$  - é a diferença do critério entre as alternativas  $a$  e  $b$ .

## 3) Uso dos Métodos AHP e PROMETHEE em Estudos Relacionados

Esta parte da pesquisa buscou encontrar os principais estudos com a adoção dos métodos AHP e o Promethee a casos envolvendo o tema de acidentes de trânsito ou assuntos semelhantes, os quais estão resumidos abaixo:

- Bernardini et al. [20] desenvolveram uma metodologia multicritério para realizar uma avaliação integrada e preventiva de projetos rodoviários considerando seus desempenhos de segurança e alguns aspectos econômicos e ambientais. O autor aplicou, com base em um projeto real de infraestrutura rodoviária rural na Bélgica, uma técnica de agrupamento de critérios baseada no método PROMETHEE para identificar grupos de alternativas similares suportadas por uma estrutura parcialmente ordenada.
- Bauza et al. [21] estudaram os sistemas de veículos cooperativos desenvolvidos para melhorar a segurança e gestão do tráfego e apresentou uma nova técnica de cooperação baseada em comunicações veículo-veículo (V2V) e lógica difusa para detectar o congestionamento do tráfego rodoviário sem a necessidade de implantar sensores de infraestrutura.
- Turcksin et al. [22] recomendaram um pacote de políticas multifatorial ao governo belga para reduzir as externalidades ambientais, incentivando as pessoas a fazerem uma escolha de veículo mais sustentável. Para isso, o autor propôs uma abordagem integrada para o problema de decisão combinando os métodos AHP e PROMETHEE, permitindo uma avaliação dos cenários políticos com seus pontos fortes e fracos e fornecendo uma classificação que facilita a seleção final para o tomador de decisões.
- Pitchipoo et al. [23] analisaram a questão da visibilidade dos motoristas em veículos pesados e a relação com os acidentes rodoviários, através da otimização dos parâmetros utilizados no projeto do espelho retrovisor. Foi utilizada uma abordagem de tomada de decisão multicritérios (MCDM) chamada técnica COPRAS (Análise Proporcional Completa de Alternativas) e os pesos dos parâmetros do projeto foram calculados usando três abordagens diferentes, como AHP, FARE (Relação Fatorial) e Medição de Entropia.
- Babae et al. [24] avaliaram a segurança e mobilidade de motoristas idosos, usando o conceito de indicadores

**Tabela 2.** Critérios gerais para o Promethee

Critério	$F(a, b) =$	Parâmetro		Descrição	Gráfico
		$q$	$p$		
Usual	$0$ se $\Delta(a, b) > 0$ $1$ se $\Delta(a, b) \leq 0$	✗	✗	não há parâmetros a ser definido	
Quase-Critério	$0$ se $\Delta(a, b) \leq q$ $1$ se $\Delta(a, b) > q$	✓	✗	limiar de indiferença	
Limite de preferência	$1$ se $\Delta(a, b) > p$ $\Delta(a, b)/p$ se $0 < \Delta(a, b) < p$ $0$ se $\Delta(a, b) < 0$	✗	✓	limiar de preferência	
Pseudocritério	$0$ se $\Delta(a, b) \leq q$ $0,5$ se $q < \Delta(a, b) \leq p$ $1$ se $\Delta(a, b) > p$	✓	✓	$q$ : limiar de indiferença $p$ : limiar de preferência	
Área de indiferença	$0$ se $\Delta(a, b) \leq q$ $\Delta(a, b)/(p - q)$ se $q < \Delta(a, b) \leq p$ $1$ se $\Delta(a, b) > p$	✓	✓	$q$ : limiar de indiferença $p$ : limiar de preferência	
Área de indiferença	$0$ se $\Delta(a, b) \leq 0$ $1 - e^{-\frac{[\Delta(a,b)]^2}{2\delta^2}}$ se $q < \Delta(a, b) > 0$	✗	✗	o desvio padrão deve ser fixado	

Fonte: Autores

Notas:  $\Delta(a, b) = g_i(a) - g_i(b)$

compostos (ICs) e abordagem multicritérios, com a realização de testes para verificar aspectos psicológicos e físicos, bem como conhecimento de sinais rodoviários. Para a agregação e atribuição de pesos a cada subindicador que afeta a qualidade e a confiabilidade dos ICs, foram utilizados os métodos de Análise Envoltória de Dados (DEA) e MCDA, que serviram para elaborar um ranking dos motoristas em termos de condução. Vale destacar que o autor concluiu que o PROMETHEE II foi um método MCDA bastante robusto em termos de resultados.

- Kukadapwar e Parbat [25] elaboraram uma metodologia baseada em decisões multicritérios para encontrar o caminho ideal na rede rodoviária urbana, considerando sete variáveis que influenciam a rede de tráfego e definem a função de custo. O método AHP foi empregado para apoiar a melhor decisão com o caminho de menor custo (caminho ideal) e considerada uma ferramenta eficaz para lidar com o problema de seleção de rota.
- Hu et al. [26] analisaram o risco de trânsito ferroviário urbano na China considerando as fontes de perigo das ferrovias, aspectos de avaliação como frequência, extensão do dano e descontrole, e a ordem de diferentes fontes de risco. Foi utilizado o método Dempster-Shafer/Analytic Hierarchy Process (DS/AHP) melhorado para a avaliação de risco em um ambiente de informação incompleta e incerta.
- Zhao et al. [27] avaliaram a segurança rodoviária através de um indicador formado por fatores pontuados na

escala Lee e utilizou o método de AHP para estabelecer os pesos dos indicadores. Os resultados verificados com os casos de avaliação baseados no modelo AHP indicam maior eficácia e precisão no gerenciamento de segurança de manutenção de estradas.

- Zhang et al. [28] realizaram estudo sobre o índice de avaliação do transporte rodoviário de baixo carbono na China, por meio do método Fuzzy com AHP e Delphi, considerando fatores como infraestrutura rodoviária, tráfego rodoviário, gestão rodoviária e políticas e regulamentos relacionados, para os quais foi estabelecida uma matriz de julgamento com vetores de peso e avaliada a consistência do índice.
- He e Du [29] propuseram um modelo quantitativo de classificação de emergência durante o acidente, considerando fatores como a garantia de trânsito de resgate, emergência e impacto na sociedade, cujos pesos foram determinados através do processo hierárquico analítico difuso (Fuzzy AHP).
- Korchagin et al. [30] desenvolveram uma abordagem objetiva para tomar decisões com o objetivo de reduzir o número de feridos e danos provocados por acidentes de trânsito, com a criação de um sistema baseado no conhecimento de especialistas para melhorar a segurança rodoviária. A sugestão do autor é que a teoria dos conjuntos difusos (fuzzy logic) pode ser uma ferramenta de pesquisa mais racional considerando aspectos como o comportamento dos participantes do trânsito e as condições das rodovias.

### III. METODOLOGIA

Dentre os diversos métodos existentes para o apoio multicritério decidiu-se, após a revisão bibliográfica, utilizar os métodos AHP e PROMETHEE como ferramentas de apoio à decisão, pois ambos fazem comparação das alternativas par a par. Após a obtenção da aplicação de ambos métodos, faz-se uma comparação dos resultados com a metodologia UPSm, que é utilizada pela maioria dos Detrans do país.

Para se atingirem os objetivos estipulados, foram coletadas informações no DETRAN do Distrito Federal - DF sobre acidentes de trânsito com feridos e com mortes ocorridos em suas vias no ano de 2016. Esses dados são oriundos do Sistema de Acidentes de Trânsito (SAT-DF) que envolve diversos órgãos do Governo do DF como a Polícia Civil, a Polícia Militar e Secretaria de Saúde, dando maior credibilidade às informações divulgadas. O processamento das bases, tratamento de variáveis e realização dos cálculos para aplicação dos métodos AHP e PROMETHEE foram feitos no programa Statistical Analysis System (SAS). As tabelas e gráficos foram realizados com o aplicativo Microsoft Excel.

#### A. OBJETIVO, CRITÉRIOS, SUBCRITÉRIOS E ALTERNATIVAS

Um acidente de trânsito pode ser do tipo sem vítima ou com vítima, sendo o último subdividido em com ferido e com morte. Os dados disponibilizados não tratam de acidentes sem vítima, o que levou a adoção dos critérios com morte e com ferido.

Para completar os critérios foi necessário um atributo que revelasse mais detalhe a respeito do acidente, nesse intuito a variável natureza foi escolhida como subcritério, pois ela caracteriza perfeitamente o acidente de trânsito. O atributo em questão foi estruturado de forma a atender a modelagem, sendo necessário em alguns casos fazer um agrupamento deles, conforme apresentado na Tabela 3:

Tabela 3. Naturezas de acidentes de trânsito

Natureza do acidente	Sigla
Atropelamento de animal	ATA
Atropelamento de pedestre	ATP
Capotamento	CAP
Choque com objeto fixo	CHO
Colisão	COL
Queda de Pessoa	QPE
Queda de Veículo	QVE
Demais tipos*	DMA

Fonte: Detran/DF

Nota: \* Demais naturezas não listadas nesta tabela.

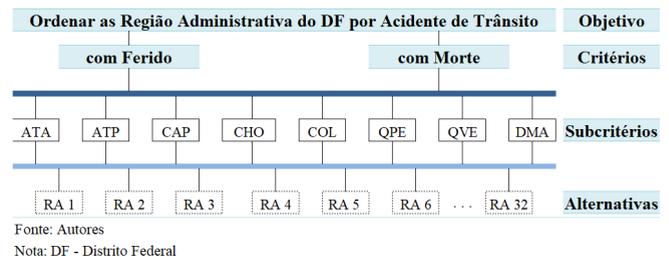
Para os casos em que havia mais de uma natureza para o acidente, ficou estabelecida a primeira natureza para termos de classificação, visto que esta natureza gerou as demais. Por exemplo: em uma colisão seguida por um atropelamento e um capotamento, a natureza deste acidente é classificada como colisão (ANUÁRIO ESTATÍSTICO) [3].

O Distrito Federal possui vias urbanas e rodovias, aquelas são controladas pelo DETRAN-DF e estas pelo Departamento de Estradas e Rodagem (DER). O trabalho foi delimitado para o estudo das vias urbanas. Porém, o foco principal não é a via em si, mas a Região Administrativa (RA) onde ela está localizada. Com isso, tem-se o objetivo que é ordenar as RA's no contexto de acidentes de trânsito.

Com a exposição do problema apresentado, passa-se à especificação dentro dos métodos escolhidos, primeiro no AHP e em seguida no PROMETHEE.

#### B. APLICAÇÃO DO MÉTODO AHP

Neste método devem ser adotados os critérios, subcritérios e alternativas já descritos anteriormente, que adaptando para o AHP clássico ter-se-á a estrutura mostrada na Figura 1. Deve ficar claro que os subcritérios se repetem para cada critério, e que as alternativas também se repetem para cada subcritério, formando-se um produto cartesiano entre eles.



Fonte: Autores

Nota: DF - Distrito Federal

Figura 1. AHP Hierárquico proposto para este estudo

1) Determinação das matrizes de pesos e vetores de eigen Os pesos dos critérios utilizados foram definidos pelo DENATRAN através do indicador Unidade Padrão de Severidade (UPS) [4] conforme Equação (10).

$$UPS = (\text{n}^\circ \text{ de acidentes sem vítima} * 1) + (\text{n}^\circ \text{ de acidentes com ferido} * 5) + (\text{n}^\circ \text{ de acidentes com morte} * 13) \quad (10)$$

Porém o DETRAN-DF não disponibiliza dados de acidentes de trânsito sem vítima, desta maneira, a Equação 5 sofreu uma pequena alteração conforme mostra a Equação (11).

$$UPS = (\text{n}^\circ \text{ de acidentes com ferido} * 5) + (\text{n}^\circ \text{ de acidentes com morte} * 13) \quad (11)$$

Portanto, a matriz de pesos para os critérios ficou da forma mostrado na Figura 2.

No caso de se ter apenas dois critérios não há necessidade de calcular o índice de inconsistência – IC.

Após essa etapa, calculam-se os vetores para os subcritérios baseados nos critérios com ferido e com morte, conforme estabelecido na Figura 1. Importante notar que a matriz de pesos foi obtida utilizando a ideia do AHP Referenciado, conforme Gomes et al. [31], ou seja, baseado nas frequências relativas de cada item que compõe o respectivo subcritério.

CRITÉRIOS	Com Ferido	Com Morte	Pesos
Com Ferido	1	5/13	5
Com Morte	13/5	1	13

Após 2 multiplicações

CRITÉRIOS	Com Ferido	Com Morte	Vetor de Eigen
Com Ferido	2,000	0,769	0,27778
Com Morte	5,200	2,000	0,72222

Fonte: Autores

Figura 2. Cálculo do vetor de eigen para os critérios: com morte e com ferido

Para melhor entendimento do cálculo dos pesos apresentados na Tabela 4, considere o subcritério ATA (frequência = 11) e CHO (frequência = 335), o peso relativo considerado é calculado pela divisão da frequência de ATA pela de CHO (11/335 = 0,033).

Tabela 4. Matriz de pesos baseados na frequência – com ferido

Subcritérios	ATA	ATP	CAP	CHO	COL	QPE	QVE	DMA	Freq
ATA	1,000	0,009	0,162	0,033	0,003	0,115	0,028	0,647	11
ATP	112,455	1,000	18,191	3,693	0,287	12,885	3,156	72,765	1,237
CAP	6,182	0,055	1,000	0,203	0,016	0,708	0,173	4,000	68
CHO	30,455	0,271	4,926	1,000	0,078	3,490	0,855	19,706	335
COL	391,636	3,483	63,353	12,860	1,000	44,875	10,990	253,412	4,308
QPE	8,727	0,078	1,412	0,287	0,022	1,000	0,245	5,647	96
QVE	35,636	0,317	5,765	1,170	0,091	4,083	1,000	23,059	392
DMA	1,545	0,014	0,250	0,051	0,004	0,177	0,043	1,000	17

Fonte: Autores

Para obtenção do vetor eigen final, a matriz de pesos deve ser elevada ao quadrado e normalizada para obtenção do 1º vetor de eigen. Em continuação, pega-se o resultado da primeira multiplicação da matriz e eleva ao quadrado novamente e normaliza para obter o 2º vetor de eigen. Em seguida, compara-se o 1º vetor de eigen calculado com o 2º, se não houver diferença entre eles (diga-se até quatro casas decimais), encerra-se o cálculo; caso contrário, prossegue o processo até a diferença pré-estabelecida entre os vetores eigen ser atendida.

Para a obtenção do vetor de eigen da matriz de subcritérios, considerando o critério com ferido, foram necessárias duas iterações, o resultado é mostrado na Tabela 5.

Tabela 5. Cálculo do vetor de eigen para os subcritérios – com ferido

Subcritérios	ATA	ATP	CAP	CHO	COL	QPE	QVE	DMA	Vetor de Eigen
ATA	8,00	0,07	1,29	0,26	0,02	0,92	0,22	5,18	0,00170
ATP	899,64	8,00	145,53	29,54	2,30	103,08	25,24	582,12	0,19137
CAP	49,45	0,44	8,00	1,62	0,13	5,67	1,39	32,00	0,01052
CHO	243,64	2,17	39,41	8,00	0,62	27,92	6,84	157,65	0,05183
COL	3.133,09	27,86	506,82	102,88	8,00	359,00	87,92	2.027,29	0,66646
QPE	69,82	0,62	11,29	2,29	0,18	8,00	1,96	45,18	0,01485
QVE	285,09	2,54	46,12	9,36	0,73	32,67	8,00	184,47	0,06064
DMA	12,36	0,11	2,00	0,41	0,03	1,42	0,35	8,00	0,00263

Fonte: Autores

Utilizando a Equação (6) obtemos o valor 0,00 que é menor que 0,1 o que indica admissibilidade dos pesos atribuídos.

Para os subcritérios do critério com morte, um ajuste na frequência foi necessário, uma vez que não se pode dividir um número por zero.

N	Lambda Máximo	IC	IR	RC	Resultado
8	8	0	1,41	0	ADMISSÍVEL

Fonte: Autores

Figura 3. Matriz de pesos baseados na frequência – com morte

Portanto, somou-se 1 (um) a todos os subcritérios, conforme é mostrado na matriz de pesos da Tabela 6.

Tabela 6. Matriz de pesos baseados na frequência – com morte

Subcritérios	ATA	ATP	CAP	CHO	COL	QPE	QVE	DMA	Freq
ATA	1,000	0,015	0,111	0,042	0,027	0,333	0,100	1,000	1
ATP	66,000	1,000	7,333	2,750	1,784	22,000	6,600	66,000	66
CAP	9,000	0,136	1,000	0,375	0,243	3,000	0,900	9,000	9
CHO	24,000	0,364	2,667	1,000	0,649	8,000	2,400	24,000	24
COL	37,000	0,561	4,111	1,542	1,000	12,333	3,700	37,000	37
QPE	3,000	0,045	0,333	0,125	0,081	1,000	0,300	3,000	3
QVE	10,000	0,152	1,111	0,417	0,270	3,333	1,000	10,000	10
DMA	1,000	0,015	0,111	0,042	0,027	0,333	0,100	1,000	1

Fonte: Autores

Novamente, após duas iterações foi obtido o vetor de eigen mostrado na Tabela 7.

Tabela 7. Cálculo do vetor de eigen para os subcritérios – com morte

Subcritérios	ATA	ATP	CAP	CHO	COL	QPE	QVE	DMA	Vetor de Eigen
ATA	8,00	0,12	0,89	0,33	0,22	2,67	0,80	8,00	0,00662
ATP	528,00	8,00	58,67	22,00	14,27	176,00	52,80	528,00	0,43709
CAP	72,00	1,09	8,00	3,00	1,95	24,00	7,20	72,00	0,05960
CHO	192,00	2,91	21,33	8,00	5,19	64,00	19,20	192,00	0,15894
COL	296,00	4,48	32,89	12,33	8,00	98,67	29,60	296,00	0,24503
QPE	24,00	0,36	2,67	1,00	0,65	8,00	2,40	24,00	0,01987
QVE	80,00	1,21	8,89	3,33	2,16	26,67	8,00	80,00	0,06623
DMA	8,00	0,12	0,89	0,33	0,22	2,67	0,80	8,00	0,00662

Fonte: Autores

Para as alternativas o processo de cálculo dos vetores de eigen é semelhante aos dos subcritérios. Omitiram-se os cálculos intermediários pelo fato de se terem muitas alternativas, no caso 32. Mas criou-se um resumo com os vetores de eigen finais para cada subcritério versus alternativa, conforme Tabelas 8 e 9.

Após obtenção dos vetores de eigen (autovalores) dos critérios (Cri), subcritérios (Subc), critérios versus subcritérios e subcritérios versus alternativas (Alt), a pontuação pelo método AHP é obtida multiplicando todos os vetores e somando para cada alternativa, conforme a Equação (12).

$$Pont. Alternativa_k = \sum_{i=1}^2 \sum_{j=1}^8 Crit_i \times Subc_j \times Alt_k \quad (12)$$

Em que:

- $i = 1$  (com ferido) e  $2$  (com morte). Os subcritérios são  $j = 1$  (ATA),  $2$  (ATP),  $3$  (CAP),  $4$  (CHO),  $5$  (COL),  $6$  (QPE),  $7$  (QVE) e  $8$  (DMA). As alternativas  $k$  são as Regiões Administrativas conforme especificado na Tabela 13.

Dessa maneira, a pontuação da Região Administrativa de Brasília (BSB) foi obtida como mostra a Tabela 10. Cabe

**Tabela 8.** Vetores de Eigen para alternativas dos subcritérios do critério com ferido

ALTERNATIVA	SUBCRITÉRIOS - com Ferido							
	ATA	ATP	CAP	CHO	COL	QPE	QVE	DMA
ACL	0,0233	0,0355	0,0100	0,0191	0,0390	0,0391	0,0307	0,0204
BAN	0,0465	0,0055	0,0100	0,0027	0,0053	0,0078	0,0094	0,0204
BRZ	0,0698	0,0079	0,0200	0,0245	0,0123	0,0547	0,0165	0,0408
BSB	0,0233	0,1757	0,1700	0,1826	0,1657	0,1328	0,1085	0,0816
CDG	0,0233	0,0032	0,0100	0,0054	0,0016	0,0234	0,0024	0,0204
CEI	0,0698	0,1946	0,1200	0,1662	0,1567	0,2188	0,1745	0,0408
CRZ	0,0233	0,0055	0,0100	0,0136	0,0107	0,0156	0,0094	0,0408
EST	0,0233	0,0032	0,0100	0,0027	0,0037	0,0078	0,0118	0,0612
FER	0,0233	0,0016	0,0100	0,0027	0,0002	0,0078	0,0047	0,0204
GAM	0,0698	0,0591	0,1100	0,0654	0,0578	0,0234	0,0566	0,0408
GUA	0,0233	0,0339	0,0100	0,0409	0,0279	0,0234	0,0236	0,0204
ITA	0,0465	0,0181	0,0100	0,0054	0,0107	0,0156	0,0165	0,0204
JBO	0,0233	0,0055	0,0200	0,0109	0,0060	0,0078	0,0047	0,0204
LGN	0,0233	0,0008	0,0100	0,0109	0,0044	0,0078	0,0094	0,0204
LGS	0,0233	0,0024	0,0100	0,0191	0,0095	0,0078	0,0071	0,0204
NOR	0,0233	0,0032	0,0100	0,0027	0,0007	0,0078	0,0047	0,0204
PAR	0,0233	0,0189	0,0200	0,0027	0,0142	0,0078	0,0071	0,0204
PLA	0,0698	0,0496	0,0200	0,0518	0,0724	0,0234	0,0708	0,0204
PWY	0,0233	0,0024	0,0100	0,0136	0,0070	0,0078	0,0071	0,0204
REM	0,0233	0,0362	0,0400	0,0463	0,0367	0,0313	0,0401	0,0408
RF1	0,0233	0,0095	0,0200	0,0082	0,0088	0,0078	0,0047	0,0408
RF2	0,0233	0,0055	0,0100	0,0082	0,0056	0,0234	0,0118	0,0204
SAI	0,0233	0,0095	0,0100	0,0109	0,0086	0,0156	0,0142	0,0204
SAM	0,0233	0,0820	0,0800	0,0845	0,0659	0,0703	0,0896	0,0612
SB1	0,0233	0,0087	0,0400	0,0191	0,0230	0,0234	0,0236	0,0204
SB2	0,0233	0,0150	0,0100	0,0136	0,0128	0,0234	0,0189	0,0204
SMA	0,0233	0,0307	0,0100	0,0681	0,0399	0,0156	0,0542	0,0408
SOC	0,0233	0,0055	0,0300	0,0082	0,0077	0,0078	0,0071	0,0204
SSE	0,0233	0,0339	0,0500	0,0136	0,0369	0,0234	0,0590	0,0204
TAG	0,0233	0,1198	0,0800	0,0681	0,1332	0,0859	0,0731	0,0612
VAR	0,0233	0,0055	0,0100	0,0027	0,0016	0,0078	0,0047	0,0204
VPI	0,0465	0,0118	0,0100	0,0054	0,0135	0,0234	0,0236	0,0408

Fonte: Autores

salientar que todos os valores para efetivação do cálculo estão distribuídos nas figuras com vetores de eigen.

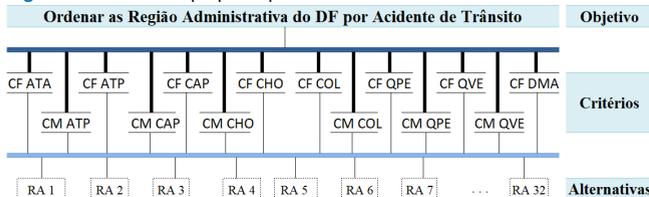
O resultado com a ordenação das alternativas é apresentado na Tabela 13 em que se destacam as seguintes regiões administrativas:

- Primeiras posições: Brasília (BSB) e Ceilândia (CEI).
- Últimas posições: Noroeste (NOR) e Fercal (FER).

A seção seguinte descreve a aplicação do método Promethee e cálculo da UPSm.

### C. APLICAÇÃO DO PROMETHEE

**Figura 4.** Promethee proposto para este estudo



Fonte: Autores  
Nota: DF - Distrito Federal

Usando uma estrutura semelhante ao AHP, porém transformando a hierarquia critério-subcritério para uma estrutura de somente critério conforme mostra a Figura 4 e Tabela 10.

**Tabela 9.** Vetores de Eigen para alternativas dos subcritérios do critério com morte

ALTERNATIVA	SUBCRITÉRIOS - com Morte							
	ATA	ATP	CAP	CHO	COL	QPE	QVE	DMA
ACL	0,0313	0,0103	0,0250	0,0182	0,0441	0,0588	0,0488	0,0313
BAN	0,0313	0,0103	0,0250	0,0182	0,0294	0,0294	0,0244	0,0313
BRZ	0,0313	0,0103	0,0250	0,0545	0,0147	0,0294	0,0488	0,0313
BSB	0,0313	0,2062	0,1000	0,0545	0,1324	0,0294	0,0488	0,0313
CDG	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
CEI	0,0313	0,1546	0,0250	0,1273	0,0441	0,0588	0,0732	0,0313
CRZ	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
EST	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
FER	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
GAM	0,0313	0,0309	0,0500	0,0364	0,0147	0,0294	0,0244	0,0313
GUA	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
ITA	0,0313	0,0309	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
JBO	0,0313	0,0103	0,0500	0,0182	0,0294	0,0294	0,0244	0,0313
LGN	0,0313	0,0103	0,0250	0,0182	0,0294	0,0294	0,0244	0,0313
LGS	0,0313	0,0103	0,0250	0,0182	0,0294	0,0294	0,0244	0,0313
NOR	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
PAR	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
PLA	0,0313	0,0309	0,0250	0,0364	0,0294	0,0294	0,0488	0,0313
PWY	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0488	0,0313
REM	0,0313	0,0412	0,0250	0,0545	0,0735	0,0294	0,0244	0,0313
RF1	0,0313	0,0206	0,0250	0,0182	0,0147	0,0294	0,0488	0,0313
RF2	0,0313	0,0206	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
SAI	0,0313	0,0206	0,0250	0,0545	0,0294	0,0294	0,0244	0,0313
SAM	0,0313	0,0515	0,1000	0,0364	0,0441	0,0294	0,0244	0,0313
SB1	0,0313	0,0103	0,0250	0,0182	0,0441	0,0294	0,0244	0,0313
SB2	0,0313	0,0412	0,0250	0,0545	0,0147	0,0294	0,0244	0,0313
SMA	0,0313	0,0206	0,0250	0,0909	0,0588	0,0294	0,0488	0,0313
SOC	0,0313	0,0103	0,0250	0,0182	0,0294	0,0294	0,0244	0,0313
SSE	0,0313	0,0206	0,0250	0,0182	0,0294	0,0294	0,0244	0,0313
TAG	0,0313	0,1134	0,0250	0,0182	0,0735	0,0294	0,0244	0,0313
VAR	0,0313	0,0103	0,0250	0,0182	0,0147	0,0294	0,0244	0,0313
VPI	0,0313	0,0206	0,0250	0,0182	0,0294	0,0294	0,0244	0,0313

Fonte: Autores

**Tabela 10.** Cálculo da pontuação para a Região Administrativa de Brasília (BSB)

Critério item	autovalor	Subcritério		Alternativa		Pontual parcial
		item	autovalor	item	autovalor	
com ferido	0,277778	ATA	0,001702	BSB	0,023256	0,000011
		ATP	0,191368		0,175729	0,009341
		CAP	0,010520		0,170000	0,000497
		CHO	0,051825		0,182561	0,002628
		COL	0,666460		0,165738	0,030683
		QPE	0,014851		0,132813	0,000548
		QVE	0,060644		0,108491	0,001828
		DMA	0,002630		0,081633	0,000060
com morte	0,722222	ATA	0,006623	BSB	0,031250	0,000149
		ATP	0,437086		0,206186	0,065087
		CAP	0,059603		0,100000	0,004305
		CHO	0,158940		0,054545	0,006261
		COL	0,245033		0,132353	0,023422
		QPE	0,019868		0,029412	0,000422
		QVE	0,066225		0,048780	0,002333
		DMA	0,006623		0,031250	0,000149
<b>Pontuação final</b>					<b>0,147725</b>	

Fonte: Autores

Nota: Vetor Eigen = autovalor

As duas primeiras letras dos critérios indicados nesta tabela significam: CF (com ferido) e CM (com morte).

O método exige o conhecimento de alguns parâmetros utilizados em sua fase de construção sobre relação de sobre-classificação. Portanto, devem-se identificar as alternativas e

os critérios; e dentro destes, identificar o peso e a respectiva função de preferência conforme informado na Tabela 11.

**Tabela 11.** Alternativas e critérios para aplicação do Promethee

Alternativas	Critérios													
	CF_A TP	CF_C HO	CF_C OL	CF_Q QPE	CF_Q QVE	CM_ COL	CM_ QPE	CM_ QVE	CF_A TA	CF_C AP	CF_Z DM	CM_ CHO	CM_ ATP	CM_ CAP
ACL	44	6	168	4	12	2	1	1						
BAN	6	23		3	1				1					
BRZ	9	8	53	6	6			1	2	1	1	2		
BSB	222	66	714	16	45	8		1		16	3	2	19	3
CDG	3	1	7	2										
CEI	246	60	675	27	73	2	1	2	2	11	1	6	14	
CRZ	6	4	46	1	3						1			
EST	3	16		4							2			
FER	1	1												
GAM	74	23	249	2	23				2	10	1	1	2	1
GUA	42	14	120	2	9									
ITA	22	1	46	1	6				1				2	
JBO	6	3	26	1	1	1				1				1
LGN		3	19	3	1									
LGS	2	6	41	2	1									
NOR	3	3		1										
PAR	23	61		2						1				
PLA	62	18	312	2	29	1		1	2	1		1	2	
PWY	2	4	30	2				1						
REM	45	16	158	3	16	4				3	1	2	3	
RF1	11	2	38	1				1		1	1		1	
RF2	6	2	24	2	4								1	
SAI	11	3	37	1	5	1						2	1	
SAM	103	30	284	8	37	2				7	2	1	4	3
SB1	10	6	99	2	9	2				3				
SB2	18	4	55	2	7							2	3	
SMA	38	24	172	1	22	3		1			1	4	1	
SOC	6	2	33	2	1					2				
SSE	42	4	159	2	24	1				4				1
TAG	151	24	574	10	30	4				7	2		10	
VAR	6	7		1										
VPI	14	1	58	2	9	1			1		1			1

Fonte: Autores

As alternativas correspondem às 32 Regiões Administrativas.

Na definição dos parâmetros e dos critérios envolvendo vítimas com morte, ficou claro que um óbito era suficiente para diferenciar uma alternativa da outra, por isso a função de preferência ficou definida com o critério usual. Para os demais critérios se observou a distribuição dos dados. Na definição dos pesos foi observada a frequência dentro da especificação com morte e com ferido e depois multiplicado pelos pesos especificados na Equação (6).

**Tabela 12.** Parâmetros utilizados para comparar as alternativas

Critério <sub>i</sub>	Função de Preferência	q	p	w <sub>j</sub>
CF_ATA	Critério Usual			5,01
CF_ATP	Área de indiferença	4	10	5,96
CF_CAP	Pseudocritério		3	5,05
CF_CHO	Pseudocritério		5	5,26
CF_COL	Pseudocritério		15	8,33
CF_QPE	Quase-critério	1		5,07
CF_QVE	Área de indiferença	2	6	5,30
CF_ZDM	Critério Usual			5,01
CM_ATP	Critério Usual			18,91
CM_CAP	Critério Usual			13,73
CM_CHO	Critério Usual			15,09
CM_COL	Critério Usual			16,27
CM_QPE	Critério Usual			13,18
CM_QVE	Critério Usual			13,82

Fonte: Autores

Em que :

- $q$  representa o limite de indiferença, o maior valor para  $[g_i(a) - g_i(b)]$  abaixo do qual existe uma indiferença.
- $p$  representa o limite de preferência, o menor valor para  $[g_i(a) - g_i(b)]$  acima do qual existe uma preferência estrita.
- $w_j$  é o peso de cada critério e significa a importância que o critério tem.

Calculam-se as diferenças par a par das alternativas considerando os critérios e parâmetros estabelecidos. Após o cálculo do índice de preferência, fluxos de superação positivo e negativo, chega-se a uma pré-ordem, ou ordem parcial (Promethee I). Para obtenção de uma ordem completa (Promethee II) deve-se calcular o fluxo líquido que é a diferença entre os fluxos positivo e negativo. O resultado deve ser ordenado em ordem decrescente em que a primeira alternativa será a de maior pontuação, conforme Tabela 13, com destaque para as seguintes posições:

- Primeiras posições: Ceilândia (CEI) e Brasília (BSB).
- Últimas posições: Noroeste (NOR) e Fercal (FER).

Para comparação com os métodos AHP e Promethee, foi utilizada a Equação 2 para o cálculo do indicador UPSM que é utilizado pela maioria dos Anuários de Acidente de Trânsito.

#### D. APLICAÇÃO DA UPSM

Para calcular o indicador, substituiu-se o número de acidentes pelo número de vítimas na Equação 11 tornando-se a Equação 13.

$$UPS_M = (\text{n}^\circ \text{ de vítimas com ferido} * 5) + (\text{n}^\circ \text{ de vítimas mortas} * 13) \quad (13)$$

O indicador apresentou semelhanças com AHP em suas ordenações, conforme Tabela 13.

#### IV. RESULTADOS DO ESTUDO DE CASO

O resultado com a ordenação das alternativas é apresentado na Tabela 13.

De maneira geral obteve-se:

- Primeiras posições: Brasília (BSB) e Ceilândia (CEI).
- Últimas posições: Noroeste (NOR) e Fercal (FER).

#### V. DISCUSSÃO

As comparações feitas neste tópico fazem referências à Tabela 13 e à Figura 5.

Como pode ser verificado pela análise da tabela e visualmente da figura, existem de fato alterações na ordenação de desempenho das Regiões Administrativas quando comparados os métodos AHP e Promethee, porém pode-se notar que nas duas primeiras posições as Regiões Administrativas são as mesmas. O mesmo acontece com as cinco últimas posições, havendo coincidência nas duas últimas. Considerando as maiores diferenças, nota-se concentração no centro do ordenamento, onde Brazlândia destaca-se como de maior diferença.

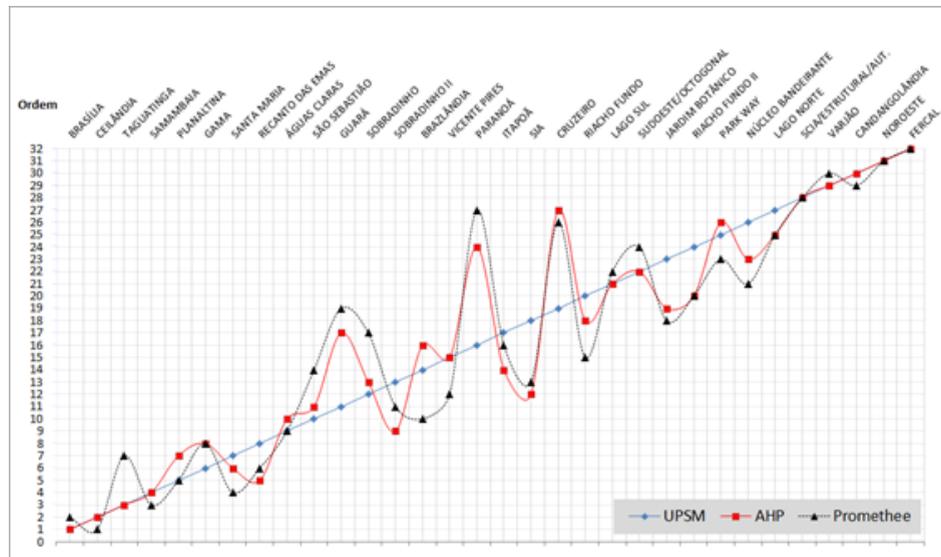


Figura 5. Gráfico comparativo das ordenações: UPSM, AHP e Promethee

Tabela 13. Comparação dos métodos

Sigla	Alternativa Descrição	Ordenação			UPS <sub>M</sub>			AHP			PROMETHEE		
		UPS <sub>M</sub>	AHP	Promethee	Morte	Ferido	valor	valor	Q <sup>+</sup>	Q <sup>-</sup>	Q		
CEI	CEILÂNDIA	2	2	1	325	5475	5800	0,12	26,26	3,43	22,83		
BSB	BRASÍLIA	1	1	2	429	5410	5839	0,15	25,32	3,09	22,24		
SAM	SAMAMBAIA	4	4	3	130	2355	2485	0,05	20,73	5,69	15,05		
SMA	SANTA MARIA	7	6	4	117	1290	1407	0,04	17,44	6,13	11,32		
PLA	PLANALTIMA	5	7	5	65	2130	2195	0,04	17,03	6,74	10,28		
REM	RECANTO DAS EMAS	8	5	6	117	1210	1327	0,05	16,32	6,06	10,26		
TAG	TAGUATINGA	3	3	7	182	3990	4172	0,09	16,01	5,84	10,17		
GAM	GAMA	6	8	8	52	1920	1972	0,04	16,77	8,02	8,76		
ACL	ÁGUAS CLARAS	9	10	9	52	1170	1222	0,03	13,83	8,32	5,51		
BRZ	BRAZLÂNDIA	14	16	10	39	430	469	0,02	11,23	9,20	2,03		
SB2	SOBRADINHO II	13	9	11	65	430	495	0,03	10,54	8,86	1,67		
VPI	VICENTE PIRES	15	15	12	26	430	456	0,02	9,98	9,02	0,96		
SAI	SIA	18	12	13	52	285	337	0,02	9,77	8,81	0,96		
SSE	SÃO SEBASTIÃO	10	11	14	26	1175	1201	0,03	9,66	8,71	0,96		
RF1	RIACHO FUNDO	20	18	15	26	270	296	0,02	8,21	10,27	-2,06		
ITA	ITAPOÁ	17	14	16	26	385	411	0,02	7,68	10,01	-2,32		
SB1	SOBRADINHO	12	13	17	26	645	671	0,02	7,05	9,76	-2,71		
JBO	JARDIM BOTÂNICO	23	19	18	26	185	211	0,02	6,96	10,97	-4,01		
GUA	GUARÁ	11	17	19		935	935	0,02	5,25	10,91	-5,66		
RF2	RIACHO FUNDO II	24	20	20	13	190	203	0,02	5,22	10,95	-5,74		
BAN	NÚCLEO BANDEIRANTE	26	23	21	13	165	178	0,02	5,06	11,09	-6,03		
LGS	LAGO SUL	21	21	22	13	255	268	0,02	4,35	11,13	-6,78		
PWY	PARK WAY	25	26	23	13	190	203	0,01	4,75	11,62	-6,87		
SOC	SUDOESTE/OCTOGONAL	22	22	24	13	225	238	0,02	4,16	11,07	-6,91		
LGN	LAGO NORTE	27	25	25	13	125	138	0,01	4,06	11,44	-7,38		
CRZ	CRUZEIRO	19	27	26		305	305	0,01	3,27	11,36	-8,09		
PAR	PARANOÁ	16	24	27		435	435	0,01	3,41	11,91	-8,50		
EST	SCLA. ESTRUTURAL AUTOMÓVEL	28	28	28		125	125	0,01	3,29	12,00	-8,71		
CDG	CANDANGOLÂNDIA	30	30	29		65	65	0,01	2,64	12,33	-9,69		
VAR	VARJÃO	29	29	30		70	70	0,01	2,22	12,59	-10,37		
NOR	NOROESTE	31	31	31		35	35	0,01	2,19	12,72	-10,52		
FER	FERCAL	32	32	32		15	15	0,01	2,19	12,82	-10,62		

Fonte: Autores

Quando se compara o AHP com UPS<sub>M</sub>, que é o indicador usado pela maioria dos Detran's do Brasil a exemplo do Anuário Estatístico [3], as diferenças começam somente após a quarta posição. Com relação às últimas posições, a mesma ordem de classificação coincide nas cinco finais. Análise semelhante pode ser feita entre o Promethee e a UPS<sub>M</sub>. De maneira geral, observa-se que as maiores diferenças permanecem no centro da ordenação.

## VI. CONCLUSÃO

O ordenamento das Regiões Administrativas do DF, com base nas estatísticas de acidentes de trânsito, é uma importante informação de apoio ao processo de decisão envolvendo as políticas públicas para um trânsito mais seguro.

Os resultados obtidos mostram o potencial da metodologia MCDA para auxiliar na estruturação do problema e no desenvolvimento dos métodos de apoio à decisão com vários critérios. Para este estudo de caso, os métodos AHP e Promethee mostraram comportamentos semelhantes quanto à ordenação, porém o primeiro foi mais complicado de se estruturar em relação ao segundo.

Uma possível vantagem do AHP sobre o Promethee neste trabalho foi a interpretação que se pode ter com relação aos subcritérios. Supondo que o tomador de decisão queira avaliar as três primeiras posições do Promethee (Ceilândia, Brasília e Samambaia), como seria a definição de prioridade nessas RA's, para a implementação de ações de prevenção de acidentes de trânsito?

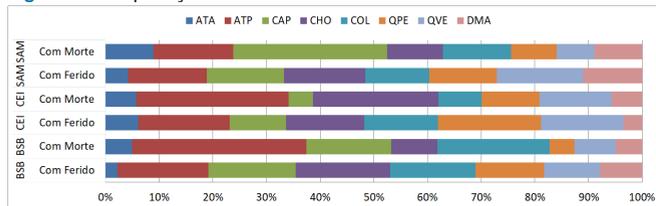
Em Ceilândia e Brasília fica claro que, com relação aos acidentes com morte, a ênfase deve ser dada para os atropelamentos de pedestres, com ações como priorizar estudos e vistorias para identificar os possíveis problemas das vias.

No caso da Samambaia, a prioridade seria a observância das vias que estão ocorrendo capotamentos com morte para verificar o real problema.

O método AHP pode revelar informações importantes com relação à hierarquia critério-subcritério. A Tabela 5 mostra que 66,7% dos acidentes com feridos são derivados de colisão, já para os acidentes com morte Tabela 7, 43,7% são derivados de atropelamento de pedestre.

Este trabalho aplicou dois métodos de MCDA para ordenar as Regiões Administrativas do DF segundo os acidentes de trânsito e apresentou resultados promissores para apoiar o processo de decisão relacionado às políticas, estratégias e

**Figura 6.** Interpretação dos subcritérios calculados no AHP



Fonte: Autores

ações de segurança viária. Como sugestão de trabalhos futuros, seria interessante estudar outros critérios que poderiam influenciar na avaliação dos acidentes de trânsito, como por exemplo, acidentes sem vítima. Além disso, poderiam ser testados métodos de apoio à decisão mais sofisticados, como a lógica nebulosa (Fuzzy Logic).

### Referências

- [1] World Health Organization. Global Status Report on Road Safety. Technical report, WHO, 2015.
- [2] Estatísticas Vitais DATASUS 2016. <http://datasus.saude.gov.br/>. Acessado em 10 de dezembro de 2017.
- [3] Governo do Distrito Federal. Anuário Estatístico de Acidentes de Trânsito. DETRAN, Brasília-DF, Brasil, 2016.
- [4] Ministério da Justiça. Manual de Identificação, Análise e Tratamento de Pontos Negros. Departamento Nacional de Trânsito (DENATRAN), 1987.
- [5] A. M. Novoa, K. Perez, E. Santamarina-Rubio, M. MariDell'Olmo, R. Cozar, J. Ferrando, R. Peiro, A. Tobias, P. Zori, and C. Borrell. Road safety in the political agenda: the impact on road traffic injuries. *Journal of Epidemiology & Community Health*, 2011.
- [6] Guoqing Hu, Timonthy Baker, and Susan P Baker. Comparing road traffic mortality rates from police-reported data and death registration data in China. *Bull World Health Organ*, pages 41-45, 2011.
- [7] Faan Chen, Jianjun Wang, and Yajuan Deng. Road safety risk evaluation by means of improved entropy TOPSIS-RSR. *Safety Science*, 2015.
- [8] Catherine A. Staton, Vijitha De Silva, Elizabeth Krebs, Luciano Andrade, Stephen Rulisa, Badra Chandanie Mallawaarachchi, Kezhi Jin, Joao Ricardo Vissoci, and Truls -stbye. High road utilizers surveys compared to police data for road traffic crash hotspot localization in Rwanda and Sri Lanka. *BMC Public Health*, 2015.
- [9] Lucia Maria Miana Mattos Paixão, Eliane Dias Gontijo, Eliane de Freitas Drumond, Amelia Augusta de Lima Friche, and Waleska Teixeira Caia a. Acidentes de trânsito em Belo Horizonte: o que revelam três diferentes fontes de informações, 2008 a 2010. *Revista Brasileira de Epidemiologia*, 2015.
- [10] Yikai Chen, Kai Wang, Mark King, Jie He, Jianxun Ding, Qin Shi, Changjun Wang, and Pingfan Li. Differences in factors affecting various crash types with high numbers of fatalities and injuries in China. *PLoS ONE*, 2016.
- [11] Firas Hasan and Alwan Asad. Road Traffic Accidents in Iraq: A Review of Evidence - Based Literature. 7(2), 2017.
- [12] Areti Boulieri, Silvia Liverani, Kees de Hoogh, and Marta Blangiardo. A space time multivariate Bayesian model to analyze road traffic accidents by severity. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 2017.
- [13] Erico Veras Marques e Marco Aurélio Machado. Identificação dos fatores relevantes na decisão da alocação dos recursos econômicos visando um trânsito seguro. 44(6):1379-1404, 2010.
- [14] Valerie Belton, Glasgow Scotland, and Theodor J Stewart. *Multiple Criteria Decision Analysis - An Integrated Approach*. 2002.
- [15] Carlos A. Bana e Costa. *Readings in Multiple Criteria Decision Aid*. Springer-Verlag, Berlin, 1st edition, 1990.
- [16] Bernard Roy. *Multicriteria Methodology for Decision Aiding*. Springer Science-Business Media Dordrecht, 1st edition, 1996.
- [17] Saaty, T. L. Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 83 - 98. 2008.

- [18] ISHIZAKA, A.; NEMERY, P. *Multi-Criteria Decision Analysis - Methods and Software*. 1ª. ed. Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom: Wiley, 2013.
- [19] BRANS & MARESCHAL. *Promethee Methods*. In: Figueira et.al. *Multi-criteria Decision Analysis - State of the Art Surveys*. Ed: Springer Science + Business Media, Inc. 2005.
- [20] Annalia Bernardini, Quantin Hayez, Cathy Macharis, Yves De Smet. Multicriteria analysis of policy options scenarios to reduce the aviation climate impact - an application of the PROMETHEE based D-SIGHT software. OR52 Annual Conference, At Royal Holloway University of London, 2010
- [21] Ramon Bauza, Javier Gozalvez, and Joaquin Sanchez-Soriano. Road traffic congestion detection through cooperative Vehicle-to-Vehicle communications. In *Proceedings - Conference on Local Computer Networks, LCN*, 2010.
- [22] Laurence Turckin, Annalia Bernardini, and Cathy Macharis. A combined AHP-PROMETHEE approach for selecting the most appropriate policy scenario to stimulate a clean vehicle fleet. In *Procedia - Social and Behavioral Sciences*, 2011.
- [23] P. Pitchipoo, D. S. Vincent, N. Rajini, and S. Rajakarunakaran. COPRAS decision model to optimize blind spot in heavy vehicles: A comparative perspective. In *Procedia Engineering*, 2014.
- [24] Seddigeh Babae, Maryam Bagherikahvarin, Renaud Sarrazin, Yongjun Shen, and Elke Hermans. Use of DEA and PROMETHEE II to assess the performance of older drivers. In *Transportation Research Procedia*, 2015.
- [25] Surendra Kukadapwar and Dhananjay Parbat. Estimation of Optimal Path on Urban Road Networks Using AHP Algorithm. *International Journal for Traffic and Transport Engineering*, 6(1):13 - 24, 2016.
- [26] Xiaoping Hu, Xiamiao Li, and Yin Huang. Urban Rail Transit Risk Evaluation with Incomplete Information. In *Procedia Engineering*, 2016.
- [27] Jianyou Zhao, Xiaoyu Fu, and Yunjiao Zhang. Research on Risk Assessment and Safety Management of Highway Maintenance Project. In *Procedia Engineering*, 2016.
- [28] Wendan Zhang, Jian Lu, and Yi Zhang. Comprehensive Evaluation Index System of Low Carbon Road Transport Based on Fuzzy Evaluation Method. In *Procedia Engineering*, 2016.
- [29] Yaqin He and Shengpin Du. Classification of Urban Emergency Based on Fuzzy Analytic Hierarchy Process. In *Procedia Engineering*, 2016.
- [30] Viktor Korchagin, Anatoly Pogodaev, Vladimir Kliavin, and Vitali Sinitnikov. Scientific Basis of the Expert System of Road Safety. In *Transportation Research Procedia*, 2017.
- [31] Gomes, L. F. A. M.; Araya, M. C. G.; Carignano, C. *Tomada de decisão em cenários complexos: introdução aos métodos discretos do apoio multicritério à decisão*. São Paulo: Pioneira Thomson Learning, 2004.



**JOÃO VICENTE PEREIRA** Funcionário do Banco do Brasil S.A, atua na área de gerenciamento de riscos e de capital, é graduado em Estatística pela Universidade de Brasília (UnB), e em Ciências Contábeis pela Fundação Universidade Tocantins, com pós-graduação em Estatística pela Universidade Paulista, e MBA em Gestão de Crédito pela Fundação Getúlio Vargas (FGV). Trabalhou no Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (INEP) como consultor estatístico de 1998-2004. Trabalhou no Departamento de Trânsito do Distrito Federal (Detran-DF) de 2001-2008 como analista de trânsito.



**RÔMULO DE MEDEIROS PALMEIRA** Possui graduação em Ciências Econômicas pelo Centro Universitário de Brasília - UNICEUB (2001) e pós-graduação em Estatística Aplicada pela Associação de Ensino Unificado do Distrito Federal - AEUDF (2013). Gerenciou microempresa do segmento esportivo no período de 1998 a 2003 (Academia de Ginástica NEW LIFE). É funcionário do Banco do Brasil desde abril de 2000 com a seguinte experiência: Escriturário na Agência Setor Militar Urbano, de 2000 a 2003, exercendo atividades de atendimento aos clientes do segmento pessoa física; Analista Júnior na Diretoria de Crédito, na Divisão de Apoio à Gestão, de 2003 a 2005, apoiando a gestão dos recursos humanos, materiais, tecnológicos e orçamentários da dependência; Assessor Pleno na Diretoria de Crédito, na Gerência de Acompanhamento do Risco de Crédito, de 2005 a 2013, com experiência em análise de risco setorial e de precificação do risco de crédito; Assessor Empresarial no Projeto Basileia II Risco de Crédito, na Divisão de Integração dos Parâmetros de Risco IRB na Gestão, de 2013 a 2016, com experiência nos processos de análise do Escopo IRB, estudos de impacto do cálculo de capital (QIS IRB), estruturação/homologação de dados corporativos para capital, modelagem da volatilidade para capital econômico e apoio à integração na gestão dos subprojetos de capital regulatório/econômico, teste de estresse, painel de risco e monitoramento de modelos. Exerce desde janeiro de 2017 a função de Assessor Empresarial na Diretoria de Gestão de Riscos, Gerência de Modelagem do Risco de Crédito.

Assessor Empresarial no Projeto Basileia II Risco de Crédito, na Divisão de Integração dos Parâmetros de Risco IRB na Gestão, de 2013 a 2016, com experiência nos processos de análise do Escopo IRB, estudos de impacto do cálculo de capital (QIS IRB), estruturação/homologação de dados corporativos para capital, modelagem da volatilidade para capital econômico e apoio à integração na gestão dos subprojetos de capital regulatório/econômico, teste de estresse, painel de risco e monitoramento de modelos. Exerce desde janeiro de 2017 a função de Assessor Empresarial na Diretoria de Gestão de Riscos, Gerência de Modelagem do Risco de Crédito.



**ANA CARLA B. REIS** Atua na área de Pesquisa Operacional e Gestão da Produção, possui graduação em Engenharia Elétrica pela Universidade Pernambuco (2005), mestrado (2007) e doutorado (2011) em Engenharia de Produção pela Universidade Federal de Pernambuco. Atualmente é professora adjunta do Departamento de Engenharia de Produção da Universidade de Brasília e do mestrado de Computação Aplicada da UnB. Tem experiência na área de Engenharia de Produção,

com ênfase em Pesquisa Operacional, atuando principalmente nos seguintes temas: Modelos de Decisão, Decisão Multicritério, Programação e Controle da Produção e Melhoria de processos.

...

)

# Modelando um esconde-esconde no *Minecraft* utilizando *Hidden Markov Model*

RONALDO E SILVA VIEIRA,  
MATHEUS RODRIGUES LEAL e  
LUIZ CHAIMOWICZ.

Universidade Federal de Minas Gerais (UFMG) (e-mail: {ronaldo.vieira, matheus.leal, chaimo}@dcc.ufmg.br

• **RESUMO** - A criação de agentes inteligentes é uma área da inteligência artificial que busca, dentre outros objetivos, construir entidades capazes de desempenhar ações semelhantes às de seres humanos. Dessa forma, é possível modelar cenários e avaliar a tomada de decisão de entidades envolvidas nestes cenários. Este trabalho consiste na construção de agentes inteligentes para participarem da brincadeira de esconde-esconde. Para isso, é desenvolvida uma abordagem probabilística baseada em um *Hidden Markov Model*, utilizando o jogo *Minecraft* como plataforma para sua aplicação. Os experimentos realizados demonstram que a estratégia desenvolvida permite a interação entre os agentes de forma adequada às regras da brincadeira.

• **PALAVRAS-CHAVE** - Agentes Inteligentes, Sistemas Multiagente, Hidden Markov Model, *Minecraft*.

## I. INTRODUÇÃO

Um objetivo fundamental da inteligência artificial (IA) é criar agentes inteligentes que realizem tarefas ou resolvam problemas de forma autônoma, com eficiência no mínimo equiparável à de seres humanos. Em termos gerais, tais agentes obtêm informações de seu ambiente através de seus sensores e as interpretam inteligentemente para escolher as ações adequadas para atingir seus objetivos [6].

Dessa maneira, não é incomum a proposta de criação de agentes que procurem agir de forma tão similar quanto possível à de seres humanos, em detrimento de estrita eficiência em alcançar algum objetivo. Neste contexto, o presente trabalho apresenta uma tentativa à simulação com múltiplos agentes da brincadeira tradicional de esconde-esconde através de uma abordagem probabilística.

O esconde-esconde inicia com uma pessoa aguardando, de olhos vendados, outras se esconderem pelas redondezas. Ao som de um sinal ou após um tempo acordado previamente, esta pessoa sai à procura das outras. As outras, no entanto, possuem como objetivo chegar sem ser pegas pela primeira no local onde ela aguardou – a base – a fim de "se salvar". Caso alguém seja pego, esta pessoa ajudará seu algoz a encontrar e pegar as outras. A brincadeira geralmente acaba quando todas as pessoas foram pegas ou salvas.

Neste trabalho, uma versão simplificada da brincadeira de esconde-esconde é usada. Nela, apenas dois agentes se movimentam por um cenário, utilizando funções de utilidade para definir seus próximos destinos. Cada agente considera em sua função de utilidade (i) uma estimativa probabilística

da localização do outro agente obtida a partir de um *Hidden Markov Model* (HMM), a fim de possibilitar que os agentes persigam ou evitem o outro; e (ii) a distância do agente em questão para a base, a fim de encorajá-lo a escolher destinos próximos da base para guardá-la ou "se salvar".

A seção II aborda o referencial teórico necessário para o restante do trabalho. Na seção III, alguns trabalhos relacionados são listados. Após, na seção IV, é descrita detalhadamente a metodologia utilizada para reproduzir a brincadeira em meios computacionais utilizando os conceitos de inteligência artificial. Em seguida, na seção V, os resultados obtidos são apresentados e discutidos. Por fim, a Seção VI apresenta as conclusões obtidas a partir dos resultados do trabalho e define etapas futuras que podem ser desenvolvidas para aprimorá-lo.

## II. REFERENCIAL TEÓRICO

No contexto de inteligência artificial, processos Markovianos são quaisquer processos estocásticos cuja determinação do próximo estado depende apenas de seu estado atual. O *Hidden Markov Model* (ou, em português, Modelo Oculto de Markov) se trata de uma forma de modelar um processo cujo estado atual não é observável mas pode ser inferido através de evidências indiretas, se baseando na suposição de que é um processo Markoviano. [6]

Neste trabalho, o HMM é utilizado para inferir a localização do agente adversário a partir de evidências: os locais já observados até o momento. Além da lista de locais possíveis e de evidências, o modelo utiliza como parâmetros

(i) uma distribuição de probabilidades inicial da localização do agente adversário; (ii) uma matriz de transição, que descreve como a distribuição se altera ao longo do tempo; e (iii) uma matriz de observação, que descreve como a distribuição se altera de acordo com novas evidências.

Dessa forma, a cada passo de tempo, utiliza-se de operações matriciais eficientes para atualizar a estimativa de localização do agente adversário. Para a criação do cenário gráfico onde os experimentos são executados, utiliza-se o projeto Malmö.

O projeto Malmö [2] é uma plataforma de código aberto para experimentos de IA construída sobre o jogo *Minecraft*. A plataforma oferece a possibilidade da criação de agentes que atuam como um personagem do jogo em um mundo aberto, dando acesso a informações de sua visão e permitindo ações como andar, pular, olhar para direções específicas ou usar itens.

Embora a plataforma ofereça maior suporte a tarefas de aprendizado por reforço, qualquer aplicação de IA com um ou múltiplos agentes é realizável. Em [9], por exemplo, um agente é treinado com aprendizado por reforço profundo para sobreviver contra uma horda de zumbis, e em [5] outro agente é treinado para escapar de labirintos gerados aleatoriamente.

### III. TRABALHOS CORRELATOS

Com o recente avanço no desenvolvimento de pesquisas na área de inteligência artificial, tornou-se necessária a busca por cenários adequados e desafiadores para experimentação de novas técnicas. Nos últimos anos, um dos cenários que vem sendo utilizados para esta finalidade é o jogo *hide-and-seek* (em português, esconde-esconde). A partir disso, os trabalhos buscam aprimorar as estratégias dos agentes envolvidos no jogo, ou modelar um problema real de acordo com as regras do jogo e utilizar estratégias já conhecidas para resolvê-los.

Em [1], o jogo foi utilizado para estudar a interação entre robôs e humanos, com ênfase na tarefa de busca envolvida no jogo. Para o desenvolvimento da estratégia, foi utilizado um Processo de Decisão de Markov de Observação Mista (MOMDP) com o intuito de reduzir o espaço de busca do problema. A partir disso, foram realizadas algumas simulações para comparar o desempenho de métodos padrão com uma nova heurística também criada pelos autores. Os resultados comprovam que a heurística desenvolvida supera as estratégias concorrentes em praticamente todos os cenários.

Em [3], o jogo também foi utilizado para estudar novas estratégias. No entanto, nesse contexto as características e as regras do jogo foram incrementadas. A principal diferença é relacionada ao papel do agente que se esconde, denominado *Hider*, que precisa desempenhar algumas atividades para conseguir “sobreviver” enquanto tenta fugir. Além disso, todas as atividades desempenhadas por ambos os agente emitem um “ruído”, que pode ser identificado pelo agente adversário. Dado o cenário proposto e após um conjunto de simulações, foi obtida uma solução analítica para todos os casos.

Além de trabalhos cujos objetivos são encontrar melhores formas de jogar o jogo, existem trabalhos focados na modelagem de problemas reais em cenários do jogo, a fim de aplicar as estratégias já conhecidas para resolvê-los. Esse é o caso de [8], que consiste em uma ferramenta de modelagem específica para o esconde-esconde. Essa ferramenta permite definir o cenário do jogo, assim como a quantidade de agentes participantes e os seus respectivos objetivos. Além disso, também fornece suporte para simulação, visualização e análise das estratégias de ambos os tipos de agentes sob diversos ambientes.

### IV. METODOLOGIA

Embora a brincadeira original possa envolver uma quantidade indeterminada de pessoas em ambientes diferentes, aqui limita-se a participação em apenas dois agentes: um para procurar, denominado *Seeker*, e outro para se esconder, denominado *Runner*. Além disso, utiliza-se um cenário 2D em formato de labirinto como local do esconde-esconde. Como plataforma gráfica, utiliza-se o jogo *Minecraft* [4] através da plataforma Malmö.

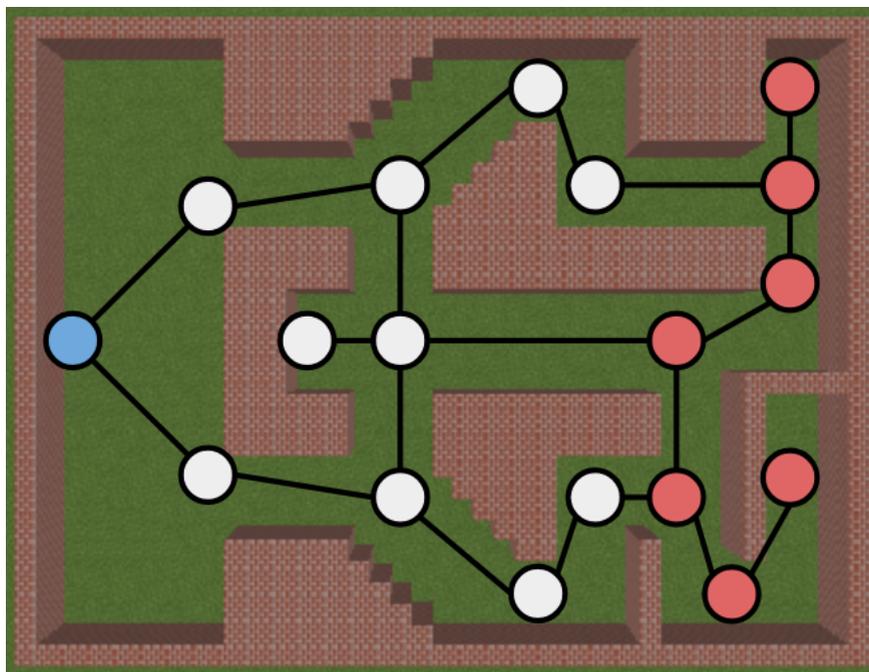
Como forma de discretizar o ambiente contínuo fornecido pelo jogo, foi criado um grafo onde cada nó representa um local possível do cenário e arestas definem quais locais são acessíveis desde quais. Da mesma forma, guarda-se ainda a informação de visibilidade dos locais, isto é, quais locais são visíveis a partir de quais. A Figura 1 mostra o cenário 2D e o grafo propriamente dito.

Na simulação, o agente *Seeker* tem como posição inicial sua base, disposta em uma das extremidades do labirinto, enquanto o agente *Runner* inicia em um dos locais mais próximos da extremidade oposta. O *Seeker* possui como objetivo encontrar o *Runner*, enquanto o *Runner* deseja chegar à base sem ser pego. É importante notar que, embora o cenário tenha sido discretizado, o tempo e a movimentação dos agentes ainda é contínua e “ser pego” é interpretado como ambos os agentes estando perto o suficiente.

Ao iniciar a simulação, os agentes devem escolher sequeencialmente novos locais para ir de acordo com algum critério. Deseja-se que este critério culmine em movimentos inteligentes dos agentes em prol de seus objetivos (chegar à base ou encontrar o outro). Para tal, foi sugerida uma abordagem probabilística a ser descrita a seguir.

Um agente pode se mover apenas para algum dos locais vizinhos do local em que está. Dada essa restrição, permanece ainda a escolha de para qual local ir. E para realizar essa decisão, atribui-se a cada local vizinho um valor de utilidade, isto é, um valor que determina quão bom é ir para um determinado local, segundo o próprio agente com as informações que possui. O local então é escolhido de forma probabilística com valores de probabilidade proporcionais aos respectivos valores de utilidade.

O valor de utilidade de um local para os agentes é calculado com base em onde ele acredita que o outro está e o quão longe o local é da base. Mais especificamente, na probabilidade do outro agente estar no local considerado e



**Figura 1.** Cenário 2D utilizado como ambiente da simulação e grafo de locais. Em azul e vermelho, respectivamente, as posições iniciais dos agentes *Seeker* e *Runner*. O último inicia em uma das posições destacadas escolhida de forma aleatória.

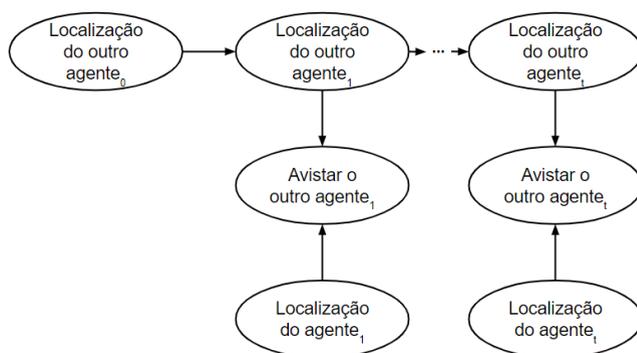
na distância euclidiana deste para a base. Formalmente, o valor de utilidade  $U(s)$  de um estado  $s$  é dado pela seguinte equação:

$$U(s) = \begin{cases} \alpha P(\text{loc. o. ag.} = s) + \frac{\beta}{d(s, \text{base})} & \text{para o Seeker} \\ \alpha P(\text{loc. o. ag.} \neq s) + \frac{\beta}{d(s, \text{base})} & \text{para o Runner} \end{cases}$$

Onde  $\alpha$  e  $\beta$  representam pesos a serem dados aos dois fatores envolvidos no cálculo. É importante notar que, enquanto o *Seeker* deseja encontrar o outro agente, este não deseja ser encontrado. Por isso, o *Runner* considera o inverso da probabilidade utilizada pelo *Seeker*. Por outro lado, ambos utilizam o inverso da distância do local até a base, de forma a encorajá-los a ir para locais próximos da base – um para guardá-la e outro para atingir seu objetivo.

Enquanto o cálculo da distância euclidiana entre locais seja trivial e possa até ser pré-calculado, o mesmo não é válido para a probabilidade, por sua natureza dinâmica. Portanto, para calcular tal probabilidade, foi utilizado um *Hidden Markov Model* [7].

Em sua modelagem, utiliza-se como evidências a localização do agente em questão e dos locais que ele consegue enxergar de acordo com as informações de visibilidade do grafo. Com essas evidências, deseja-se inferir  $P(\text{loc. o. ag.})$ , isto é, a distribuição de probabilidades que indica a localização do outro agente. A Figura 2 apresenta a modelagem em forma gráfica, com a inclusão do fator temporal.



**Figura 2.** Modelo markoviano utilizado para o cálculo da distribuição de probabilidades da localização do agente adversário.

O primeiro dos parâmetros necessários para um HMM é distribuição de probabilidades  $P(\text{loc. o. ag.}_0)$ , isto é, a probabilidade para cada local no início da simulação. Estas são definidas de forma homogênea, pois sem nenhuma evidência, o outro agente pode estar em qualquer um dos dezoito locais com a mesma probabilidade.

Outro parâmetro é o modelo de observação. Ele se refere à aplicação das evidências: a forma como as probabilidades de localização do agente adversário mudam conforme os locais que o agente em questão consegue enxergar. O modelo de observação é definido, informalmente, como:

- (i) Se o agente adversário está em um dos locais que o

agente em questão consegue enxergar, é atribuída probabilidade 1 ao local onde ele está e 0 para todos os outros.

- (ii) Se o agente adversário **não** está em um dos locais que o agente em questão consegue enxergar, é atribuída probabilidade 0 aos locais visíveis pelo agente, enquanto os locais não visíveis mantêm sua distribuição de probabilidades.

Dessa forma, ao ver o outro agente, o modelo se torna certo de sua posição. Em contrapartida, caso não o esteja vendo, pode-se apenas ter certeza de que ele não está nos locais visíveis.

Por último, é necessário definir o modelo de transição – como a distribuição de probabilidades se comporta com o passar do tempo. Em outras palavras, deseja-se calcular  $Ploc. o. ag. t. loc. o. ag. t-1$  e, para tal, é necessário calcular uma quantidade de probabilidades do tipo  $Ploc. o. ag. = s'loc. o. ag = s$ , para todos os estados  $s'$  e  $s$ .

Conforme o tempo passa, é improvável que o agente esteja no mesmo local e mais provável que esteja em algum de seus vizinhos, com peso maior para os vizinhos mais próximos. Dessa forma, o modelo de transição possui o comportamento de decair a probabilidade do agente estar em um local, se ele estava nele anteriormente, e elevar a probabilidade do agente estar em um estado vizinho. Esta alteração se dá de forma proporcional à distância entre os dois locais em consideração.

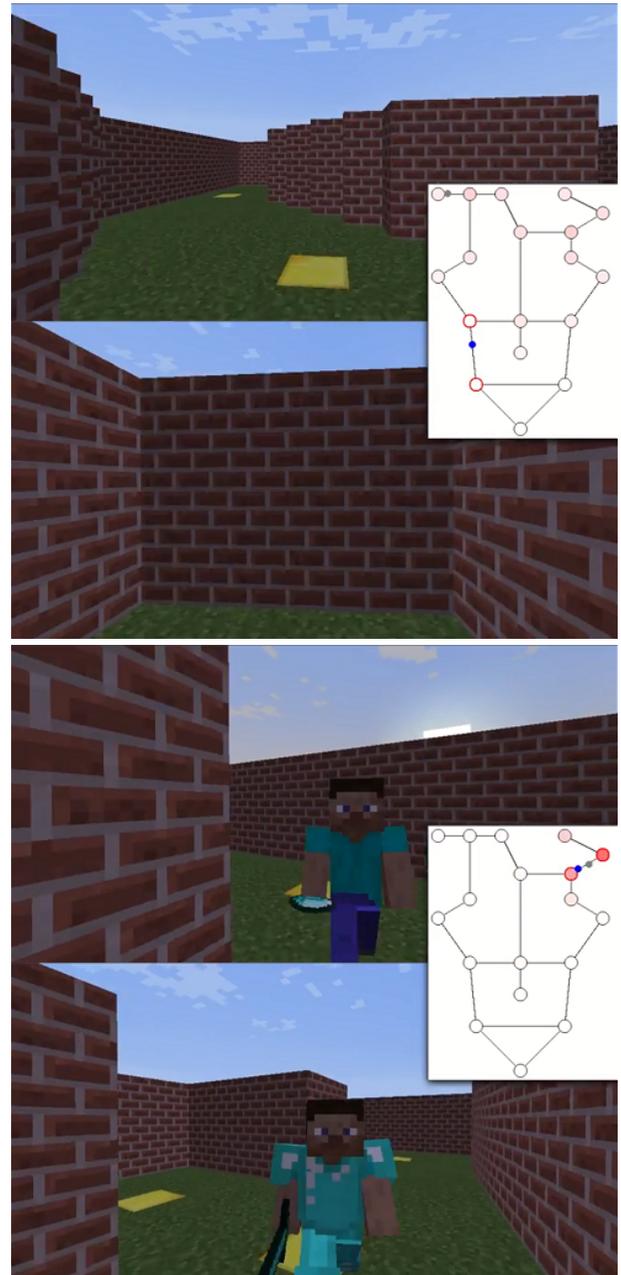
Com a probabilidade inicial e os modelos de observação e transição especificados, todos os componentes necessários para a execução do algoritmo estão definidos. No início da execução, os agentes são dispostos no labirinto em suas respectivas posições iniciais e recebem como conhecimento inicial sobre a localização do adversário uma distribuição de probabilidades uniforme sobre todos os estados. Os agentes realizam os seguintes passos:

- 1) Cada agente verifica se o adversário está em seu campo de visão.
- 2) Caso esteja, a matriz do modelo de observação é definida para indicar a certeza da localização do adversário em um estado específico.
- 3) Se não, a matriz do modelo de observação é definida para indicar a certeza de que o adversário não está nos estados vizinhos.
- 4) Em seguida, a matriz do modelo de transição é utilizada para atualizar a distribuição de probabilidades de acordo com a possível movimentação do adversário.
- 5) Por último, a matriz do modelo de observação é multiplicada pela distribuição de probabilidades resultante, atualizando o conhecimento do agente de acordo com a sua observação.
- 6) A partir disso, são calculados os valores de utilidade dos estados vizinhos. A próxima localização dos agentes é então selecionada probabilisticamente.

A cada iteração, estes passos são repetidos até que o *Seeker* alcance o *Runner* ou o *Runner* consiga chegar à base.

## V. RESULTADOS

Pela natureza do problema, uma análise qualitativa foi realizada para avaliação da simulação. Esta análise foi executada diversas vezes e observações de comportamentos adequados ou inadequados foram notadas. Os parâmetros  $\alpha$  e  $\beta$  que culminaram em maior realismo foram, respectivamente, 0,9 e 0,1. A Figura 3 mostra alguns momentos das simulações. Duas execuções representativas de ambos os agentes completando seus objetivos está disponível em <https://youtu.be/SkSCwN-mF44> e [https://youtu.be/Er4B11D6\\_3E](https://youtu.be/Er4B11D6_3E).



**Figura 3.** Dois momentos capturados nos experimentos. Em cada momento, as visões de cima e de baixo são, respectivamente, do *Seeker* e do *Runner*. A janela mais à direita contém uma representação visual do grafo de locais, probabilidades e posicionamento dos agentes.

Pôde-se perceber que a aplicação da abordagem probabilística proposta resulta em dois comportamentos desejados: o agente *Seeker* procurando ativamente o *Runner* e este, por sua vez, procurando fugir – principalmente quando visto. No entanto, o processo de fuga frequentemente termina no agente encurralado por seu adversário, apesar de ser encorajado a ir em direção da base.

Percebe-se que a fonte de alguns comportamentos inadequados como fugir em direção a um beco sem saída e não aproveitar chances de chegar próximo à base advém da capacidade do agente de apenas considerar probabilidades localmente, isto é, não levar em consideração a probabilidade dos vizinhos de um local para calcular sua probabilidade.

Além disso, por ser uma abordagem probabilística, é incomum, porém possível, que os agentes tomem decisões claramente ruins. Por outro lado, este pode ser um fator benéfico ao que a abordagem almeja: realismo em vez de perfeição. Dado que na maioria absoluta das vezes o comportamento esperado dos agentes seria também o comportamento esperado de seres humanos na brincadeira.

## VI. CONCLUSÃO

Foi proposta uma abordagem probabilística para simular uma brincadeira de esconde-esconde com agentes inteligentes. Modelou-se um cenário 2D em formato de labirinto com estratégias para discretizar o espaço e o tempo. Os agentes no cenário se movimentam em função de um cálculo de qual é o melhor local para ir que, por sua vez, utiliza um *Hidden Markov Model* e cálculos de distâncias.

A simulação atinge os resultados esperados, embora faça emergir algumas seqüências de comportamentos improváveis em um cenário real. Uma melhoria natural para a escolha do próximo local a ir é a utilização de um Processo de Decisão de Markov para que se efetuem cálculos de utilidade de forma global.

Além disso, apesar da simplicidade obtida ao discretizar o espaço através de um grafo, isto restringe a quantidade de movimentos possíveis dos agentes e também acarreta no comportamento improvável de nunca mudar de direção uma vez que inicia um percurso entre um local e outro. Um próximo passo sugerido é ampliar o escopo da simulação para utilizar mais agentes. Este passo naturalmente necessitaria de implementações de políticas de cooperação e comunicação entre agentes, como descrito em [10].

O código-fonte deste trabalho está disponível em <https://github.com/j-ufmg/mine-and-seek>.

## Referências

- [1] A. Goldhoorn, A. Sanfeliu, and R. Alquézar. Comparison of momdp and heuristic methods to play hide-and-peek. In CCIA, 2013.
- [2] M. Johnson, K. Hofmann, T. Hutton, and D. Bignell. The malmo platform for artificial intelligence experimentation. In IJCAI, pages 4246–4247, 2016.
- [3] G. McCormick and G. Owen. A composite game of hide and seek. Central European Journal of Operations Research, 27(1):1–14, Mar 2019.
- [4] Mojang. Minecraft, 2009.
- [5] M. Monfort, M. Johnson, A. Oliva, and K. Hofmann. Asynchronous data aggregation for training end to end visual control networks. In Proceedings

of the 16th Conference on Autonomous Agents and MultiAgent Systems, pages 530–537. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

- [6] S. J. Russell and P. Norvig. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited., 2016.
- [7] R. L. Stratonovich. Conditional markov processes. Theory of Probability & Its Applications, 5(2):156–178, 1960.
- [8] A. Tandon and K. Karlapalem. Medusa: Towards simulating a multi-agent hide-and-peek game. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 5871–5873. International Joint Conferences on Artificial Intelligence Organization, 7 2018.
- [9] H. Udagawa, T. Narasimhan, and S.-Y. Lee. Fighting zombies in minecraft with deep reinforcement learning. Technical report, Technical report, Stanford University, 2016.
- [10] M. Wooldridge. An introduction to multiagent systems. John Wiley & Sons, 2009.



Inteligência Artificial Aplicada a Jogos e Aprendizado por Reforço.

**RONALDO E SILVA VIEIRA** Possui graduação em Ciência da Computação pela Universidade Federal Rural do Rio de Janeiro (2017). Atualmente é mestrando em Ciência da Computação pela Universidade Federal de Minas Gerais, pesquisador, sob a orientação do professor Luiz Chaimowicz, no Laboratório de Inteligência Artificial Aplicada a Jogos do Departamento de Ciência da Computação na mesma instituição. Tem experiência na área de Inteligência Artificial, com ênfase em



com ênfase em Processamento de Linguagem Natural e Representação do Conhecimento.

**MATHEUS RODRIGUES LEAL** Possui graduação em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (2017). Atualmente é mestrando em Ciência da Computação pela Universidade Federal de Minas Gerais, pesquisador, sob a orientação da professora Gisele Lobo Pappa, no Laboratório de Inteligência Computacional do Departamento de Ciência da Computação na mesma instituição. Tem experiência na área de Ciência dos Dados e Aprendizado de Máquina,



aplicada a jogos.

**LUIZ CHAIMOWICZ** Professor Associado do DCC - UFGM e Bolsista de Produtividade em Pesquisa Nível 2 do CNPq. Ele é doutor em Ciência da Computação pela UFGM (2002) e realizou um pós-doutorado em Robótica na Universidade da Pennsylvania, EUA (2003-2004). Suas áreas de interesse incluem robótica móvel, coordenação de múltiplos robôs, plataformas para programação e simulação de robôs móveis e desenvolvimento de jogos digitais, especialmente inteligência artificial

...

# Classifying Descriptions of Goods with Artificial Neural Networks

VINICIUS DI OLIVEIRA and  
MARCELO LADEIRA.

Computer Science Department, University of Brasília-(UnB), P.O. Box 4466, 70910-900, Brasília-DF, Brazil  
(e-mail:vinidiol@gmail.com, mladeira@unb.br)

Autor Correspondente: Vinicius Di Oliveira (e-mail: vinidiol@gmail.com).

• **ABSTRACT** - The present study aims to evaluate the performance of an artificial neural network in the classification of merchandise descriptions indicated in electronic bills, legal document used to record all commercial transactions in Brazil. For this, a significant sample of the actual descriptions will be used as well as a overlook about the performance of the neural network with a KNN and a GBM algorithms forecasting the category of the merchandise each description refers. This paper brings a method for classifying descriptions of goods with Artificial Neural Networks. The descriptions are small non structured texts, maximum of 120 characters, relating to goods traded in commercial transactions.

• **KEYWORDS** - Text mining, classifying, neural networks

## I. INTRODUCTION

This paper proposes to study a way to classify goods indicated in electronic invoices issued in Brazil (NFe) in order to contribute to tax inspection by automating the analysis of these notes. The development of a computational method capable of identifying the tax classification to which a merchandise belongs in an electronic invoice (NFe) would bring a significant gain of efficiency to the tax administration. Artificial neural networks have been shown to be an effective tool in the classification of texts observed in the field of text mining. In the case studied, the good description is presented in the NFe in a string-type field limited to 200 characters, but free filled, so the ways of describing the same good vary substantially according to the NFe issuer.

The methodology adopted was CRISP-DM, as it is widely known and used in the academy as well as in the industry. A real database was provided by the tax administration of Distrito Federal/Brazil was made up of 15,000 NFe records, 5,000 from each classification group, regarding to invoices issued in 2017. At the end of the study, a satisfactory result in the classification (AUC = 0.95) was achieved, matching the methodology officially used by the state currently. The tax administrations of the Brazilian states have sought to improve their control processes in a substantial way, this study aims to contribute to the continuous search for improvement. In this sense, the development of an artificial intelligence capable of reading an invoice and indicates the correct tax treatment to be dispensed for that merchandise, this study aims to present

the “starting point” in that reach.

## II. STATE OF ART REVIEW

### A. CRISP-DM

In the context of the literature studied, the CRISP-DM project structure proved to be the most robust in terms of use and easy understanding. The great permeability of CRISP-DM in market and academic environment makes its use more experienced and documented. Thus, for the purpose of this paper, we selected the CRISP-DM standard as more suitable for this data mining project.

According to Azevedo (2008) [1], CRISP-DM consists of a six-step cycle: 1) Understanding the business - understanding the objectives and requirements of the project from a business perspective, defining the problem and a preliminary plan to achieve the objectives; 2) Understanding data - collecting data and understanding it, identifying problems, seeking insight into relevant data and/or subsets to reveal hidden information; 3) Data preparation - activities to build the final data set; 4) Modeling - various modeling techniques are selected and applied and their parameters are calibrated; 5) Evaluation - the model obtained is evaluated in detail and revised to achieve the business objectives; 6) Deployment - usually not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge acquired must be organized and presented in a useful way for the user/client.

## B. TEXT MINING

The text mining is a kind of knowledge discovery about texts via data processing, however, a different way of data mining in general due to the nature of unstructured textual data, Bezerra (2010) [2]. Allayari (2017) [3] says that this simple but not trivial activity is based on the analysis of frequency of terms in each document and the construction of tables with their attributes and values, thus enabling the use of supervised and unsupervised machine learning algorithms for processing this information.

According to Nishanth et al (2012) [4], text mining is an intensive process of knowledge in which a user interacts with a collection of documents using a set of analytical tools. Text mining algorithms operate on representations of attributes present in documents. Characters, words, terms, and concepts are the potential resources used to represent documents. Text mining often involves the prior conversion of unstructured content into structured content before applying the usual data mining techniques. The data in the databases may be lost due to data entry errors, system crashes at the time of data recovery, or various other reasons. Thus, text mining becomes an effective tool in classifying texts and detecting patterns, categorization, or errors.

## C. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks are mathematical models inspired by the organic neural structure of intelligent living beings, which learn by experience, applied in computational techniques, Ninness (2013) [5]. With the rise of the importance of social networks and their impact on social and economic relations, the techniques of text mining have evolved systematically. In particular, as regards the mining of posts and comments, which are short and unstructured texts, artificial neural networks have stood out as a reference in mathematical modeling in the prediction of these data, Ali (2013) [6]. According to Guerrero et al (2017) [7], artificial neural networks (ANNs) can be used to classify the different concepts extracted in a semantic category.

## III. UNDERSTANDING THE BUSINESS

### A. SCENERY

All commercial transactions in the purchase and sale of goods in Brazil are obligatorily registered with electronic bills, whose data are stored in the state databases. By law [8], this issuance is done through a system controlled by the Treasury, all the data reported are available to the State. On the other hand, there is a representative aspect in these records that can distort them in relation to the real activities: The tax evasion and the errors of filling. Due to diverse cultural and socioeconomic factors, companies avoid taxes, therefore they could omit or alter information regarding their transactions in order to reduce in part or in a whole the tax burden.

### B. DETERMINING THE APPLICATION'S OBJECTIVES

The verification of the tax due is done electronically in large masses of data. Such analyzes depend on reading the

filled NFe fields. The field for product description, named "XPROD", is the most reliable because it is printed and the goods purchaser usually checks it for compliance reasons or guarantee terms. The XPROD field is a string type, up to 120 characters, with no padding pattern, so its reading and classification becomes unfeasible by traditional means of data crossing. In order to solve the problem of merchandise identification, it is used the field called NCM (Mercosul Common Nomenclature Product Code) as pattern, since it is a standardized 8 digits field (e.g. "28470000" - type: integer). However, it is also verified in the completion of these field, albeit rarer, an NCM that is totally different from the nature of the product described, such as in a observed case of "hydrogen peroxide", NCM "28470000", that was registered as NCM "01012100" which in the Mercosul pattern refers to "Live horses, asses and mules - Purebred breeding animals."

## C. PURPOSE OF APPLICATION

- Given the above is defined as a proposal of this work to solve the following problem: "Is it possible to classify goods by text mining from an electronic bill through the Description field?"
- It is important to note in this study that there is a difference between this text mining and the others verified in relation to its prediction characteristic, not of understanding or identifying the subject matter to which the text is treated, but rather of an objective categorization through the description of the object;
- And there is still a need for high accuracy in the prediction.

## D. APPLICATION SUCCESS CRITERIA

The criterion to be used to measure the success of the application will be the accuracy in the classification of the goods descriptions. The method currently used by the Department of Finance of Brasília, which uses the NCM field has an accuracy of 0.95 accuracy, so for the implementation of the proposed system this should overcome this assertiveness.

## IV. UNDERSTANDING DATA

The database provided by the Transit Goods Inspection Bureau contains the record of a 15,000 (fifteen thousand) rows by 3 (three columns) worksheet. There are 5,000 records classified for each category. The visual representation of the worksheet is shown in Table 1.

- "XPROD" - Description of the product in NFe, type text, string of 120 characters. The texts has no standard of filling. Is a field of free registration by the issuer of the note, how to write or describe a product is free, there is no criticism or rules in the NFe issuance systems;
- "NCM" - The NCM code, 8 digits (integers numbers);
- "Item" (38, 39 or 40) - a reference to the tax legislation indicating the taxation form, which varies according to the item. "38" for cosmetics, "39" for cleaning material and "40" for foods.

Tabela 1. Sample Database Table

XPROD	NCM	Item
- REF.: (01254) - PENTE PROMO - SENS	96151100	38
0037540 ST DEO COL BLUE OCEAN	33030020	38
0040002 ST DEO COLONIA PINK DIAM	33049910	38
0041416 ST RIMEL EXTRA VOLUME	33042010	38
0041436 ST 8 SABONETES CORACAO	34013000	38
...	...	...

## V. DATA PREPARATION

In this work the data preparation has the main focus to formatting the XPROD fields. As that is a text-type field, free and no padding pattern, it should be treated for use in modeling. The R platform was used in conjunction with the R Studio interface to carry out the work developed in this project. The database in a Excel spreadsheet was imported to R and a data base vector was created.

The NCM column has been deleted, the tax class column (Item) remained, 38 for Cosmetics, 39 for Cleaning Materials and 40 for Foods. In the "Item" column the numbers 38, 39 and 40 have been replaced respectively by the corresponding names "cosmeticos", "matlimpeza" and "alimentos" as that will be used as the classification parameter.

In the XPROD field the accent, punctuation, numbers, double spaces and special characters were removed. All text has been converted to lowercase. The stop words were also withdrawn. The "Item" column was renamed to "Category" and the "XPROD" column to "Text". The Table 2 shows how the database was after treatment. Such operations were made in R.

Tabela 2. Database after treatment

Text	Category
ref pente promo sens	cosmeticos
st deo col blue ocean	cosmeticos
st deo colonia pink diam	cosmeticos
st rimel extra volume	cosmeticos
st sabonetes coracao	cosmeticos
...	...

## VI. MODELING

In order to better evaluate the performance of artificial neural networks, the k-Nearest Neighbor - KNN and Gradient Boosting Machine - GBM prediction models will be evaluated in a comparative way. Therefore, some manipulations at the base of the model are still necessary. The inclusion of the "Category" column in the data frame as a reference for the test and training classification. The index is the lines order maintained until then.

For a clear overview at the modeling possibilities, the Word Clouds were made for each of the three categories. This is a visual resource that plots the most recurring words in classes where their size print is proportional to their frequency. The result is shown in Figures 1, 2 and 3.



Figura 1. Word cloud for Cosmetics (cosmeticos).



Figura 2. Word cloud for Foods (alimentos).

### A. ALTERNATIVE MODELS - KNN AND GBM

For the construction of the training and validation bases a reference column is inserted with the correct classification for each description. The training base has a random sample of 0.7 factor and the validation base has the remaining data, 0.3 fraction of the total initial base.

The KNN model was set in the native default form of the package used in R language (library: "class"). The GBM model was set with 1,000 trees and depth of 3. This arrangement was founded by a test combination trial: 300, 500, 1000 and 1500 trees, crossing 3, 5 and 7 depth. The H2O package for R was loaded for that task (library: "h2o"). The Confusion Matrices resulted is shown in Tables 3 and 4.



Figura 3. Word cloud for Cleaning Materials (*matlimpeza*).

Tabela 3. Confusion Matrix - KNN Model

Predictions	alimentos	cosmeticos	matlimpeza	Real
alimentos	968	20	4	992
cosmeticos	49	825	32	906
matlimpeza	41	137	924	1.102
Total	1.058	982	960	Acc: 90,56%

### B. ARTIFICIAL NEURAL NETWORKS MODEL

The training and validation bases of the artificial neural network model are the same used for the GBM model. This can show a more effectively comparison of performance between the two models. A network with 3 layers of artificial neurons was dimensioned, being 100 in the first layer, 200 in the hidden layer and 100 in the output layer. The epochs were 20 (twenty). This set was determined by testing several variants of these configurations, where layers of 50, 100, 200 and 300 neurons were disposed in 2, 3 and 4 layers, as well as training cycles, or epochs, of 20, 40 and 60. The set chosen was the one with the lowest error rate in predictions. The neural network confusion matrix was generated and its results are expressed in Table 5.

Based on these results, the superior performance of artificial neural network - ANN can be seen at looking to the KNN and GBM models in this specific case of text mining. At least, as regards the prediction accuracy of category described goods, the ANN model showed success rate higher than the other two models, although it is important to observe the tendency of greater error in the KNN for the "matlimpeza"(cleaning materials) category, GBM and the ANN models obtained a more uniform error between the categories as observed in their confusion matrices.

The  $R^2$  statistical term of the ANN and GBM models can also be compared for a more accurate assessment. The GBM model obtained an  $R^2$  of 0.9418, whereas the ANN model obtained  $R^2 = 0.9701$ . The parameter in the initial

Tabela 4. Confusion Matrix - GBM Model

Predictions	alimentos	cosmeticos	matlimpeza	Real
alimentos	878	103	13	994
cosmeticos	15	934	25	974
matlimpeza	10	120	867	997
Total	903	1.157	905	Acc: 90,35%

Tabela 5. Confusion Matrix - ANN Model

Predictions	alimentos	cosmeticos	matlimpeza	Real
alimentos	945	43	6	994
cosmeticos	16	939	19	974
matlimpeza	3	75	919	997
Total	964	1.057	944	Acc: 94,54%

purpose of this study, set out in item III, D, as 0.95 success rate in predicting came close to being hit with only the table used in the ANN model, which presented a sample of 5,000 descriptions for each category of merchandise. With the construction of a new base model with a more representative sample, the currently success rate can be overcome.

## VII. CONCLUSIONS

In view of the results presented, it was found that the classification of goods can be made by the descriptions given in the electronic bills through artificial neural networks. This study shows a satisfactory performance of the ANN algorithm for this specific text mining task. It is noteworthy that the text is not structured and has no fill rule, but it has small size, a maximum of 120 characters.

The classification of this type of text has its differential highlighted in relation to textual ratings commonly seen in the literature as well as in technical reports widely available on the Internet, is descriptive of the goods itself, not the meaning or significance of the analyzed text. When the most frequent need for cases of text mining shows the classification of the meaning of the text verified, in this issue the goal is the straight classification of merchandise which is being described in the text.

This type of categorization of goods can be expanded in future studies to more classes or even for all 99 classes listed in the Mercosul Common Nomenclature - NCM, which would help the FISCO identify filling errors, tax classification or even fraud.

## Referências

- [1] A. Azevedo, M. F. Santos, KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW, IADS European Conference Data Mining, 2008.
- [2] E. Bezerra, R. Goldschmidt, Classification in Text Mining. journal Sistemas de Informação, n 5, 2010, pp. 42-62.
- [3] M. Allahyari, P. Seyedamin, A. Mehdi, E. D. Trippe, J. B. Gutierrez, K. Kochut, A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. KDD Bigdas, Halifax, Canada, 2017.
- [4] K. J. Nishanth, V. Ravi, N. Ankaiah, I. Bose, Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Systems with Applications, vol. 39, pp. 10583-10589, 2012.
- [5] C. Ninness, R. Marilyn, C. Logan, L. David, J. T. Lacy, S. Halle, R. McAdams, S. Parker, D. Forney, Neural Network and Multivariate Analyses:

Pattern Recognition in Academic and Social Research, Behavior & Social Issues, vol. 22, 2013, pp. 49-63.

- [6] M. P. J. Ali, and N. M. S. Surameery, A. R. M. Yunis, Gender Prediction of Journalists from Writing Style, ARO. The Scientific Journal of Koya University, vol. 1, 2013, pp. 22-28.
- [7] J. I. Guerrero, C. Len, I. Monedero, F. Biscarri, J. Biscarri, Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection, Knowledge-Based Systems, vol. 71 2014, pp. 376-388.
- [8] Sistema Integrado de Informações Econômicas e Fiscais- SINIEF. Ministry of Finance, Convênio s/n, National Council of Finance Policy CONFAZ, 1970.



**VINICIUS DI OLIVEIRA** Possui graduação em Engenharia Civil pela Universidade Federal de Goiás (1997). Especialista em Modelagem Estatística e Dinâmica Econométrica pela UnB, mestrando em Computação Aplicada na UnB e Atualmente é Auditor Fiscal da Receita do DF - Secretaria de Estado de Fazenda do Distrito Federal. Tem experiência na área de Fiscalização Tributária, atualmente com ênfase em Gerenciamento de Risco de Projetos.



**MARCELO LADEIRA** Possui graduação em Engenharia Mecânica pela Universidade de Brasília(1976), especialização em Curso Avançado de Operação de Sistema Hidrotérmico pela Universidade Federal do Rio de Janeiro(1981), mestrado em Análise de Sistemas e Aplicações pelo Instituto Nacional de Pesquisas Espaciais(1981), doutorado em Computação pela Universidade Federal do Rio Grande do Sul(2000) e pós-doutorado pela Universidade de Lisboa(2004). Atualmente é Professor

Adjunto 2 da Universidade de Brasília, Membro de Comitê Consultivo do Ministério da Ciência, Tecnologia, Inovações e Comunicações, Revisor de periódico da Journal of Information and Data Management - JIDM e Membro de comitê assessor do Conselho Nacional de Desenvolvimento Científico e Tecnológico. Tem experiência na área de Ciência da Computação, com ênfase em Metodologia e Técnicas da Computação. Atuando principalmente nos seguintes temas: Inteligência Artificial, raciocínio probabilístico, diagramas de influências, redes bayesianas, Teoria da decisão e Representação de conhecimento incerto.

...

## APPENDIX

### R coding for the ANN and the GBM models.

```
set.seed(321)
library(readxl)
library(tm)
library(h2o)
h2o.init()
df <- read_excel("yourdatabase.xlsx")
docs <- Corpus(VectorSource(df\$$Text))
docs <- tm_map(docs,
content_transformer(tolower))
docs <- tm_map(docs,
removeNumbers)
docs <- tm_map(docs,
removeWords, stopwords("pt"))
docs <- tm_map(docs,
removePunctuation)
docs <- tm_map(docs,
stripWhitespace)
dtm <- DocumentTermMatrix(docs)
mat.df <- as.data.frame(data.matrix(dtm),
stringsAsFactors = FALSE)
mat.df <- cbind(mat.df, df\$$Category)
colnames(mat.df)[ncol(mat.df)]
<- "category"
mat.df.h <- as.h2o(mat.df)
data.split <- h2o.splitFrame
(data = mat.df.h, ratios = c(0.7, 0.2),
seed = 1234)
data.train <- data.split[[1]]
data.valid <- data.split[[2]]
data.test <- data.split[[3]]
myY <- "category"
myX <- setdiff(names(data.train),
c(myY, "ID"))
gbm.model <- h2o.gbm(myX, myY,
training_frame = data.train,
validation_frame = data.valid,
ntrees = 1000,
max_depth = 3,
model_id = "gbm_xprod_5mil")
conf.mat <- h2o.confusionMatrix
(gbm.model@model\$$validation_metrics)
write.table
(conf.mat, file="confmat_gbm_5mil.csv",
sep=";")
r2.gbm.model.5mil <-
gbm.model@model\$$validation_metrics@metrics\$r2
dl.model <- h2o.deeplearning(myX, myY,
training_frame = data.train,
hidden = c(100,200,100),
epochs = 20,
validation_frame = data.valid,
model_id = "dl_xprod_5mil")
conf.mat.dl5mil <- h2o.confusionMatrix
(dl.model@model\$$validation_metrics)
write.table(conf.mat.dl5mil,
file="confmat_dl_5mil.csv", sep=";")
r2.dl.model.5mil <-
dl.model@model\$$validation_metrics@metrics\$r2
```

# **Singular Engenharia, Tecnologia e Gestão**

**Vol. 1, N. 1, Fevereiro, 2019**

**eISSN: 2596-2604**

**<https://doi.org/10.33911/singular-etg.v1i1>**